# Towards user assistance in Data Mining

Cristina Oprean


# Towards user assistance in Data Mining

- Master thesis -

Author: Cristina OPREAN
Advisor: Philippe LENCA

August 13, 2011

cristina.oprean@telecom-bretagne.eu

# Contents

# List of Figures



# List of Tables

***Abstract*** *- The need for a unifying and comprehensive methodology is considered one of the major challenges to Data Mining. Indeed, the existing methodologies and data mining tools say what to do, but not how. The user is thus faced with many choices and no help. In the best situation we can find "local" support for a sub-task, for a sub-problem etc. without any interaction with previous and next steps in the process. There is therefore no global vision over the development of the process and in the worst case many projects fail (due to lack of resources, their complexity) or the results are suboptimal. The Data Mining software is also increasingly complex and the need for assistance for advanced users amplifies. The "general public" solutions (tools for analyzing data on Smartphones, for example) strengthens the interest of thinking about software based on methodologies for guiding users.*
*It is proposed first to describe an architecture that provides the appropriate environment for the assistant and then to describe the chain of possible tasks during the Business Understanding and Data Understanding steps of CRISP-DM process. The assistance is provided taking also in account the past similar situations. The proposal for the conceptualization of these two steps will be exemplified through a case study.*
***Advantage****: it's offering assistance during the first steps of Data Mining processes by accessing the organizational information.*

**Keywords**: Data Mining, Knowledge Discovery in Databases, assistance, process, CRISP-DM, methodology.

# 1. Introduction and problem statement

"*We are drowning in information , but starved for knowledge*" (John Naisbitt, 1982)

Nowadays, with the explosion of information, Data Mining has become one of the top ten emerging technologies that will change the world [1]. "Data Mining is most sought after..." according to Information Week Survey [2].

There are several definitions for Data Mining, but the following are the most used by the scientific community: *Data Mining is a decision support process where the users are looking for the interpretation of the data patterns* [3]. *"Data mining is the process of discovering meaningful new correlations, patterns and trends by sifting through large amounts of data stored in repositories, using pattern recognition technologies as well as statistical and mathematical techniques"* [1]. *Data Mining is an interactive and iterative process of analyzing a large set of data which extracts valuable knowledge by the data-analysts that play a central role* [4].

Data Mining (DM) and Knowledge Discovery in Databases (KDD) are both terms which define the process that has as objective information and knowledge extraction from large volumes of data [5]. Actually, the complete process of knowledge extraction is KDD, and DM is just a step in the entire process [6]. Still, in the scientific literature both terms DM and KDD are used as synonyms [7]. In order to avoid confusion, the name "Modeling" will be used instead of DM for the KDD step and DM and KDD as synonyms.

[1] http://www.gartner.com

## 1.1. Motivation and problem statement

Ever since the recent beginning of the discipline, most efforts have been focused on improving DM algorithms. These form only a link in the chain of data processing, from data recovery to result integration, in the organization information system. Between these two extremes however, the path is long and in order to help users/analysts, some efforts have been made to develop processes for DM([5], [8]).

The few existing solutions propose to structure the DM process in 6 to 8 major phases (see figure 1). Even if they are a first user support they are far from being satisfying. Processes recommend the actions to be taken but they do not say how to execute them, leaving the user with many choices (choice of an algorithm and associated parameters, protocol selection and validation measures, etc.). As a result, many available DM tools offer, in the best case, "wizard-like" assistance to help users during a DM process [55]. Even advanced users (analysts, statisticians, etc.) are sometimes lost. Moreover, the processes not specific to one application area have changed quite a bit since their origins. This is the case of CRISP-DM 1.0 [8], the most used. Version 2.0 of CRISP-DM, though promised for 2007 and to take into account the evolution of DM, has never seen the light of day [2].

