Ontology Building: an Application in Food Risk Analysis

Sandrine Blanchemanche[1], Patrice Buche[1], Juliette Dibie-Barthélemy[1,2], Eve Feinblatt Mélèze[1], Liliana Ibanescu[1,2] and Akos Rona-Tas[3]

[1]Met@risk – INRA, [2]UFR Informatique, AgroParisTech
16 rue Claude Bernard, F-75231 Paris Cedex 05
[3] University of California, San Diego, 9500 Gilman Dr., La Jolla, CA, 92093-0533
Sandrine.Blanchemanche@paris.inra.fr,
Patrice.Buche@paris.inra.fr, Juliette.Dibie@agroparistech.fr,
Eve.Feinblatt@agroparistech.fr,
Liliana.Ibanescu@agroparistech.fr, aronatas@ucsd.edu

Abstract: This paper presents the methodological steps we followed to build an ontology in the food risk analysis domain. In the first, or preparation phase, we defined the domain and scope. In the second or building phase, first, an ontology was created using a representative subset of food risk assessment reports. Then, the ontology was refined taking into account the entire set of 74 reports. Finally, in the third or refinement phase we mapped our ontology onto existing classifications, validated them by outside experts and double coded them by independent inside coders. The entire process was aided by a database prototype in Access and a simple text processing tool.

Key words: Ontology building, annotation

1. Introduction

The development of ontologies is a lively research area in many different domains of computer science including knowledge engineering, computational linguistics, information retrieval and semantic web research. There are many definitions of what ontology is. Gruber (Gruber, 1993) defines ontology as "an explicit specification of a conceptualization". Other definitions depend on the purpose of the ontology. Ontologies can be used to represent domain knowledge or to mediate between knowledge expressions (Bachimont, 2007) or to help software and human agents communicate and share domain knowledge (Maedche, 2002). The ontology that we propose allows human agents to share information in a domain, by a machine-readable representation of the domain knowledge, with the help of a set of concepts characterized by properties and linked by relations. There are many methodologies (Bachimont & Isaac *et al.*, 2002), (Gangemi, 2005), (Bendaoud & Napoli *et al.*, 2008), (Mondary & Deprès *et al.*, 2008) for building ontologies and there are various tools to assist users in this task. This variety reflects the complexity of such a task and the ambiguity inherent in any definition of ontology (see Corcho & Fernandez-Lopez *et al.*, 2003 for a state of art).

This paper presents the methodological steps we followed to build an ontology in the food risk analysis domain. The ontology represents the scientific uncertainty expressed in risk analysis reports written by scientific panels at the request of risk managers who have to make decisions about protecting the public against the food risk analyzed. Indeed, in recent years, a large body of work emerged on the role of scientific uncertainty in policy making. This role is strongly influenced by institutional and regulatory conditions that vary from country to country. Comparisons between the United States and the European countries show that they follow different paradigms in food safety partly because of their institutional, political and legal differences. Except for a few case studies, there has not been any systematic *empirical* research on how uncertainties of various kinds are actually expressed in risk assessment reports, how prevalent each type is, and what kind of actual consequences they have for future assessments and for the rest of the regulatory process. To better understand the role of uncertainty in policy making, such an inquiry was necessary and we developed the HolyRisk Project involving scientists from sociology, economics, risk analysis and computer science. This project is a US/EU comparative empirical study that investigates the ways uncertainties are perceived, handled and expressed by experts throughout the food risk analysis process (risk assessment, risk management and risk communication). We are especially interested in national (or US/EU) differences in the way uncertainty is reported and the manner in which regulatory decisions are influenced by the level and kind of uncertainty reported in the risk assessment documents. Finally, our research aims to provide useful knowledge and computer tools to assist risk assessors with conducting their scientific inquiry and presenting their findings, and risk managers with their decision making process. The construction of an ontology of uncertainty concepts appeared to be the core of the representation and management of knowledge needed in the HolyRisk project.

