INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

# *Team orpailleur*

# *Extraction de connaissances*

*Lorraine*

THEME 3A

# Table of contents

# 1. Team

**Head of the team**

Amedeo Napoli [Directeur de Recherche CNRS]

**Administrative assistant**

Antoinette Courrier [Technicienne CNRS]

**Staff members**

Marie-Dominique Devisgnes [Chargée de Recherche CNRS]
Florence Le Ber [Professeure, ENGEES Strasbourg]
Jean Lieber [Maître de conférences, Université Henri Poincaré — Nancy 1]
Jean-François Mari [Professeur, Université de Nancy II]
Emmanuel Nauer [Maître de conférences, Université de Metz]
Malika Smaïl [Maître de conférences, Université Henri Poincaré — Nancy 1]
Yannick Toussaint [Chargé de Recherche INRIA]

**Doctorants**

Rim Al Hulou [ATER Université de Nancy 2, Thèse 05-11-2003]
Mathieu d'Aquin [doctorant, bourse MENRT]
Martine Cadot [doctorante, PRAG, Université Henri Poincaré — UHP Nancy 1]
Fairouz Chakkour [ATER Université Henri Poincaré — Nancy 1, Thèse 11-12-2003]
Hacène Cherfi [doctorant, ATER Université Henri Poincaré — UHP Nancy 1]
Nicolas Jay [doctorant avec co-encadrement, interne des hôpitaux de Nancy]
Sébastien Hergalant [doctorant avec co-encadrement, bourse INRA-Région]
Sandy Maumus [doctorante avec co-encadrement, bourse INSERM-Région]
Jean-Luc Metzger [doctorant, ATER Université de Nancy 2]
Laszlo Szathmary [doctorant, bourse KVM-Région]

**Post-doctoral fellow**

Huaizhong Kou [Post Doctorant, ACI MDA (01-12-2003)]

**Visiting scientist**

Dietmar Janetzko [Professor, Universität Freiburg, novembre et décembre 2003)]

**Technical staff**

Sébastien Brachais [Ingénieur associé INRIA]

# 2. Overall Objectives

The "orpailleur" is a French word for a person who is searching for gold in the rivers. In the present case, gold nuggets are in correspondence with knowledge units. These knowledge units may have two major different sources: explicit knowledge that can be given by domain experts, or implicit knowledge that must be extracted from databases of different kinds, e.g. rough data or textual documents. Moreover, the knowledge units must be represented in adequate formalisms for being used in a number of processes such as information retrieval or problem-solving.

The main objective of the members of the Orpailleur team is to extract knowledge units from different sources and to design structures for representing the extracted knowledge units. Closing the loop, knowledge-based systems may then be designed, to be used in a number of application domains such as agronomy, biology, chemistry, medicine, the Web...

The research work of the Orpailleur team may be considered from three main interrelated viewpoints: knowledge extraction, knowledge representation, and semantic Web. First, the data sources are prepared to be processed, then they are mined, and finally, the extracted information units are interpreted for becoming knowledge units. These units are in turn embedded within a representation formalism to be used within a

knowledge-based system. The mining processes are based on the *classification* operation, e.g. lattice-based classification, frequent itemset search. The mining process may be guided by a domain *ontology*, that is considered as a domain *model*, used for interpretation and reasoning.

The whole transformation process from rough data into knowledge units is based on the underlying idea of *classification*. Classification is a polymorphic tool involved in a number of tasks within the transformation process from data to knowledge: within the mining process, the modeling of the domain for designing a domain ontology (possibly extended with extracted knowledge units), and knowledge representation and reasoning. Finally, the knowledge extraction process and the associated knowledge base can be used for problem-solving and for achieving different tasks within the framework of the Semantic Web, e.g. Web mining, intelligent information retrieval, content-based document mining...

*2.1.1. Note on the organization of the report.*

Regarding the organization of this report, for convenience, applications and scientific results are not presented in specific sections, but, instead, follow the theoretical topics on which they are based.

# 3. Scientific Foundations

## 3.1. Knowledge Discovery in Databases

**Key words:** *knowledge discovery in databases*, *data mining methods*, *lattice-based classification*, *frequent itemset search*, *association rule extraction*, *hidden Markov models for data mining*, *text mining*, *bioinformatics*.

**Participants:** Martine Cadot, Fairouz Chakkour, Hacène Cherfi, Sébastien Hergalant, Huaizhong Kou, Florence Le Ber, Jean-François Mari, Sandy Maumus, Amedeo Napoli, Emmanuel Nauer, Yannick Toussaint.

*Glossary*

**knowledge discovery** is a process for extracting information units from large databases, units that can be interpreted to become knowledge units to be reused.

### *3.1.1. Symbolic Methods in Knowledge Discovery*

*Knowledge discovery in databases* (KDD) consists in processing a huge volume of data in order to extract useful and reusable knowledge units from these data. An expert of the data domain, called the *analyst*, is in charge of guiding the extraction process, on the base of his objectives and of his domain knowledge. The extraction process is based on data mining methods returning information units from the considered data. The analyst selects and interprets a subset of the units for building "models" that will be further considered as knowledge units with a certain plausibility.

The KDD process is performed with a KDD system based on four main components: the databases, a domain ontology (associated with a knowledge-based system), data mining modules (either symbolic or numerical), and interfaces for interactions with the systems, e.g. editing and visualization. A KDD system is aimed at handling huge volume of data in a given domain. For achieving this task, the system may take advantage of domain knowledge, i.e. an ontology, and the problem-solving capabilities of a knowledge-based system working in the domain of data. In turn, the knowledge units extracted by the KDD system may be integrated within the ontology to be reused by the knowledge-based system for future problem-solving operations.

*3.1.1.1. Lattice-based classification, frequent itemset search, and association rule extraction.*

Lattice-based classification can be considered as a symbolical data mining technique that can be used for extracting from a database or a set of rough data a set of concepts organized within a hierarchy (i.e. a partial ordering), frequent itemsets i.e. sets of properties (characteristics of data), occurring together with a certain frequency, or association rules with a given confidence (association rules emphasize correlations between sets of properties). Lattice-based classification relies on the analysis of boolean tables relating a set of individuals with a set of properties (or characteristics), where *true* stands for the individual `i` has the property `j` (the

relation between individuals and properties can be read as follows: the individual `i` includes or does not include the property `j`). The lattice may be built according to the so-called *Galois* correspondence, classifying within a formal concept a set of individuals, i.e. the extension of the concept, sharing a same set of properties, i.e. the intension of the concept.

In a parallel way, the extraction of frequent itemsets consists in extracting from boolean tables sets of properties occurring with a support or frequency, i.e. the number of individuals sharing the properties, greater than a given threshold. From the frequent itemsets, it is possible to generate association rules of the form `A` $\rightarrow$ `B` relating the subset of properties `A` with the subset of properties `B`, and that can be interpreted as follows: the individuals including `A` include also `B` with a certain support and a certain confidence. The number of rules that can be extracted is very large, and there is a need of pruning the sets of extracted rules for interpretation (most of the time, the analyst is in charge of interpreting the results of the rule extraction process). This is why a number of measures has been set on, mainly based on probability theory e.g. the so-called *statistical implication* between `A` and `B`, that can be read as `when A almost B`. Moreover, a probability distribution of rules can also be studied. Actually, this kind of work is under investigation, and preliminary results can be found in [30][29][5][14][9]. Among others, there is a study on the separation of rules mainly due to hazard and the other rules linked with the data domain.

*3.1.1.2. Association rule extraction in a biological database.*

Relying on the same principles, a research work is currently under investigation in the domain of biology, for searching for correlations between biological factors and diseases in a given population of individuals [26]. A first study has been carried out on a real-world individual database, namely the STANISLAS COHORT, consisting in 1006 caucasian families supposed to be healthy and from homogeneous origin, recruited for medical examination at the "Centre de Médecine Préventive de Vandœuvre-Lès-Nancy" since 1993 (thus allowing longitudinal studies). The cohort is explored for searching for genotypes and intermediate phenotypes of cardiovascular diseases (multifactorial pathologies resulting from gene-gene and gene-environment interactions). In other words, there is a need for extracting implicit or new potential risk factors of cardiovascular diseases, within an always growing volume of data (mainly due to the development of technologies such as multiplex technologies or microarrays). In the STANISLAS COHORT, information hold on environmental data, clinical data, biological data, genetic data, etc. Experiments are currently undertaken and have given first results that still must be improved. At present, the results are in accordance with already known statistical results, and may give new research insights for further investigations.