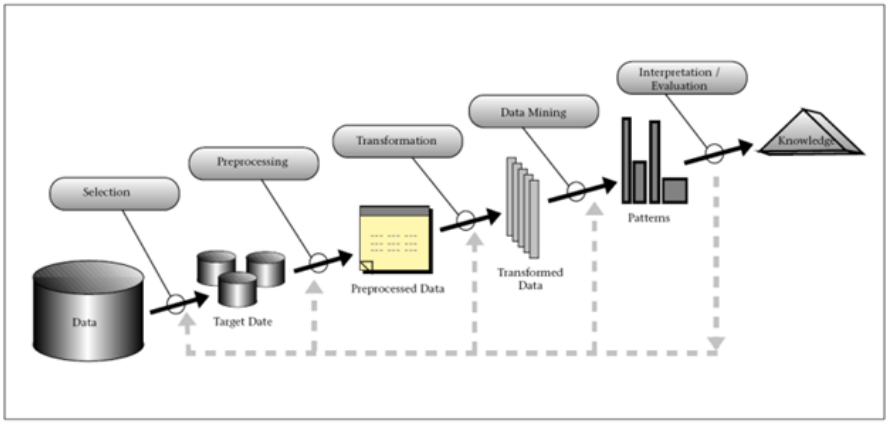


**Figure 1:** Knowledge Discovery in Databases [3]

Faced with the increasing complexity of studies, new problems, and the exponential increase of data, the need for methodological assistance is becoming stronger, especially that now, non-sophisticated users are "targeted". The lack of knowledge about new methods and new technologies in organizations is the main cause of project failures in the field of information technology and especially in Enterprise Resource Planning (ERP) — that contain all the information required to manage business [9]. As the failure rate for the DM projects is more than half of all data mining projects [3], the purpose of this thesis is to offer assistance to a DM user in order to offer help in taking the right choices and thus to avoid project failure.

The need for assistance is increasingly urgent. Indeed, providers of business intelligence applications have begun to place their software on mobile devices, and emphasize the benefits that mobile workers would have: access to analytical data and organization reports in real time from anywhere [4],[5]. So everyone could become a data miner or an analyst (with the condition to

[2] http://www-01.ibm.com/software/analytics/spss/

[3] http://www.analyticbridge.com/main/search/search?q=Search+AnalyticBridge&page=4

[4] http://www.lemondeinformatique.fr/actualites/lire-sas-pousse-la-bi-vers-l-ipad-et-l-iphone-33105.html

[5] http://www.lemagit.fr/article/mobilite-decisionnel-bi-ipad/8347/1/tribune-mobile-information-bout-des-doigtsfinalement/? utm_source=essentielIT&utm_medium=email&utm_content=new&utm_campaign=20110318&xtor=ES-6

master the process or to be assisted). This is the case, for example, for the farmers for whom ICT (Information and Communication Technology) is becoming essential. In their case, the sensors — located on farms, and animals — continue to produce information, which needs to be analyzed according to the type of forming operations. The seller who has access to billings, the sales give him the possibility to define a market, the analysis could also give him the possibility to reach an agreement with a potential client, etc.

The development of a methodology that automates the sequence of different operations required in a DM process is considered to be one of the 10 challenges of research in the domain [10]. It is proposed to contribute to this research by contributing to the development of a theoretical and methodological approach in order to provide a platform for user support and to suggest "how to", not just "what to do."

To summarize, the main identified problems in DM domain are: the lack of assistance for the end-users, the increase awareness of organization about the DM, the increase complexity of studies, data, etc. The impact of these problems in DM is devastating: more than half of the DM projects fail or end-up by being abandoned[6]. So, this thesis tackles the problem of reducing the fail rate of the projects in this domain. At the end of this paper we should be able to answer the following question: "How the processes could be enriched and help to reduce the fail rate of DM projects?"

## 1.2. Master's thesis structure

The paper is organized as follows:

**Section 2** presents a state of the art of the existing processes and tools in DM, from the first generation to the third generation processes analyzing their advantages and their inconveniences. There are two types of DM processes: research and industrial oriented. A comparison between them is made.

**Section 3** presents a different approach for DM processes. Methodologies are discussed, because the processes are enriched with assistance. After doing a more detailed analysis of the most popular DM tools, the shortcomings of these tools will be identified in order to highlight the need for a DM assistant. I will have a critical view concerning the static representation of the existent processes (ontologies) and assistants (workflow generation), by proposing pertinent criteria for comparing them. At the end of the section recommendations are made for an appropriate process representation and a helpful way of assisting the user during the DM process.

**Section 4** enters into more detail and proposes a general architecture which constitute the foundation for the future assistant. As there is no formal representation for the first steps of the DM process, a conceptualization for the first two steps of the CRISP-DM process is proposed. This conceptualization represents the first step towards the formalization of Business Understanding (BU) and Data Understanding (DU).