In this paper, we do not propose a new ontology building method, but present an example of a step by step empirical construction of an ontology in a particular application: food risk analysis. Several characteristics of this field make the building of an ontology challenging: i) risk assessors are experts from different disciplines (toxicology, statistics, epidemiology, chemistry, etc.) and use different tools and frameworks to deal with uncertainty, ii) the food safety field itself covers several domains of hazards (contaminants, microorganisms, GMOs[1]) which require the use of different sources of data and methods and use of expert conventions, iii) American and European experts do not proceed in exactly the same way in evaluating risk and presenting conclusions (including uncertainties) in the reports.

We were inspired by the guidelines of the simple knowledge engineering methodology presented in (Noy & McGuinness, 2001). Our empirical approach for ontology building can be decomposed in three main phases which are presented in the next three sections: preparation, building and refinement. Then we present the

[1] GMO stands for genetically modified organism

structure of the obtained ontology. Finally we conclude and present the planned future trajectory of this work.

2. Preparation phase

In the preparation phase the main task is the definition of the domain and the scope of the ontology, the gathering of sources in the domain and the identification of existing ontologies.

The definition of the domain and the scope of the ontology are essential to lay the correct foundations of the work and help the people involved in the building of the ontology by giving them the necessary guidelines. In our application, the ontology has been defined to represent expressions of scientific uncertainties in food risk analysis. It will be used to determine in what ways scientific uncertainty expressed in food risk analysis influences the decision taken by risk managers. The domain and the scope of the ontology were identified by the social scientists and the risk analysts involved in the HolyRisk project, who are considered experts in the ontology building process. The corpus of the relevant documents was identified and they were located and collected.

Ontologies for the expression of uncertainty in food risk analysis do not exist but partial elements could be extracted from three types of sources: the scientific literature on uncertainty, the existing classifications of uncertainty developed by food safety agencies (institutional literature) and risk assessment reports.

In the scientific literature, there are different terms used according to the field of research. Among these terms appears a common distinction between variability and uncertainty (Thomson, 2002), (Suter & Barnthouse *et al.*, 1987), (Rothmans & Van Asselt, 2001), (Walter & Harremoës *et al.*, 2003), (Natke & Ben-Haim, 1996), (Morgan & Henrion, 1990) and (Kandlikar & Ramachandran *et al.*, 2007). This distinction relates to the role of additional information: uncertainty can be reduced by more knowledge while variability is an inherent property of a physical, chemical, biological or social system which cannot be further reduced by additional data. In order to take into account the subsumption links between the terms, we proposed to build a hierarchical system composed of concepts representing different levels of generality. The process moves down from the most general distinction between variability and uncertainty to more and more specific categories stopping at the level of generality beyond which there is no more specific category that could accommodate the expression of uncertainty. Through this hierarchy, we tackled uncertainties that arise in the different disciplines like epidemiology, microbiology, toxicology or exposure assessment, involved in the food risk assessment process (Grandjean & Budz Jorgensen, 2007), (Kang & Kodell *et al.*, 2000), (Dorne & Renwick, 2005) and (Kroes & Muller *et al.*, 2002). This literature helped us to be more specific especially in taking into account the uncertainties resulting from inferences (such as inference from in vitro observations to in vivo situations, from the general to a special - e.g., sensitive - population, from laboratory animals to humans etc.) when conclusions must be drawn from evidence only indirectly addressing the issue at hand.