With respect to statistical work (more generally used in such a context), the general idea of this research work is the following: mining the cohort for extracting itemsets that are in turn considered as hypotheses for being validated by statistical tests. In the next future, data are going to be analyzed with respect to their temporal dimension[1].

Finally, let us mention another experiment based on itemsets and association rule extraction. This research work has been carried out in the field on organic chemistry: databases of chemical reactions have been mined in order to extract the so-called "synthesis methods", that can be seen as kinds of meta-rules for guiding the chemist when he is searching for a new synthesis plan [12][21].

### *3.1.2. Hidden Markov Models for Data Mining*

We present in this section the research work based on higher-order stochastic models – namely second-order hidden Markov models (HMM2) – that aims at discovering spatial and temporal dependencies in databases. HMM2 are able to map sequences of data into a Markov chain in which the transitions between the states depend on the *two* preceding states. HMM2 are based on probability and statistics theories. Their main advantage is the existence of a non-supervised training algorithm (the EM algorithm), that allows the estimation of the parameters of the Markovian model from a corpus of observations and an initial model. The resulting Markovian model is able to segment each sequence of data into stationary and transient parts.

[1] A first research work in the medical domain has already been carried on in the team, namely by Nicolas Jay, Recherche et interprétation de motifs séquentiels fréquents dans une base de données médicales, Mémoire de DEA, Université Henri Poincaré Nancy 1, septembre 2003.

We focused our effort on two points: (1) The elaboration of a process for mining spatial and temporal dependencies for knowledge acquisition. This process involves a non-supervised classification of data. (2) The specification of adequate visualization tools giving a synthetic view of the classification results to the experts, who have to interpret the classes and/or specify new experiments.

Below, we describe three main applications, developed using a generic data-mining system for spatio-temporal data, based on HMM2, and named CARROTAGE (the CARROTAGE system is a free software with a GPL license). The two first experiments are in concern with knowledge discovery in the domain of agronomy (done in collaboration with agronomists), one for a better understanding of the farmer work, and the other for a better management and prediction of water needs [11][25]. The third experiment holds on gene segmentation and interpretation in the domain of bioinformatics.

*3.1.2.1. Crop rotations in the Seine river watershed.*

For thirty or forty years, the hydrosystem of the Seine river has been gradually degraded, regarding water quality and biological population, due to human activities (domestic, industrial, agricultural activities). The nitrate contamination of cave and surface waters is mainly caused by the evolution of agricultural activities, and related to their nature and to their organization inside the river watershed. The objective of the interdisciplinary research program PIREN-Seine (*Programme Interdisciplinaire de Recherche en ENvironnement sur la Seine*) [33] is to develop a tool for predicting the water quality in the Seine river watershed, based on assumptions on agricultural changes. In this research work, members of the INRA ("Institut de la recherche en agronomie") team in Mirecourt analyze the agricultural activities in the watershed, with respect to their dynamics and their spatial organizations. They particularly focus on the crop (temporal) rotations that may explain the risk of nitrate dissemination. In this application we use a French national database related to land use, named `Ter Uti`, that describes the land use at two levels: the first one is defined by a grid of aerial pictures, and the second level is defined by 6x6 matrices of sites located in the pictures. Land use (wheat, corn, forest, ...) is checked every year on each site.

HMM2 have been used for computing the average crop distribution during a given time period (here from 1992 to 1999), for viewing the main annual transitions between crops, and for listing all types of crop rotations in each region. Such an analysis has been carried out for small agricultural regions in the Seine watershed. The regions are then clustered according to their main crop rotations and their evolutions [33]. This classification has appeared to yield meaningful results for domain experts, especially for specifying simulation models of nitrate dissemination.

*3.1.2.2. Helping the interpretation of satellite images of the Midi-Pyrénées Region.*

Our approach has been used by researchers of the INRA research center in Toulouse that works on the prediction of irrigation needs in the Midi-Pyrénées region (South-West of France). Usually, irrigation needs are estimated using annual land-use maps based on satellite data. This method is not always satisfying since data are not necessarily available at the moment the prediction has to be done: the satellite images are obtained at the beginning of the cropping season (in spring) and this does not allow to recognize all the crops of a given region. Whenever the crop rotations are known, they can be used for recognizing the crops themselves, based on the land-use map of the year before. Knowing the crop in a plot at year $n-1$, the potential crops in the same plot at year $n$ can be inferred, and their number reduced using the available satellite images.

We perform a spatial clustering by defining a fractal scanning of the images with the help of a Hilbert-Peano curve, that introduces a total order on the sites, preserving the relation of neighborhood between the sites [25]. Spatial and temporal classifications are simultaneously processed by means of two HMM2 measuring the *a posteriori* probabilities to map a temporal sequence of images onto a set of hidden states. In this case, we have adopted a Bayesian point of view for measuring the uncertainty of a classification by a probability.

The models built for the spatial segmentation and the temporal segmentation have been used in two complementary ways. The first one allows the definition of homogeneous and stable areas regarding the crop rotations, while the second one allows a more specific study of each area. This method has appeared to be very interesting, although the scale of `Ter Uti` data may be insufficient for precisely recognizing the potential crops in an irrigation basin.

*3.1.2.3. An application in bioinformatics.*

A long term data mining research project in bioinformatics is carried out in collaboration with the Laboratory of Genetics of the "Université Henri Poincaré Nancy 1" (thesis of Sébastien Hergalant). The biological material is the soil-dwelling filamentous bacteria belonging to the genus *Streptomyces*, that is the largest source of antibiotics amongst microorganisms. In particular, the *Streptomyces coelicolor* chromosome (8,7M bases) is entirely sequenced and annotated. We are interested in detecting "genome heterogeneity islands", and inter-sequences dependencies by means of Hidden Markov Models, without prior knowledge. Initially, we have focused on the understanding of horizontal transfer phenomena, but two other areas of interest in genetics are currently investigated: the detection of intragenomic DNA repeats, and the detection of promoters. So far, the mining of the DNA sequences is performed under the supervision of Sébastien Hergalant, who is specifying a toolbox for computing, displaying and analyzing the *a posteriori* probabilities of the HMM2 hidden states, in various adapted graphical standard environments. It becomes then possible to correlate the output signal of HMM2 with the biological annotations of long segmented DNA sequences (more than 50 000 nucleotides).

We are investigating the pertinence of various methods for segmenting and classifying the chromosome. At present, we are using a *Fast Fourier Transform* to extract the periodic components, and to have another point of view on the stationary/transient behaviors of the process. We also try to determine an appropriate distance, e.g. the Mahalanobis distance, between a codon in a gene, and the class of codons specific to the species.

Moreover, we have elaborated several learning methods for specific analyses:

- The horizontal transfer understanding. Markovian models with respect to "species specific homogeneities" have been designed and coupled with the transform filters described above. Their behavior generates regions with different statistical properties allowing the user to separate "foreign DNA regions" with the proper DNA regions of the studied species. In *S. coelicolor*, the regions with statistical consensus have been extracted and correlated with potential events of horizontal transfer.
- The intragenomic reiteration detection. In previous work, we have developed Markov models for processing DNA sequences without prior knowledge of their content, searching for DNA subsequences that show strong homology. The graphical signal was used to detect different forms of repeats, e.g. tandem, direct or reversed, with various length and localization in the *S. coelicolor* genome. The variability captured during the learning step by the EM algorithm allows the detection of degenerated repeats. This interesting feature is not present in general in the string algorithms searching for exact matches.
- The detection of promoters. We have studied the detection of short DNA patterns in the chromosome, appearing frequently (and abnormally) at non-random locations. Identifying the underlying functions linked to these short DNA patterns may be of great interest in the decryption of genome organization and regulatory functions. First DNA patterns identified with the present method have been identified as promoters [20] (promoters are regulatory sequences essential for the cell life, constituting target sites for specific proteins and implied in the gene expression mechanisms). We are currently working on the automatic extraction of these patterns. The resulting set of classes, defined by the underlying DNA sequences, could characterize new promoters and thus define new sets of co-regulated genes.

### *3.1.3. Text Mining*

The goal of a text mining process is to find in a large set of texts new and useful knowledge units. If text mining relies on the principles of KDD, it shows specific characteristics due to the fact that texts are written in natural language. The information in a text is expressed in an informal way, following linguistic rules, making the discovery process more complex. To avoid information dispersion, a text mining process has to take into account paraphrases, ambiguities, specialized vocabulary, and terminology. Moreover, the interpretation of a text relies on a common knowledge shared by the authors and the presumed readers. Part of this background knowledge is expressed in the texts and should not be extracted by the mining process as new knowledge. Part of it is not expressed but may be useful to relate notions in a text that, at a first glance, seem to be disconnected.