**Section 5** takes the proposition made in previous section and exemplifies it though a study case of a Portuguese wine organization, in order to highlight the needs of such an approach.

**Section 6** offers a summary of the master thesis work, the encountered problems and presents future work.

[6] http://www.analyticbridge.com/main/search/search?q=Search+AnalyticBridge&page=4

# 2. Processes in Data Mining - State of the art

In the previous section we have explained the difference between Data Mining and Knowledge Discovery in Databases. Before presenting the state of the art of the existing processes, we mention that in the scientific literature there is a lot of confusion between the terms "**process**" and "**methodology**". A **process** is represented by a sequence of steps executed in order to produce a certain result. A **methodology** is defined as an instance of a process, by specifying the tasks that should be executed, the inputs, the outputs and the way the tasks should be executed. In brief, a process gives the user the tasks that should be executed and a methodology tells the user also "how to" perform those tasks.

One of the problems identified by Qiang Yang and Xindong Wu in [10] is the importance of building a process that automates the composition of DM operations, in order to avoid the mistakes made by users in this area. By automating or semi-automating certain steps of the KDD process, the user effort will be reduced. There is a fairly large range of processes in DM domain, but none of these processes give any indication on "how" the task should be fulfilled.

In [11], Cios et al. explains why we need a process in KDD: (1) an unstructured application produces useless results; (2) a process "should have a logical, cohesive, well-thought-out structure and approach that can be presented to decision-makers who may have difficulty understanding the need, value, and mechanics"; (3) "Knowledge discovery projects require a significant project management effort that needs to be grounded in a solid framework"; (4) the processes in KDD should follow the other mature domains (as Software Engineering); (5) need for standardization.

This section deals with presenting an overview of the existing processes and a comparison between them, in order to highlight their shortcomings.

## 2.1. An Overview of Knowledge Discovery in Databases processes

The first KDD process was proposed by Fayyad [12] in 1996. This process consists of several steps that can be executed iteratively. Starting from Fayyad's model, lately, a lot of efforts were oriented towards finding and proposing a DM processes. Further in the section, the processes will be divided in two categories: research oriented processes and industrial oriented processes. In each section, a summary of the existing processes is made and only the most important process are presented.

### Research oriented processes for Data Mining

The process proposed by Fayyad in 1996, has met the needs of businesses and quickly became very popular. Knowledge Discovery in Databases (KDD) has as its goal the extraction of knowledge, of valid, useful and usable patterns from the large amounts of data by using automated or semi-automatic methods [13].

The KDD process is iterative and interactive. The process is iterative, which means that sometimes it may be necessary to repeat the previous steps. The problem with this process, as with all the existing processes for DM, is the lack of user guidance. The user has to make at each step decisions, being presented a certain number of choices and without any help. A bad decision in one step generates iterations in the previous steps, which means a waste of time and resources. The process offers details on the data analysis technical aspects, but no business guidance [11].

The nine steps of KDD are [14] (see figure 1):

1. **Developing and understanding the application domain** - is the initial step of this process. It prepares the user for understanding and developing the objectives of the application.

2. **Creating a target data set** - the user has to establish what data should be used, which attributes are important for the DM task.

3. **Data cleaning and preprocessing** - this step includes operations such as remove the noise and the outliers and, if necessary, strategies for handling the missing values, etc.

4. **Data reduction and projection** - This step is very important for the success of the project and must be adapted according to each database and each project's objectives. In this step are searched the correct methods for representing the data. These methods include reducing the number of used attributes. Once all these steps are completed, the following steps will be linked to the DM part, with a focus on algorithmic aspects.

5. **Choosing the DM task** - the DM objectives must be chosen (e.g. classification, regression, clustering, etc.).

6. **Choosing the DM algorithm** - In this step it is necessary to determine the specific methods for pattern searching, deciding the appropriate algorithms and parameters.

7. **Data Mining** - the selected algorithms with their parameters will be implemented. Maybe it will be necessary to apply the algorithm several times to get the expected result.