Beside this literature, we used two main institutional documents: the opinion of the Scientific Committee of EFSA entitled Uncertainties in Dietary Exposure Assessment (EFSA, 2006) and the WHO Draft guidance document on Characterizing and Communicating Uncertainty in Exposure Assessment (WHO, 2007). Both documents call for a tiered approach to the characterization of uncertainty: qualitative, deterministic or probabilistic, according to the result of the assessment. These classifications were very helpful but they had three main limits for our purposes. First, their goal is to provide guidelines for experts on how to describe uncertainties according to their source and nature. This "normative" approach did not correspond to our own objective which was to obtain an ontology able to represent how uncertainties are, in fact, expressed in the US/EU final reports. For instance, "contradictory findings between experts" by definition cannot be found in these prescriptions but are of great importance for us. Second, the number of uncertainty categories described in these guidelines was far too big for our objective of comparison among documents and cases and for the practical task of coding the texts. For instance, the EFSA document globally describes 126 categories of uncertainty combining the different sources of uncertainty with the different types of uncertainty. Third, both guidelines focus on exposure assessment, a specific step of the risk assessment process, directly related to risk management. We wanted our ontology to be exhaustive and to take into account the whole process of risk assessment based on the commonly used four-step framework: hazard identification, exposure assessment, hazard characterization, risk characterization.

3. Building phase

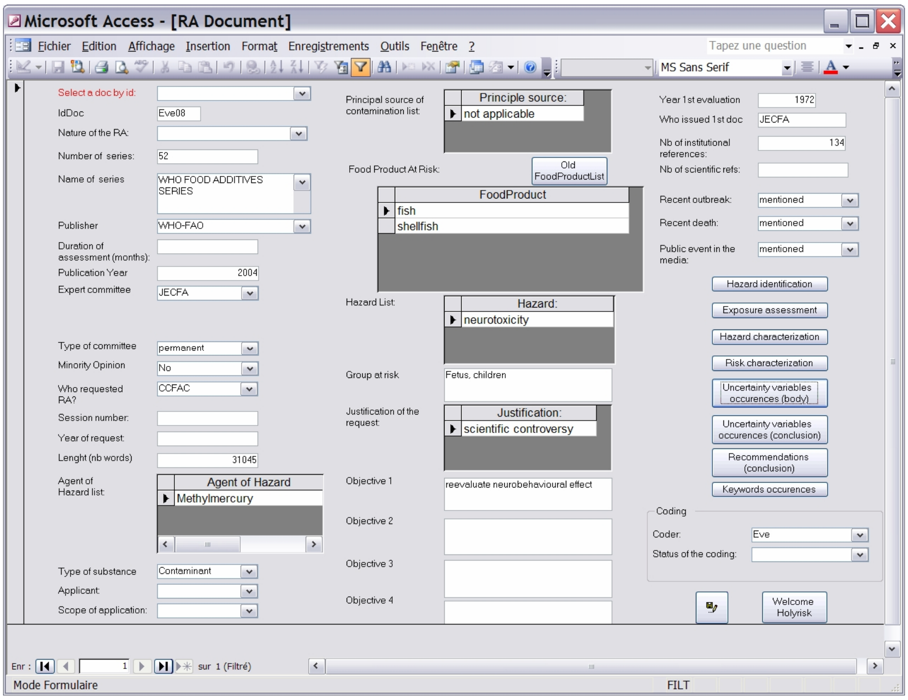
In the building phase the task is the construction of the ontology. It consists of enumerating important terms, organizing them in taxonomy and defining their properties. This task leads to a "first version" of the ontology that must be tested on data sources unused in the task.

3.1 Initial step

To overcome the limits of existing classifications, we elaborated an empirical approach in annotating the risk assessment reports. Four experts (from sociology) worked in a collaborative way to define "manually" a first version of the ontology, called HOLY1. The expression of uncertainty was extracted from 74 risk assessment reports covering three categories of hazard (GMO, chemical and microbial contaminants) and three national sources of reports (international, EU, US). Reports are documents ranging from a dozen to several hundred pages written by scientists from different disciplines. In order to help experts to extract expressions of uncertainty from such complex documents, two tools were developed: a modeling of the report including the representation of uncertainty, and a semi-automatic text processing tool to extract the paragraphs of the report expressing uncertainty.