To carry out researches on text mining, the Orpailleur team is interested in linguistics resources: actual texts in actual contexts with robust tools, contrasting other works dealing with specific phenomena in the language. The language is considered as the way to access information, and not an object to be studied *per se*. Thus, the text mining process is considered as involved in a loop, where the process can be used to improve linguistic resources, and where linguistic resources can be used, in turn, to improve the information extraction process for guiding a kind of model-based text mining process (the model makes reference to the available knowledge on the domain of texts).

*3.1.3.1. The process of text mining.*

The expression "text mining" is widely used in the literature to name very different experiments, starting from information retrieval or question answering to ontology building or technological watch. From our point of view, we define text mining as a specific process of knowledge discovery in databases. An analyst, expert in a scientific or technical domain, is in charge of guiding the mining process of a large amount of texts. The very first steps are dedicated to linguistic knowledge acquisition: lexicon, terminology, markers of semantic relations, discourse markers, specific syntactic or semantic structures...The following steps aim at identifying or structuring the background knowledge for extracting new knowledge units.

*3.1.3.2. Extracting semantic relations using case-based reasoning.*

Text mining at the term level is at present relatively well known. However, due to the weakness in the semantic description of a term, to the lack of semantic relations between terms, and to the lack of syntactic dependencies between terms, the analyst has a high degree of freedom, and thus is not efficiently guided in the analysis of the results of the text mining process. He has to refer to the texts to avoid errors due to misunderstanding in the interpretation of the results. For improving the text mining process, there is a need to represent texts by their content, terms in the text content, and relations between terms. These complex structures make then possible to represent "events", such as "the mutation of the gene GyrA", or "the resistance of a bacterium to quinolone", and to calculate correlations between these events (rather than between "simple" terms).

In this context, the research work presented in the thesis [2] aims at analyzing semantic relations from scientific or technical texts. It is assumed that there are syntactic and semantic regularities in the texts. A case-based reasoning approach is used to identify and to explain the semantic relations and their arguments in a sentence, with both ontological and linguistic information.

*3.1.3.3. Extraction of association rules from texts.*

We have performed a number of experiments on the extraction of association rules from texts, in the context of scientific and technological watch. One major problem is that the extraction process generates a very large number of rules. Then, selecting an "interesting rule" involving new knowledge units is a rather complex task for an analyst. Thus, the extraction process is considered from two points of view: ranking and evaluation of the rules.

- Ranking association rules: texts are indexed using a merge of several thesauri (we are currently working on medical texts). The support and the confidence are the two main indices that are associated to the rules, and they play a major role in the reduction of the calculation time and of the number of the generated rules. Research work taking advantage of these indices are presented in [5][17][16][15], together with other indices, e.g. the interest, the dependency, the novelty. The rules, ranked according to these indices, have been presented to the analyst. It turns out that some combinations of these indices allow the analyst to identify complex semantic relations between terms or synonyms. However, most of the extracted rules are in accordance with the present domain knowledge, and do not provide any new knowledge unit. Thus, the next objective is to be able to extract association rules providing new effective knowledge units.

- Introducing a knowledge model. We are currently working on the explicit exploitation of knowledge for pruning and ranking the association rules, with respect to a "degree of novelty" that a rule may include, compared to the existing background knowledge. In this context, we are studying Bayesian networks on the one hand, and on the combination of association rules and meta-rules on the other hand (meta-rules based on background knowledge).

## 3.2. Knowledge Representation and Knowledge Systems

**Key words:** *knowledge representation*, *object-based representation systems*, *description logics*, *classification-based reasoning*, *case-based reasoning*, *lattice-based classification*, *qualitative spatial reasoning*.

**Participants:** Mathieu d'Aquin, Sébastien Brachais, Florence Le Ber, Jean Lieber, Jean-Luc Metzger, Amedeo Napoli, Laszlo Szathmary.

*Glossary*

**knowledge representation** is a process for representing knowledge within knowledge representation formalisms, giving knowledge units a syntax and a semantics.

### *3.2.1. Classification-based Systems and Reasoning*

A knowledge system relies on a knowledge base and a reasoning module for problem-solving and knowledge management in a given domain. Knowledge units are represented within a knowledge representation formalism where they have a syntax and an associated semantics. Inference can be drawn from already known knowledge units (or facts) for deriving new facts, that are useful for solving the current problem. Moreover, the units extracted from data by data mining procedures have also to be represented within a knowledge representation formalism to be taken into account in the framework of a knowledge system.

In the team Orpailleur, two kinds of formalisms are particularly studied, namely object-based knowledge representation (OBKR) systems, and description logic (DL) systems, together with classification-based reasoning. The function of such a system is to represent knowledge units within concepts (also called classes), attributes (that can be properties of concepts, or relations, also called roles in DL) and individuals. The hierarchical organization of concepts relies on a subsumption relation that is a partial ordering. Such a system provides a representation and an organization of knowledge units, and a number of inference services. Among the inference services, let us mention concept and individual classification. The first operation is used to insert a concept at the right place in the concept hierarchy (searching for its most specific subsumers and its most general subsumees). The second operation is used for recognizing the concepts an individual may be instance of. In both cases, subsumption and classification are the main operations: this is why these systems are denoted here by "classification-based systems".

Case-base reasoning (CBR) relies on three main operations: retrieval, adaptation, and memorization. A source case `(srce,Sol(srce))` lies in a case base, and can be seen as a problem statement `srce` together with its solution `Sol(srce)`. Then, given a new target problem, say `tgt`, retrieval consists in the search for a memorized case whose problem statement `srce` is similar to the target problem `tgt`. Then, when `srce` exists, its solution `Sol(srce)` is adapted to fulfill the constraints attached to `tgt`. When there is enough interest, the new pair `(tgt,Sol(tgt))` can be memorized as a new case for further problem solving. In the context of concept hierarchy, case-base reasoning can be seen as a natural extension of classification-based reasoning. Retrieval and adaptation may be based both on classification and on searching for paths in the concept hierarchy. Moreover, a number of studies within the Orpailleur team has been carried out on CBR, especially on "adaptation-guided retrieval", that consists in searching for a source case whose solution will be adaptable for the target problem, giving a kind of guarantee regarding the building of the solution of the source case.

In parallel with knowledge representation, knowledge management is oriented toward the management of what could be called the "cycle" of knowledge, including acquisition, memorization, retrieval, maintenance, dissemination (or exchange) of knowledge. There is also a need for coupling knowledge with data, with

respect to representation and management. This means in particular that, besides knowledge extraction from databases, there are some other needs such as e.g. information retrieval, for helping a reasoning process. Thus, there must exist channels between the knowledge representation universes and the document (or data universe). This is particularly important in the framework of the semantic Web (introduced in the next section). These kinds of channels can rely on a coupling of a knowledge representation formalism and a description language for documents, such as XML. In this way, knowledge representation units can be associated to document descriptions units: the management of documents (or data) is performed within the document description language, and reasoning is performed within the knowledge representation formalism. Moreover, additional coupling between information retrieval and knowledge extraction can be set on. This view of knowledge management is of primary importance, mainly because of the Web, and the always growing need of disseminating information and knowledge.

### 3.2.2. Spatial Knowledge Representation and Spatial Reasoning

In this framework, we work on two major themes, the representation of spatial structures in knowledge-based systems, and the design of reasoning models on these structures e.g. hierarchical classification, CBR. This research work is applied to answer agronomical questions regarding the recognition and the analysis of farmland spatial structures.

*3.2.2.1. Lattice-based classification of spatial relations.*

This work was initiated during the thesis of Ludmila Mangelinck (1995–98) in collaboration with the INRA BIA laboratory in Nancy. It has been carried out in the context of the design of a knowledge-based system for agricultural landscape analysis [4]. The main objective of this system, called LOLA, is to recognize *landscape models* on land-use maps extracted from satellite images. Landscape models are abstract models describing agricultural spatial structures as sets of spatial entities and qualitative spatial relations between these entities. They are used to classify *zones* extracted from the maps. A zone is a collection of raster regions, i.e. connected sets of pixels with the same label denoting the land-use category, e.g. crops, meadows, forest, buildings, etc. From an implementation point of view, an object-based knowledge representation system, equipped with a classification process, has been used. In this framework, the exploitation of land-use maps for landscape analysis may be considered as an *instance classification* problem, where landscape models correspond to classes, while zones correspond to instances that have to be classified according to landscape model classes.

Following these needs, we have designed a hierarchical representation of topological relations based on a *Galois lattice* –or *concept lattice structure*– relying on the Galois lattice theory [7][22]. A Galois lattice is a multi-faceted tool for designing hierarchies of concepts: it allows the construction of a hierarchical structure both for representing knowledge and for reasoning. In a concept lattice structure, a concept may be defined by an *extension*, i.e. the set of individuals being instances of the concept, and by an *intension*, i.e. the set of properties shared by all individuals. In our framework, the extension of concepts corresponds to topological relations between regions of an image, and the intension of concepts corresponds to properties computed on that image regions (*computational operations*). Thus, a concept lattice structure emphasizes the correspondence between qualitative models, e.g. topological relations, and quantitative data, e.g. vector or raster data.

Currently, this work is continuing with a deeper study of Galois lattices for linking qualitative topological relations and computational operations on numerical (raster or vector) data. In particular, we focus on the comparison of lattices built on different sets of relations or computational operations.

*3.2.2.2. CBR on spatial organization graphs.*

This work has been undertaken in the framework of J.-L. Metzger thesis in collaboration with INRA SAD and ENGREF. The objective is to develop a knowledge-based system, called ROSA, for comparing and analyzing farm spatial structures. The reasoning in the ROSA system follows the principles of case-based reasoning (CBR). In our research work, CBR relies on the agronomical assumption that there exists a strong relation between the spatial and the functional organizations of farms, and thus, that similar spatial organizations correspond to similar functional organizations. According to this assumption, and given a set of previously

studied farm cases, the ROSA system has to help agronomists to analyze new problems holding on land use and land management in farms [32][28][27][8][35].

- In a first step of the present work, a model of the domain knowledge has been proposed, in accordance with agronomists. This model is based on *spatial organization graphs*, or SOG, with labeled vertices and edges. Relying on these spatial organization graphs, *spatio-functional cases* for farms have been designed: they mainly consist of a description of the land use and an associated explanation linking spatial and functional organizations.
- In a second step, the SOGs and the cases have been represented within a knowledge representation formalism, namely the description logic (DL) system RACER. In this way, reasoning in the ROSA system relies on an original combination of hierarchical classification, CBR, and qualitative spatial reasoning. In addition, spatial inference rules are used for building *similarity paths* between SOGs. These paths are used in the CBR mechanism for comparing problems and adapting the solution from a source case to a new target problem.

During this year we have focused on the representation and the manipulation of graphs in the DL framework, and on the definition of transformation rules for building similarity paths between graphs. The knowledge acquisition and modeling issue has been undertaken with the help of researchers in socio-psychology and linguistics (CODISANT, LPI-GRC, Université Nancy 2 and GRIC UMR 5612 CNRS, Lyon). The work on qualitative spatial reasoning has been carried out for the largest part under the supervision of Florence Le Ber (since the nineties), who obtained the so-called *habilitation diploma*, for being recognized as a senior researcher [3].

### *3.2.3. Knowledge Management in Medicine: the Kasimir System*

This section presents an overview of the KASIMIR research project, whose objective is decision support and knowledge management for the treatment of cancer. This is a multidisciplinary research project in which participate researchers in computer science (Orpailleur), in ergonomics (*Laboratoire d'ergonomie du* CNAM, Paris), experts in oncology (*Centre Alexis Vautrin* or CAV, Vandœuvre-lès-Nancy) and Oncolor (an association of physicians from Lorraine involved in oncology).

For a cancer localization, e.g. the breast, the treatment is based on a protocol similar to a medical guideline. This protocol is built according to evidence-based medicine principles. For most cases (about 70%), a straightforward application of the protocol is sufficient, and provides a solution, i.e. a treatment, that can be directly reused.

A case out of the 30% remaining cases is "out of protocol", meaning that either the protocol does not provide a treatment for this case, or the proposed solution raises difficulties, e.g. contraindication, treatment impossibility, etc. For such an out of protocol case, oncologists try to *adapt* the protocol (actually they discuss such a case during meetings of the so-called "breast therapeutic decision committee", including experts of all domains in breast oncology, i.e. chemotherapy, radiotherapy and surgery). In addition, protocol adaptations are currently studied from the ergonomics and computer science viewpoints. These adaptations can be used to propose *evolutions* of the protocol based on a confrontation with actual cases. The idea is then to make suggestions of protocol evolutions based on frequently performed adaptations.

#### *3.2.3.1. Knowledge representation for decision support tools.*

The KASIMIR system is currently under development and implements (at the moment) an object-based representation formalism associated with an inference engine based on hierarchical classification, and a decision support module in oncology. A number of knowledge bases corresponding to specific cancers (decision protocols) are under development. Moreover, the inference engine has been extended for taking into account a fuzzy representation of concepts and fuzzy hierarchical classification [24]: the system tries to detect and propose more than one treatment for "borderline cases".

*3.2.3.2. Going further: a semantic portal for oncology.*

The current research in computer science on the KASIMIR system follow two main directions: protocol adaptation, and the embedding of the KASIMIR system within a semantic portal for oncology [13], i.e., a Web server relying on the principles and technologies of the semantic Web for providing an intelligent access to knowledge and services in oncology.

One of the main issues of the semantic Web relies on interoperability for knowledge and applications. Thus, building a semantic portal implies a standardization of knowledge and software components of the KASIMIR system. For the knowledge bases, standardization relies on a sharable domain model, and leads to the definition of general ontologies in oncology. This kind of knowledge base re-engineering requires to replace the *ad hoc* knowledge representation formalism of KASIMIR with a knowledge representation formalism adapted to the semantic Web, such as OWL.

This work also implies a new software architecture, for the KASIMIR reasoner and the editing, visualization and maintenance modules. This architecture must take into account constraints related to the distributed and dynamic environment of the semantic Web. Moreover, since the KASIMIR inference engine is based on subsumption, a study on the integration of an extended inference engine taking into account inferences based on CBR and the integration within the semantic Web has to be carried out (first elements are given in [18]).

## 3.3. The Semantic Web

**Key words:** *Semantic Web*, *knowledge-based information access and retrieval*, *information retrieval guided by data mining*, *bioinformatics*.

**Participants:** Rim Al Hulou, Mathieu d'Aquin, Marie-Dominique Devignes, Huaizhong Kou, Amedeo Napoli, Emmanuel Nauer, Malika Smail, Laszlo Szathmary, Yannick Toussaint.

*Glossary*

**Semantic Web** is a framework for building knowledge-based systems for manipulating documents on the Web by their contents and their semantics.

### *3.3.1. Introducing Research on Semantic Web*

Today people try to take advantage of the Web by searching for information (navigation, exploration) and by querying documents using search engines (information retrieval). Then people try to analyze the obtained results, and that may be a very difficult and tedious task. To-morrow, the Web will be "semantic" in the sense that people will search for information with the help of machines, that will be in charge of posing questions, searching for answers, and interpreting the answers. The Web will become a space for exchange of information between machines, allowing an "intelligent" access and management of information. However, a machine will be able to read, understand, and manipulate information on the Web only if the knowledge necessary for achieving those tasks is available. This is why ontologies are of main importance with respect to the task of building a semantic Web. Moreover, there is a need for languages for annotating documents, i.e. describing the content of documents, and giving a semantic to this content. Knowledge representation languages are good candidates for achieving this kind of task: they have a syntax with an associated semantics, and thus they can be used in information retrieval, query answering, and reasoning processes.

The semantic Web has gained a great interest in the research work of the Orpailleur team [6]. Indeed, it constitutes a good platform for experimenting a number of ideas on knowledge representation, reasoning, knowledge management, and knowledge discovery (and especially text mining) as well. Among others, we are interested in the content-based manipulation of textual documents using annotation, ontologies and a knowledge representation language. The idea here is to build an XML-based "bridge" between documents and object-based knowledge units associated to the domain of documents. The annotations attached to documents and the queries are built with the help of a domain ontology, and have an XML syntax. Moreover, annotations, elements of the ontology, and queries, have corresponding elements within the knowledge representation language, providing a semantics for these elements. Then, the manipulation of annotations, e.g. information

retrieval, query answering, is in charge of the reasoning module associated to the knowledge representation formalism. In our case, this formalism is based on description logics [1][10].