Modeling of the reports: The aim of modeling the reports is to deliver a standard reading pattern of the reports, to provide a summary of them and to help the experts understand and analyze them. A database prototype using Microsoft Access has been created for this purpose. Access was chosen because it allows the structure of the database and the GUI (Graphical User Interface) to be designed using the AGILE method, allowing quick and reactive exchanges between the computer team and users. The risk assessment database prototype was built iteratively in about three months. This database contains about 20 tables. Figure 1 presents the main screen of the HolyRisk database where the main pieces of information about a report appear (references of the report, agent of hazard, food products at risk etc.). Sub-screens are accessible through the buttons on the right side of the main screen. Each button gives access to the information associated with the 4 steps of a risk assessment: hazard identification, exposure assessment, hazard characterization, risk characterization. These buttons also provide access to expressions of uncertainty found in the body and the conclusion of the report to be accessed.

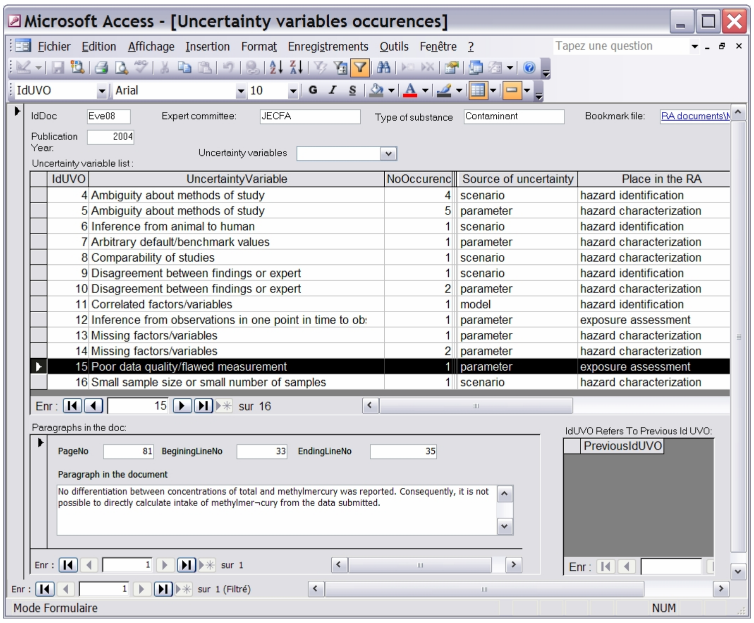
Figure 1 -- Main screen of the HolyRisk database



Developing a semi-automatic text processing tool, called FindParagr: In order to find most of the paragraphs of the documents where uncertainty is expressed, a list of 37 key-words often associated with the expression of uncertainty was established manually. The list included words like: ambiguity, assumption, uncertainty, probably, very likely, may etc. An MSWord macro was developed to search automatically for paragraphs containing a combination (conjunctive or disjunctive) of these key-words. It is important to note that the list of key-words cannot be exhaustive and that uncertainty can be expressed more or less explicitly. Additionally, some words may or may not reflect uncertainty, depending on their context. Therefore, the list of words was used as flags for the readers pointing them to possible expressions of uncertainty. Because uncertainty could be expressed without using any of the words on the list, the reader had to read the entire text not just the paragraphs flagged by text processing tool.

Example 1. In the following example extracted from a risk assessment report on Organotins (EFSA 2003), the paragraph was identified by FindParagr because of the key-word "may": *"The Panel noted that the consumption of fish, mussels and other marine animals from highly contaminated area, such as the vicinity of harbors and heavily used shipping routes,* ***may*** *lead to OTC intake that exceed the group TDI."* In this example, a modal term is used to express the possibility that the final estimation of the risk (based on the "'tolerable daily intake"' calculated in the assessment of the toxic effects of organotins and on the exposure of the population) is uncertain, since some scenarios (consumption of fish coming from highly contaminated areas) have been omitted from consideration. It is possible that for the quoted scenario the exposure exceeds the tolerable daily intake set by the experts. Thus, we move from the certainty of no risk to the uncertainty of a possible risk.

Figure 2 -- Expression of uncertainty in the body of the report



The annotation process was performed on a subset of 10 representative reports previously selected from the initial 74 documents with the help of our semi-automatic tool FindParagr. Starting from the survey of the scientific literature completed in the preparation phase, the experts established a first list of 10 terms, called "uncertainty variables". After that, each paragraph extracted from the 10 reports was annotated using this list of "uncertainty variables." The annotation process produced an updated and extended list of 18 "uncertainty variables", which corresponds to the first version of the ontology, called HOLY1. This iterative process was carried out in a collaborative manner and its results were established by the consensus of the four experts involved in the project and were stored in the database.

Figure 2 shows a part of the list of the 16 uncertainty variable occurrences found in the report on MethylMercury. Occurrence in position 15, in black, corresponds to the uncertainty variable "Poor data quality" and is associated with the paragraph extracted from page 81 of the report whose text is also stored in the database: *"No differentiation between concentrations of total and methylmercury was reported. Consequently, it is not possible to directly calculate intake of methylmercury from the data submitted."*

3.2 Second step

The second step of the building phase consisted of progressively refining the ontology HOLY1 with the help of the unused reports. Here we asked experts of different domains to annotate various documents according to the first version of the ontology and to use these annotations to revise this ontology. More precisely, this refinement involved first annotating the conclusion section of each risk assessment report using the uncertainty terms of the ontology HOLY1. The annotation was limited to the conclusion section of the reports to simplify the work of the domain experts and therefore allow them to annotate all the 74 reports. The conclusion section contains the principal message the risk assessors want to transmit to the risk managers. The results of this refining step are also stored in the Access database. The annotations and the associated ontology refinements were done by a close collaboration between domain experts and every decision was regularly discussed and justified. This collaborative work led the domain experts to modify existing ontology terms, to define new ones, to suppress others, to revise the taxonomy, and to clarify the terms and -- in particular -- their use. The domain experts have noticed that many terms were ambiguous: for several annotations, each domain expert chose a different term, each choice being justified according to his/her own interpretation. The domain experts thus proposed to add properties to each term: each term was characterized by a definition and was additionally characterized by comments and examples of use found in the reports. The result of this step was a second version of the ontology, called HOLY2, composed of 38 terms.