### *3.3.2. Intelligent Access to Information*

The availability and retrieval of information is of main importance in scientific and technical domains, for e.g. research and watch purposes. Nowadays, there is a huge quantity of data available, and this requires to implement adapted tools for exploiting and taking advantage of the data. One research work carried out within the Orpailleur team holds on the definition and implementation of an environment allowing an intelligent access to information, by combining information retrieval, hypertext navigation, and data-mining. This environment can be used for document retrieval on the Web, bibliographical search or domain analysis.

In this framework, a hypertext data-mining system for searching bibliographical data is currently under development, based on the following assumptions. The use of hypertext functionalities has to favor an exploration-based access to data. The use of data-mining functionalities, i.e. classification and association rule extraction, has to favor information retrieval and data domain understanding. The underlying idea here is that data-mining and information retrieval are complementary for accessing and analyzing data. Data-mining allows the guiding of information retrieval by taking advantage of the knowledge units extracted from the data. Conversely, information retrieval allows the guiding of the data-mining process by making available information on data that can be used for example for pruning a set of extracted rules or for providing a focus for a classification process.

In a more concrete way, in the domain of bibliographical data, the knowledge units extracted from bibliographical data can be used to guide a keyword-based document retrieval on the Web. Information retrieval can be enhanced when semantic relationships between textual data are taken into account, by making precise the context knowledge of a query, and by filtering documents on the basis of the similarity of their content, with respect to background knowledge. These are the two main ways on which relies our research work for identifying relevant documents on the Web, and in bibliographical databases.

### *3.3.3. Intelligent Access to Genomic Sources on the Web in Bioinformatics*

Web sources are widely used in bioinformatics. Scientists are getting more and more concerned with the problem of exploiting at the best the whole mass of biological information stored in the numerous, heterogeneous, and public data sources. Many functionalities are proposed by the data servers: access to the data, execution of programs such as sequence comparison and analysis tools. Some integrated systems offer a unified access to heterogeneous sources and resources (Entrez, SRS, PBIL). Mediation architectures allow in certain case-studies automatic processing of complex queries (TAMBIS). One approach studied within the Orpailleur team is based on the distinction between two types of problems generated by a complex question: firstly the identification and the characterization of relevant sources in terms of availability and query capabilities, secondly, collecting and integrating data from the selected sources [19].

Answering a complex biological question can be considered as an execution of a succession of steps aimed at querying given sources or resources. Such steps are currently under investigation, in a model involving the following functionalities: (1) selecting relevant sources (for a given step), (2) ranking sources (according to desired criteria), (3) query construction and submission, (4) extraction of useful data from returned documents, (5) iteration of steps (3) and (4) over the sources. Finally output data has to be integrated to constitute a global answer to the initial question. Chaining of the steps allows, when necessary, output data from one step to become input data for the next step. In practice, functionalities (1) and (2), as well as the description of the chaining of the steps, constitute the definition of a scenario. This process clearly involves the user cooperation, and the optimization of the process relies on the particular user knowledge. On the contrary, functions (3) to (5) as well as integration of the data into a structured document, represent the execution of the scenario, i.e. a data retrieval process that is easier to model and can lead to the development of an application. The automation of this group of functionalities presents several advantages: time-saving when answers are required for multiple entries, easy update of the answers, easier exploitation of the answers because of the structured storing format.

The design of a generic data retrieval process, within the so-called Xcollect project, is based on two models: a generic scenario model and a generic session-data model. The generic scenario model appears as a succession of steps, where the following information is specified: source name, type, and location; input name, type, and value (including parameters for appropriate query construction); output name and type; and parameters (such as regular expressions) necessary to extract the useful data from the returned document. An XML DTD has been used to represent this model. The generic session-data model describes the steps of the scenario with their respective input and output data. There is no satisfying standard solution for storing the retrieved data, and therefore, a simple generic DTD has been written on the basis of the scenario DTD. Depending on the desired usage of the data, appropriate XSL transformations allow a straightforward conversion of this generic representation of the retrieved data into various biological meaningful structures.

Moreover, the following elements are also currently under investigation: identification and characterization of relevant sources for designing query answering scenarios, management of multiple answers to individual steps of a scenario, formalization and exploitation of knowledge on sources and their contents for selecting pertinent sources and interpreting the collected data.

# 5. Software

## 5.1. Stochastic Systems for Knowledge Discovery

**Participants:** Sébastien Hergalant, Florence Le Ber, Jean-François Mari [contact person].

**Key words:** *Hidden Markov models*, *stochastic process*.

### *5.1.1. Carrotage*

One aspect of data-mining is to provide a synthetic representation of data that a domain analyst can interpret. The purpose of the CARROTAGE system is to build a partition –called the hidden partition– in which the inherent noise of the data is withdrawn as much as possible. Then spatio-temporal data are explored for extracting homogeneous classes both in temporal and spatial dimensions, giving also a clear view of the transitions between the classes.

CARROTAGE is a free software[2] under a GPL license, that takes as input an array of discrete data (the rows represent the spatial sites and the columns the time slots), and that builds a partition with the associated *a posteriori* probability. This probability may be plotted as a function of time, and is a meaningful feature for the analyst searching for stationary and transient behaviors of data. This software is currently used by INRA researchers interested in mining the successions of land use processes, in order to build models to simulate the nitrate contamination of cave and surface waters.

### *5.1.2. genExp*

GenExp [34] is an experimental software that has been developed in the framework of the "OGM Impact project", in collaboration with biologists of ESE UPRESA 8079 CNRS, Paris Sud. The objective of the system is to simulate agricultural landscape for studying the dissemination of vegetal transgenes. The system is based on the stochastic system CARROTAGE, and on computational geometry.

## 5.2. Software for Text Mining

**Participants:** Hacène Cherfi, Dietmar Janetzko, Yannick Toussaint [contact person].

**Key words:** *knowledge discovery from databases*, *text mining*, *frequent pattern extraction*, *association rule extraction*.

We are currently developing a system named RAR, standing for "Ranking Association Rules", that allows a the navigation through a large set of association rules (such as those obtained within a text mining experiment). This system is based on a user-friendly interface, and it can be easily used by non-computer scientists, e.g.

[2] http://www.loria.fr/~jfmari/App/

analysts experts in the domain of the data analyzed. The association rules are supposed to be extracted by a mining algorithm –the *Close* algorithm in our case, for extracting frequent itemsets and association rules– and should be encoded in a predefined XML format. The RAR system then stores the rules in a database, and propose eight different statistical measures, e.g. support, confidence,interest, conviction, dependence...for sorting the analyzed set of rules. It is also possible for the analyst to focus on smaller sets of rules satisfying a given set of constraints. These constraints may be expressed as operations on the values of the statistical measures, and on the content of the left/right hand side of a rule.

## 5.3. Software for the KDD Cup

**Participants:** Martine Cadot [contact person], Joseph Di Martino [Parole, LORIA].

**Key words:** *knowledge discovery from databases*, *text mining*, *frequent pattern extraction*, *association rule extraction*.

The KDD CUP 2003 is a knowledge discovery and data-mining competition held within the Ninth Annual ACM SIGKDD Conference (http://www.acm.org/sigkdd/kdd2003/). The competition of 2003 has focused on problems of network mining and usage logs analysis. Complex networks have emerged as a central theme in data-mining applications, appearing in numerous domains. In parallel, the difficulty for designing complete and accurate representations of large networks is still a serious obstacle. This year, KDD CUP was based on a large collection of research papers, that provided a framework for testing general network and usage mining techniques, to be explored through four tasks. Each task defined a separate competition, with its own specific goals. We participated in the second task, consisting in re-creating the citation graph of about 35 000 papers (1.8 GB of data), with, for each paper P in the collection, a list of cited papers $(P_1, \ldots, P_k)$ (note that P may possibly cite papers that are not in the collection). This second task can be considered as a data-cleaning task, that is one of the the most expensive tasks within the knowledge discovery process. A specific software has been developed for that participation in the KDD CUP by Martine Cadot and Joseph di Martino, who have obtained the third place in this international competition (see http://www.cs.cornell.edu/projects/kddcup/results.html for the official results).