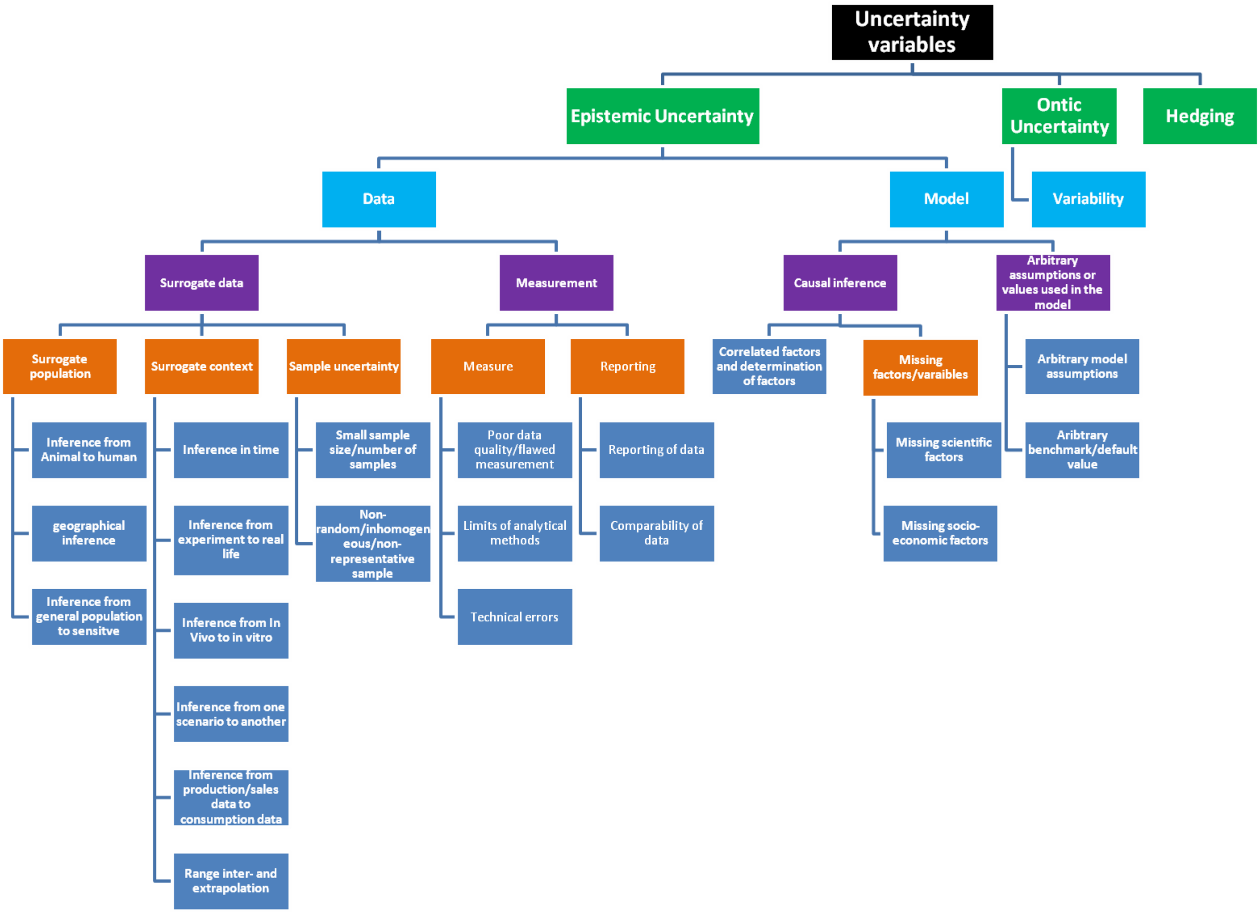
4. HOLY2 Ontology

The ontology HOLY2 is composed of a hierarchy of 38 concepts (see Figure 3), these concepts being represented by terms. The concepts are linked by subsumption relations.

Example 2. The uncertainty expressed into the paragraph given in Example 1 corresponds to the concept "geographical inference" because of the following hierarchical reasoning (see Figure 3):

1. The uncertainty could be reduced by more knowledge on highly contaminated areas: it is an **Epistemic Uncertainty**
2. Uncertainty arises because of the **data** used.
3. The experts derived a risk estimate for the general population. We don't have the right observations for a specific population. Uncertainty comes thus from a **surrogate population**, which is a sub-concept of **surrogate data**.
4. More specifically, the observations come from one set of countries and the results are not directly transferable to the situation in another country. This corresponds to the concept: **Inference from <u>data coming from one/some geographical location to a broader or other geographical location</u>** (simplified as "geographical inference" in Figure 3).

Figure 3 -- Hierarchy of concepts of the ontology HOLY2



Each concept is characterized by three properties: a definition, a comment and some examples.

Example 3. The Concept “Arbitrary benchmark/default value”, a sub-concept of the concept “Arbitrary assumptions or values used in the model”, which is a sub-concept of “Model”, in the hierarchy has the following properties:

- Definition: In the calculations certain parameters are unmeasured or unreported, and the particular quantity is estimated in a way that it is one of several potential values.
- Comment: “The assumption does not have to be a point estimate; it can be a range of values.”
- Example: (1) “Pending reduction in the uncertainty associated with various aspects of the derivation of the steady-state intake from maternal concentrations of mercury in hair, the Committee concluded that the uncertainty factor could be refined and possibly reduced”.