## 5.4. Software for Spatial Reasoning

**Participants:** Florence Le Ber, Jean-Luc Metzger [contact person].

**Key words:** *qualitative spatial reasoning*, *tpological relations*, *land organization*.

Rosa, for "Reasoning on Organization of Space in Agriculture", is a system developed in collaboration with agronomists, whose objective is to record and maintain an agronomical knowledge base on farms, and to solve problems in agronomy, based on this knowledge base. Two kinds of knowledge elements are considered: domain knowledge, and knowledge on spatial organization and functioning of specific farms. The domain knowledge is described by a hierarchy of spatial concepts and relations (spatial occupation and relations). The spatial organization of farms is described by the so-called "space organization graphs" (SOGs) that link spatial entities with spatial relations. A vertex of a SOG (either spatial entity or relation) is labeled and linked to a concept of the domain knowledge hierarchy. The functioning of farms is described within "explanations" attached to parts of SOGs. An explanation holds on a particular function in the farm organization and functioning. The association of a particular SOG with an explanation composes a case, to be used within a case-based reasoning process. The Rosa system is under development, and is implemented within the RACER description logic system.

## 5.5. The Kasimir System

**Participants:** Mathieu d'Aquin, Christophe Bouthier [ECOO research project], Sébastien Brachais, Jean Lieber [contact person], Amedeo Napoli.

**Key words:** *object-based representation system*, *classification-based reasoning*, *case-based reasoning*.

The objective of the KASIMIR system is decision support and knowledge management for the treatment of cancer. A number of tools have been developed within the KASIMIR system: mainly modules for the editing of protocols, visualization, and maintenance [31]. The ontology editor PROTÉGÉ has been customized for editing the KASIMIR protocols, and it has been connected with the KASIMIR inference engine. The use of the PROTÉGÉ editor involves a simplification of the protocol editing, and the detection of errors during the editing, thanks to the inference engine.

Two visualization modules have been integrated in PROTÉGÉ, allowing the display of the KASIMIR hierarchy of concepts from the protocol being edited: PALÉTUVIER and HYPERTREE (currently developed in the ECOO team at LORIA). The combined use of these two visualization modules, and of the classical tree widget of PROTÉGÉ, provides several useful features for hierarchy visualization, navigation, and global or focused views.

Finally, a maintenance module has been developed and integrated into PROTÉGÉ, that compares two versions of a protocol in order to separate changed and unchanged elements. This module can be used in particular during an editing session, to visualize the modifications since the beginning of the session.

## 5.6. Intelligent Access to Information for the Semantic Web

**Participants:** Rim Al Hulou, Amedeo Napoli, Emmanuel Nauer [contact person].

**Key words:** *information retrieval*, *information access*, *semantic Web*.

Two majors systems are under development. A first one, called "IntoBib", is a generic system designed for the exploitation of bibliographical data. The IntoBib system is based on a toolbox providing a number of modules, among which, hypertext navigation, retrieval of bibliographical references, extraction of correlation between references, search for equivalent references (duplicates), conceptual clustering of similar references (with respect to a given point of view), normalization of fields e.g. author name, keywords.

A second system is under development for manipulating textual documents by their content. Documents are annotated using a domain ontology (experiments have been carried out with biological documents). The annotations and the ontology are implemented within an XML-based description language. Every element with an XML-based description, i.e. a concept of the ontology or an annotation, has a corresponding knowledge structure within the associated knowledge representation system, actually based on the the RACER description logic system. In this framework, syntactic aspects are taken into account within the XML level, and the semantic aspects are taken into account within the knowledge representation level. The system is aimed at content-based document retrieval, and query management, e.g. query answering, query classification, detection of similarities on the content of documents [1][10].

## 5.7. DefineCrawler: a Generic Crawler for the Semantic Web

**Participants:** Amedeo Napoli, Emmanuel Nauer [contact person].

**Key words:** *information retrieval*, *semantic Web*, *Web crawling*.

The "DefineCrawler" system can be seen as an information retrieval "meta-system", in the sense that it can be parameterized for satisfying different information retrieval tasks. The DefineCrawler system is based, on the one hand on a thorough study of the capabilities provided by classical information retrieval architectures, and on the other hand on the search engines available on the Web. A number of parameters have been retained, to be adjusted within an XML file for implementing and controlling different information retrieval system behaviors.

- Initialization parameters (`Start`) include the maximum depth of the crawl (`Depth`), a set of starting points for navigation (`URL`, possibly making reference to the URL of a search engine), the directory where have to be stored the data collected by the crawler (`Directory`), the number of parallel processes crawling the Web (`NbThread`), a halting condition (`Stop`) making possible the specification of a maximal crawling time, and thus ensuring a termination of the information retrieval process.

- Validation parameters (`Validation`) include a set of conditions (connected by boolean operators) that must be satisfied by the documents, for eliminating documents without interest with respect to the query, e.g. documents that do not satisfy some criteria, that are not in a fixed language...
- Evaluation parameters within which additional conditions can be set, in order to evaluate the returned documents. The evaluation and validation conditions can be combined to calculate a score for a returned document. This score is then used to rank the returned documents.

Every validation and evaluation condition is defined by an external instruction, allowing the use of various commands or tools, e.g. for checking the presence of an element, for counting the occurrences of some elements, for calculating a similarity between documents...

## 5.8. Collecting and Integrating Genomic Mapping Data: Xmap

**Participants:** Charles Auffray [Genexpress, CNRS-FRE2571, Villejuif], Claude Chelala [Genexpress, CNRS-FRE2571, Villejuif], Marie-Dominique Devignes [contact person], Nicole Fayein [Genexpress, CNRS-FRE2571, Villejuif], Malika Smail.

**Key words:** *knowledge discovery from databases*, *text mining*, *frequent pattern extraction*, *association rule extraction*.

The Xmap application is based on two modules, Xmap_AUTO and Xmap_DB, that have both been integrated within the bioinformatics platform server at LORIA. These modules have been used to correlate novel human transcripts all over the genome with co-localized rare not-cloned genetic diseases, in the frame of an international annotation jamboree dealing with a collection of 42 000 sequences of ADNc (International Annotation Workshop H-invitational, JBIRC, TOKYO, Japan). Discrepancies between sources of information about genetic diseases (Genatlas, LocucLink) have been detected, and their analysis is currently under investigation.

## 5.9. Xcollect for Collecting and Integrating Biological Data from the Web

**Participants:** Philippe Collet [Université Henri Poincaré, Nancy], Lionel Domenjoud [Université Henri Poincaré, Nancy], Marie-Dominique Devignes [contact person], Malika Smail.

**Key words:** *knowledge discovery from databases*, *text mining*, *frequent pattern extraction*, *association rule extraction*.

The Xcollect system is a Java application composed of a configuration module and an execution module, for managing query answering scenarios. In the configuration module, an interface allows the user to manually input the information specifying his scenario. Input data are stored into an XML document according to the generic scenario DTD. The execution module takes as input the XML scenario document, implements each step of the scenario, and returns an XML data-session document containg the retrieved data, structured according to the generic session-data DTD. A predefined style-sheet converts the XML data-session document into an HTML document to allow visualisation of the session results. Xcollect has been instantiated as the Xprom application for the PPAR project (*Peroxysome Proliferator Activated Receptors*).

Very few target genes are known so far, and a systematic search for target genes has been undertaken in order to better understand their role, and to propose new candidate genes for complex pathologies. About a hundred of DNA fragments have been experimentally isolated on the basis of their aptitude to bind PPAR. Sequences of these fragments have been processed according to a scenario that has been carefully described as the Xprom scenario: excision of irrelevant sequences, detection of putative PPAR binding elements, position on the genome, visualisation of genomic context and neighbouring genes. Since the genome sequence data and annotations are updated every three months, the Xprom automatic application revealed very useful in updating the data of interest.

Yet another scenario has been implemented in the Xcollect application: the Xfunction scenario aims at collecting information about function of human genes from the various available Web sources. Specific effort

has been made for presenting to the user the multiple answers obtained from different homologous sources, and for identifying the quality criteria that could help the user to sort the collected data.

# 8. Other Grants and Activities

## 8.1. The European Network of Excellence Knowledge Web

Here is the abstract of the Knowledge Web proposal that has become a European network of excellence (three INRIA teams are involved in Knowledge Web: ACACIA at INRIA-SOPHIA, EXMO at INRIA-RHÔNE-ALPES and Orpailleur). The current World Wide Web (WWW) is, by its function, the syntactic Web where structure of the content has been presented while the content itself is inaccessible to computers. The next generation of the Web, the Semantic Web, aims at alleviating such problem and provide specific solutions targeted to concrete problems. The Web resources will be much easier and more readily accessible by both human and computers with the added semantic information in a machine-understandable and machine-processible fashion. It will have much higher impact on eWork and eCommerce as the current version of the Web already had. Still, there is a long way to go transferring the Semantic Web from an academic adventure into a technology provided by software industry. Supporting this transition process of Ontology technology from Academia to Industry is the main and major goal of the Knowledge Web project. This main goal naturally translates into three main objectives given the nature of such a transformation:

- Industry requires immediate support in tacking up this complex and new technology. Languages and interfaces need to be standardized to reduce the effort and provide scalability to solutions. Methods and use-cases need to be provided to convince and to provide guidelines for how to work with this technology.
- Important support to industry is provided by developing high-class education in the area of Semantic Web, Web services, and Ontologies.
- Research on Ontologies and the Semantic Web has not yet reached its goals. New areas such as the combination of Semantic Web with Web services realizing intelligent Web services require serious new research efforts.

Spoken in a nutshell, it is the mission of Knowledge Web to strengthen the European software industry in one of the most important areas of current computer technology: Semantic Web enabling eWork and eCommerce. Naturally, this includes education and research efforts to ensure the durability of impact and support of industry.

## 8.2. National initiatives

### *8.2.1. ACI "Masse de données": Knowledge Discovery and Ontology Design in Astronomy*

This research project is carried out in collaboration with the CDS in Strasbourg ("Centre de données astronomiques de Strasbourg", http://cdsweb.u-strasbg.fr/, and the IRIT computer science laboratory in Toulouse. Researchers in astronomy use everyday an information network made of journal articles available under an electronic form, and a number of databases such as the SIMBAD database recording bibliographical entries and measure sets on about three millions of astronomical objects, and the catalog server VizieR recording astronomical catalogs and measure tables published in the astronomical journals. A step further must be done at present, and interested researchers should have access to the content of documents, e.g. journal articles, astronomical object catalogs, or measure tables. Researchers in astronomy have at their disposal a base of the so-called UCD for "Unified Content Descriptors", i.e. a hierarchical database that has been extracted and designed at the CDS from the content of astronomical catalogs and tables.

The research work currently carried out within the Orpailleur team holds on the study and the design of an ontology for representing astronomical objects, starting from a collection of articles (and thus involving

text mining) and extending the UCD database. This ontology will be used for a number of important and different tasks for researchers in astronomy, such as intelligent information retrieval based on the content of documents, information manipulation for matching and comparing the content of the astronomical documents. This research work can be seen as a contribution to the research works on the Semantic Web, where the purpose is to attach semantics to astronomical documents for defining an annotation method of astronomical documents, and for a knowledge-based information retrieval method in astronomical heterogeneous sources.

### *8.2.2. CNRS TCAN Project: Traitement des connaissances, apprentissage et NTIC*

A research work on Adaptation Knowledge Acquisition (AKA) for the KASIMIR system is carried out in the framework of the CNRS interdisciplinary project TCAN[3]. The objective of AKA is to provide knowledge in the form of *adaptation meta-rules*:

- Automated AKA is based on the mining of the protocols. A protocol can be seen as a set of rules `situation` $\rightarrow$ `decision`. Knowing how the decisions change when the situations change from one rule to another rule provides a specific adaptation rule. Clustering and generalizing these specific rules produce general adaptation rules, that have to be validated by experts.
- Supervised AKA is based on the analysis of adaptations performed by experts. Interviews of experts confronted to decision problems requiring adaptation have been recorded to be afterwards analyzed and modeled within adaptation rules.

Orpailleur is involved in this TCAN project, together with the "laboratoire d'ergonomie du CNAM" and the Centre Alexis Vautrin. Beyond the application framework, this work should involve progress in AKA methodology and techniques, that is an original research area in CBR (at its very first beginning, despite its importance for knowledge-intensive approaches in CBR). A preliminary study has been carried out in [23], that has highlighted several adaptation schemas that remain to be instantiated.

### *8.2.3. Projects and Collaborations in Spatio-Temporal Reasoning*

- Géomatique (CNRS–STIC): "Modélisation, comparaison et interprétation d'organisations territoriales agricoles" (in charge of F. Le Ber).
- Impact des OGM (MENRT): "Modélisation de la dispersion de transgènes à l'échelle de paysages agricole" (in charge of F. Le Ber).
- Other collaborations: INRA (Nancy-Mirecourt, Paris-Grignon, Dijon, Toulouse), Laboratoire ESE UPRESA 8079 CNRS/Paris-Sud, Équipe Codisant, LPI GRC, Université de Nancy 2, GRIC UMR 5612 CNRS Lyon, ENGREF Clermont-Ferrand.

### *8.2.4. Other Links with CNRS: "actions spécifiques" (AS) and "réseaux thématiques pluridisciplinaires" (RTP)*

- AS "Fouille de textes" (Text Mining), and AS Discovery Challenge.
- RTP 12: "Information et connaissance : découvrir et résumer".
- AS "Intégration et Interopérabilité de sources de données génomiques", attached to RTP 41 "Bioinformatique : de la séquence génomique à la fonction biologique".
- Working group[4]. "Ontologies and Metadata for Biology" depending on the IMPG action "Informatique, Mathématiques et Physique pour la Génomique".

[3] http://www.cnrs.fr/DEP/prg/TCAN.html.
[4] http://www.impg.prd.fr/Equipes/ONTOBIO.html

## 8.3. Le Contrat de Plan État-Région (CPER)

- CPER ILD-ISTC (Ingénierie des langues et du document, information scientifique, technique et culturelle). The orpailleur team is involved within the regional research project ILD-ISTC. In this context, research work is done in association with the URI team at INIST CNRS on the design of an operational text mining platform for technological watch.
- CPER Bioinformatique. Research is carried out in the frame of the Bioinformatics theme "Exploitation des génomes - Gènes candidats". Two biology laboratories are partners in this project, namely Genexpress (FRE 2571, Villejuif), and Peroxisomes proliferators (UPRES 3446, UHP, Nancy).

# 9. Dissemination

## 9.1. Scientific Animation

- The members of the Orpailleur team are involved, as members or as head persons, in a number of national research groups, mainly in "Actions spécifiques du CNRS" as mentioned above.
- The members of the Orpailleur team are involved in the organization of conferences, as members of conference program committees, as members of editorial boards, and finally in the organization of journal special issues.

## 9.2. Teaching

- The members of the Orpailleur team are involved in teaching at all levels of teaching in the universities of Nancy (especially "Université Henri Poincaré Nancy-1" and "Université de Nancy 2"; actually, it must be noticed that most of the members of the Orpailleur team are teachers).
- The members of the Orpailleur team are also involved in student supervision, again at all university levels, from under-graduate until post-graduate students.
- Finally, the members of the Orpailleur team are involved in HDR and thesis defenses, being thesis referees or thesis committee members.

# 10. Bibliography

### Doctoral dissertations and "Habilitation" theses

[1] R. ALHULOU. *Les logiques de descriptions pour le traitement intelligent des données textuelles dans le projet Escrire.* Thèse d'université, Université Henri Poincaré – Nancy 1, 2003.

[2] F. CHAKKOUR. *Identification des relations sémantiques dans les textes scientifiques et techniques en exploitant le raisonnement à partir de cas.* Ph. D. Thesis, Université Henri Poincaré, Nancy 1, France, 2003.

[3] F. LE BER. *Représentation de connaissances et raisonnements sur l'espace. Applications au domaine agronomique.* Mémoire d'Habilitation à Diriger des Recherches, UHP Nancy 1, April, 2003.

### Articles in referred journals and book chapters

[4] J. BACHACOU, F. LE BER, L. MANGELINCK. *Analyse des paysages agricoles : définition d'indicateurs pour la reconnaissance de structures spatiales sur images satellitaires.* P. MONESTIEZ, S. LARDON, B. SEGUIN,

editors, in « Organisation spatiale des activités agricoles et processus environnementaux », series Science Update, INRA Éditions, 2003.

[5] M. CADOT, A. NAPOLI. *Une optimisation de l'extraction d'un jeu de règles s'appuyant sur les caractéristiques statistiques des données.* in « Extraction de connaissances et apprentissage », number 6, volume 16, 2003, pages 631–656, Numéro spécial sur les Méthodes d'optimisation pour l'ECA, sous le direction de C. Dhaenens.

[6] J. EUZENAT, A. NAPOLI, J.-F. BAGET. *XML et les objets (objectif XML).* in « L'Objet », number 3, volume 9, 2003, pages 11–37.

[7] F. LE BER, A. NAPOLI. *Design and comparison of lattices of topological relations for spatial representation and reasoning.* in « Journal of Experimental & Theoretical Artificial Intelligence », number 3, volume 15, 2003, pages 331–371.

[8] F. LE BER, A. NAPOLI, J.-L. METZGER, S. LARDON. *Modeling and comparing farm maps using graphs and case-based reasoning.* in « Journal of Universal Computer Science », number 9, volume 9, 2003, pages 1073–1095.