This example was extracted from a risk assessment report of Methylmercury (JECFA 2002). The paragraph was flagged by the semi-automatic text processing tool with a research on the key-word “uncertainty.” This paragraph explicitly expresses uncertainty. The uncertainty factor used is a default uncertainty factor rather than a one calculated for this specific case.

5. Refinement phase

In the building phase presented above, we have exploited the reports in two ways. First, a representative subset of 10 reports was analyzed by scanning the entire documents in order to find and classify uncertainty expressed in the 4 steps of risk assessment: hazard identification, exposure assessment, hazard characterization, risk characterization. The result was HOLY1, a first version of the ontology, which contains a list of 18 “uncertainty variables.” Second, the conclusion section of the remaining reports (64 reports) were annotated with terms of HOLY1 producing the ontology HOLY2 with 38 uncertainty variables”.

In a first refinement step, the list of 38 uncertainty variables from HOLY2 was mapped onto the list of 126 EFSA categories mentioned above in order to check if important categories were missed.

In a second refinement step, to further validate our ontology, two domain experts, not involved in the building phase of the ontology HOLY2, were asked to annotate the conclusion section of the reports with their own words. Disagreements with annotations based on ontology HOLY2 were discussed and some definitions of HOLY2 were revised.

A last refinement step is under process. It consists of a double annotation of the same reports done by two different domain experts (one from the US and the other one from EU) using the ontology HOLY2. The objective is to compare the annotations in order to detect remaining ambiguities in the definitions of concepts.

6. Conclusion

This paper presented an empirical methodology constructing an ontology to represent scientific uncertainty expressed in food risk analysis reports. Due to the complexity and the heterogeneity of the available sources, we did not use in this preliminary work methods and tools like automatic extraction of terminological relations between concepts (Mondary et al. 2008) because it would have required reasoning methods that are too complex. The building of the ontology was done in three stages. First, the ontology HOLY1 was created using a representative subset of 10 reports and mapped with institutional guidelines (EFSA and WHO) in a collaborative way, involving 4 experts from sociology. Second, the ontology was revised taking into account the entire set of 74 reports by annotating the paragraphs that express uncertainty and the ontology HOLY2 was constructed. Thirdly, annotations using the ontology HOLY2 were compared to « free » annotations done by risk analysts experts not involved into the building task of HOLY2. The entire process relies on a database prototype in Access.

We plan to build three additional tools for the creators and users of the HolyRisk data base. We want to develop machine learning methods which will be able to use manual annotations, built from the ontology and associated with paragraphs, to predict annotations of new paragraphs, thus providing more sophisticated assistance for the coding of the documents. This tool will replace FindParagr, our current text processing aid. We also plan to develop case-based reasoning methods. These will assist risk managers in comparing new cases to the collection of cases registered and annotated in our database of reports. And finally, we will build a visual tool that allows readers of the documents to navigate easily across texts to find similar types of uncertainties in different reports.

References

BACHIMONT B. (2007). Ingénierie des connaissances et des contenus. Le numérique entre ontologies et documents. Hermes science, Lavoisier, 2007.

BACHIMONT B. & ISAAC A. & TRONCY R. (2002). Semantic Commitment for Designing Ontologies : A Proposal. In Gomez-Pérez A. and Richard Benjamins V. editors, *Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management. Ontologies and the Semantic Web, EKAW'2002*. LNCS 2473.

BENDAOUD R. & NAPOLI A. & TOUSSAINT Y. (2008). Formal Concept Analysis: a Unified Framework for Building and Refining ontologies. In A. Gangemi and J. Euzenat editors, *Knowledge Engineering: Practice and Patterns – Proceedings of the 16th International Conference on Knowledge Engineering and Knowledge Management, EKAW'2008*, p. 156-171.

CORCHO O. & FERNANDEZ-LOPEZ M. & GOMEZ-PEREZ A. (2003). Methodologies, tools and languages for building ontologies. Where is their meeting point? *Data and Knowledge Engineering 46*, p. 41-64, Elsevier.

DORNE J. L. & RENWICK A.G (2005). The Refinement of Uncertainty/Safety Factors in Risk Assessment by the Incorporation of Data on Toxicokinetic Variability in Humans. *Toxicological Sciences 86*, p. 20–26.

EFSA. (2006). Guidance of the Scientific Committee on a request from EFSA related to Uncertainties in Dietary Exposure Assessment, *The EFSA Journal*, 438, p. 1-54.

GANGEMI A. (2005). Ontology Design Patterns for Semantic Web Content. In *Semantic Web – Proceedings of the 4th International Semantic Web Conference, ISWC'2005*. LNCS 3729.

GRANDJEAN P. & BUZ-JORGENSEN E. (2007). Total Imprecision of Exposure Biomarkers : Implications for calculating Exposure Limits. *American Journal of Industrial Medicine*, 50, p. 512-519.

GUBER T. A. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition 5(2)*, p. 199-220.

KANDLIKAR M. & RAMACHANDRAN G. & MAYNARD A. & MURDOCK B. & TOSCAN W.A. (2007). Health risk assessment for nanoparticles: A case for using expert judgment, *Journal of Nanoparticle Research,* p. 137-156.

KANG S. & KODELL R.L & CHEN J.J. (2000). Incorporating model uncertainties along with data uncertainties in microbial risk assessment, *Regulatory Toxicology and Pharmacology* 32, p. 68–72.

KROES R. & MULLER D. & LAMBE J. & LOWIK M. R. H., & VAN KLAVEREN J. & KLEINER J. & MASSEY R. & MAYER S. & URIETA I. & VERGER P. *ET AL.* (2002). Assessment of intake from the diet. *Food Chem. Toxicol.* 40**,** p. 327–385.

MAEDCHE A. (2002). Ontologies Learning for the Semantic Web. Springer, 2002.

MONDARY T., DESPRÈS S., NAZARENKO A., SZULMAN S. (2008). Construction d'ontologies à partir de textes: la phase de conceptualisation. 19ième Journée Francophones d'Ingéniérie des Connaissances, Nancy, France, p. 87-98.

MORGAN M.G. & HENRION M. (1990). Uncertainty: A Guide to Dealing With Uncertainty in Quantitative Risk and Policy Analysis. Cambridge, Cambridge University Press.

NATKE H.G. & BEN-HAIM Y. (1996). Uncertainty: a discussion from various points of view. In H. G. Natke and Y. Ben-Haim, eds., *Uncertainty: Models and measures*, Akademie Verlag, Berlin, Germany.

NOY N. F. & MCGUINNESS D. L. (2001). Ontology Development 101: a guide to creating your first ontology. *Stanford Knowledge Systems Laboratory Technical Report*, march 2001.

ROTHMANS J. & VAN ASSELT M.B.A. (2001). Uncertainty in integrated assessment modeling : A labyrinthic path. *Integrated assessment*, 2, p. 43-55.

SUTER G. W. I. I. & BARNTHOUSE L.W. & O'NEILL R.V. (1987). Treatment of risk in environmental impact assessment. *Environmental Management*, 11, p. 295–303.

THOMPSON K.M. (2002). Variability and Uncertainty meet risk management and risk communication, *Risk Analysis*, 22, p. 647-654.

WALKER W.E. & HARREMOËS P. & ROTMANS J. & VAN DER SLUIJS J.P. & VAN ASSELT M.B.A. & JANSSEN P. & KRAYER VON KRAUSS M.P. (2003). Defining Uncertainty A Conceptual Basis for Uncertainty Management in Model-Based Decision Support. *Integrated Assessment*, 4, p. 5-17.

WHO/IPCS (World Health Organization/International Program on Chemical Safety) (2007). Guidance Document on Characterizing and Communicating Uncertainty in Exposure Assessment.