[9] A. MASSON, M. CADOT, M. ANSSEAU. *Perfectionnisme : effets du sexe et de l'échec.* in « L'Encéphale », volume XXIX, 2003, pages 125–135.

## Publications in Conferences and Workshops

[10] R. AL-HULOU, A. NAPOLI, E. NAUER. *Une mesure de similarité sémantique pour raisonner sur des documents.* in « Troisièmes journées nationales sur les Modèles de Raisonnement (JNMR-03), Paris », INRIA, 2003.

[11] M. BENOÎT, F. LE BER, J.-F. MARI, C. MIGNOLET, C. SCHOTT. CARROTAGE*, un logiciel pour la fouille de données agricoles.* in « 3ème Colloque STIC et Environnement, Rouen, France », INSA Rouen, pages 63–66, June, 2003.

[12] S. BERASALUCE, G. NIEL, A. NAPOLI, C. LAURENÇO. *Data mining in reaction databases. Extraction of knowledge on topology-based and functionality-based transformations.* in « First Lilly Distinguished Lectureship, Namur, Belgium », 2003, Poster.

[13] S. BRACHAIS, M. D'AQUIN, J. LIEBER, A. NAPOLI. *Vers un Web sémantique en cancérologie.* in « Première journée Web sémantique médical - WSM'2003, Rennes, France », J. C. ANITA BURGUN, editor, March, 2003, http://videostream.univ-rennes1.fr/~wsm/.

[14] M. CADOT, A. NAPOLI. *Règles d'association et interaction entre variables binaires.* in « Dixièmes Rencontres de la Société Francophone de Classification (SFC-03), Neuchâtel, Suisse », Presses académiques de Neuchâtel, Y. DODGE, G. MELFI, editors, pages 87–90, 2003.

[15] H. CHERFI, A. NAPOLI, Y. TOUSSAINT. *Towards a Text Mining Methodology Using Frequent Itemsets and Association Rule Extraction.* in « Journées d'informatique Messine - JIM'03, Metz, France », E. SanJuan, INRIA Lorraine, M. NADIF, A. NAPOLI, E. SANJUAN, A. SIGAYRET, editors, pages 285–294, Setpember,

2003.

[16] H. CHERFI, A. NAPOLI, Y. TOUSSAINT. *Vers une méthodologie de fouille de textes s'appuyant sur l'extraction de motifs fréquents et de règles d'association.* in « Conférence d'Apprentissage (CAp'03, plate-forme AFIA'03), Laval, France », Association Française d'Intelligence Artificielle et ESIEA Recherche, Presses universitaires de Grenoble, R. GILLERON, editor, pages 61–76, July, 2003.

[17] H. CHERFI, Y. TOUSSAINT. *Méthodologie de sélection et de lecture de règles d'association pour la fouille de textes.* in « Atelier de fouille de textes en Génomique, (conférence Extraction et de Gestion des Connaissances – EGC'03), Lyon, France », pages 1–2, January, 2003.

[18] M. D'AQUIN. *Un modèle de connaissances en RDF(S) pour raisonner à partir de cas sur le Web sémantique.* in « Actes de l'atelier raisonnement à partir de cas, RàPC'03, plate-forme AFIA », July, 2003.

[19] M.-D. DEVIGNES, Y. NORSA, M. SMAÏL, P. COLLET, L. DOMENJOUD, M. DAUA. *A Generic Solution for Automated Collecting and Integration of Biological Data from Web Sources.* in « European Conference on Computational Biology - ECCB'03 , Paris, France », 2003, Session Posters.

[20] S. HERGALANT, B. AIGLE, B. DECARIS, J.-F. MARI, P. LEBLOND. *HMM, an Efficient Way to Detect Transcriptional Promoters in Bacterial Genomes?.* in « European Conference on Computational Biology - ECCB'2003, Paris, France », pages 417–419, September, 2003, poster in conjonction with the french national conference on Bioinformatics (JOBIM 2003).

[21] C. LAURENÇO, S. BERASALUCE, P. JAUFFRET, A. NAPOLI, G. NIEL. *Fouille de données dans des bases de données de réactions : extraction de connaissances sur les méthodes de synthèse.* in « Chimiométrie 2003, Paris », 2003, Communication.

[22] F. LE BER, A. NAPOLI. *Three Galois lattices of topological relations : design, comparison and use.* in « Journées de l'informatique Messine (JIM'2003), Metz, France », INRIA Lorraine, M. NADIF, A. NAPOLI, E. SAN JUAN, A. SIGAYRET, editors, pages 55–66, September, 2003.

[23] J. LIEBER, M. D'AQUIN, P. BEY, A. NAPOLI, M. RIOS, C. SAUVAGNAC. *Acquisition of Adaptation Knowledge for Breast Cancer Treatment Decision Support.* in « 9th Conference on Artificial Intelligence in Medicine in Europe – AIME 2003, Protaras, Cyprus », series Lecture Notes in Artificial Intelligence 2780, M. DOJAT, E. KERAVNOU, P. BARAHONA, editors, Oct, 2003, http://www.loria.fr/publications/2003/A03-R-224/A03-R-224.ps.

[24] J. LIEBER. *Raisonnement à partir de cas s'appuyant sur les classifications dure, floue et élastique dans une hiérarchie de problèmes.* in « Actes des rencontres francophones sur la logique floue et ses applications (LFA-03) », CEPADUES-Éditions, Toulouse, C. FRÉLICOT, editor, pages 99–106, 2003.

[25] J.-F. MARI, F. LE BER. *Temporal and Spatial Data Mining with Second-Order Hidden Markov Models.* in « Actes des Journées de l'Informatique Messine (JIM'2003), Metz, France », INRIA Lorraine, M. NADIF, A. NAPOLI, E. SAN JUAN, A. SIGAYRET, editors, pages 247–254, September, 2003.

[26] S. MAUMUS, A. NAPOLI, S. VISVIKIS. *A first experiment on the STANISLAS cohort using closed frequent pattern search.* in « European Conference on Computational Biology – ECCB'2003, Paris, France », 2003,

Poster Session.

[27] J.-L. METZGER. *Raisonnement à partir de cas pour expliquer des organisations spatiales.*. in « Actes du 11ième atelier sur le raisonnement à partir de cas, RàPC'03, Plate-forme AFIA, Laval », July, 2003.

[28] J.-L. METZGER, F. LE. BER, A. NAPOLI. *Éléments pour la modélisation et la représentation de structures spatiales agricoles.* in « Langages et Modèles à Objets (LMO-03), Vannes », series RSTI - L'objet 9(1-2)/2003, Hermès, Paris, J.-P. BRIOT, J. MALENFANT, editors, pages 197–210, 2003.

## Internal Reports

[29] M. CADOT, A. NAPOLI. *Règles d'association et "Paradoxe de Simpson".* Rapport de recherche, number A03-R-172, LORIA, 2003.

[30] M. CADOT, A. NAPOLI, V. NAHAMA-FOURGUETTE. *Comparaison de deux techniques d'extraction automatique de règles dans les bases de données. Illustration sur des données issues d'un questionnaire sur les peurs.* Rapport de recherche, number A03-R-052, LORIA, 2003.

[31] M. D'AQUIN, C. BOUTHIER, S. BRACHAIS, J. LIEBER, A. NAPOLI. *Knowledge Edition and Maintenance Tools for a Semantic Portal in Oncology.* Rapport de recherche, Loria, Oct, 2003.

[32] J.-L. METZGER, F. LE. BER, A. NAPOLI. *Modeling and representing structures for analyzing spatial organization in agronomy.* Rapport de recherche, LORIA, 2003.

[33] C. MIGNOLET, C. SCHOTT, J.-F. MARI, M. BENOIT. *Typologies des successions de cultures et des techniques culturales dans le bassin de la Seine.* Rapport Intermédiaire, Institut national pour la recherche agronomique (INRA), 2003.

## Miscellaneous

[34] A. GUERREIRO. *Développement d'un générateur expérimental de paysages agricoles aléatoires bidimensionnels.* Rapport de stage du DESS méthodes et outils informatiques de la chimie, UHP Nancy 1, September, 2003, Laboratoire ESE (CNRS - Université Paris Sud) & LORIA.

[35] G. ROCHE. *Étude d'une plateforme de graphes conceptuels pour représenter des structures spatiales agricoles.* Rapport de stage de 2ième année ESIAL, September, 2003, LORIA.