8. **Interpreting the mined patterns** - includes the evaluation and the interpretation of the discovered patterns. This step provides an opportunity to return to the previous steps, but also to have a visual representation of the patterns, to remove redundant or non-representative patterns and to turn them into information understandable by the user.

9. **Consolidating discovered knowledge** - includes adding the discovered knowledge in other systems, for other actions.The effect of the knowledge on the system should be measured, check and resolve potential conflicts with the prior knowledge.

**Process application**: The process is iterative and may have several loops between either of the two steps. The KDD has become itself a model for other processes. The process is included into the commercial system MineSet [15] and has been used in several areas as engineering, medicine, e-business, production, software development, etc..

**Other research processes**:

- **Anand et Buchner** [16] - hybrid process for solving the problems in cross-sales and analyzing the marketing data on the Internet. Contains eight steps: *Human Resources Identification*, *Problem Specification*, *Data Prospecting*, *Domain Knowledge Elicitation*, *Methodology Specification*, *Data Preprocessing*, *Pattern Discovery*, *Knowledge Post-Processing*. The process provides a detailed analysis for the initial steps of the process, but does not include the necessary activities to use the discovered knowledge and the project's documentation.

- **The 5 A's** - The 5 A's is a process developed by SPSS [17] to give a broader view of data analysis and DM process. The importance of this model is given by the "Automate" step, which automates the process of DM to help the non-expert users to apply already defined methods to new data. The five steps of this process are: *Asses*, *Access*, *Analyze*, *Act* and *Automate*. The main disadvantage is that the **5 A's** do not contain the steps "*Business*

*Understanding*" and "*Data Understanding*", considered very important to understand the business objectives and to test data quality. The process was abandoned in 1999.

- **Cios** [18] - it was proposed by adapting the CRISP-DM process to meet the academic research needs. The process is using technologies like XML[7], PMML[8], SOAP[9] and UDDI[10]. The model consists of six steps: *Understanding the problem domain*, *Understanding the data*, *Preparation of data*, *Data Mining*, *Evaluation of the discovered knowledge*, *Using the discovered knowledge*.

### Industry oriented processes for Data Mining

CRISP-DM [19] was proposed in 2000 to meet the needs of industrial projects in Data Mining. CRISP-DM is described as a hierarchical process consisting of several tasks with four levels of abstraction: phase, generic task, specialized task and process instance. Shortly after its appearance, CRISP-DM has become the process the most widely used for DM projects, according to the KDnuggets polls realized in 2002[11], 2004[12], 2007[13]. CRISP-DM stands for Cross-Industry Standard for Data Mining and contains a cycle of six steps:

1. **Business Understanding** - this initial phase focuses on understanding the business and DM objectives and project requirements.
2. **Data Understanding** - it begins with by collecting the initial data and continue with several activities in order to become familiar with the data, identify data quality problems, determine the meta-data.
3. **Data preparation** - this phase includes all activities necessary to construct the final database.
4. **Modeling** - in this phase are selected and applied several DM techniques. Since there are several specific forms of data construction, most of the time it is necessary to return one step back.
5. **Evaluation** - at this stage, models are evaluated and the steps taken to build the model are reviewed to ensure that the project meets the business objectives, defined at the beginning of the project.
6. **Deployment** - the creation of the model is not the end of the project. Although the initial objective of the project is to increase data knowledge, the acquired knowledge needs to be organized and presented in a certain manner to the client.

The sequence of phases is not mandatory. We can go between the phases, as suggested in figure 2 by the arrow indicating the most important and frequent dependencies between phases. The main problem of this process is the fact that it is iterative, we need to go back and redo some steps to get the desired result. CRISP-DM does not guide the user on "how" the tasks should be performed, but the model is easily understood and well documented. In general, CRISP-DM is a process extensively used in industry ("defacto standard").

[7] http://fr.wikipedia.org/wiki/Extensible_Markup_Language
[8] http://en.wikipedia.org/wiki/Predictive_Model_Markup_Language
[9] http://en.wikipedia.org/wiki/SOAP
[10] http://en.wikipedia.org/wiki/Universal_Description_Discovery_and_Integration
[11] http://www.kdnuggets.com/polls/2002/methodology.htm
[12] http://www.kdnuggets.com/polls/2004/data_mining_methodology.htm
[13] http://www.kdnuggets.com/polls/2007/data_mining_methodology.htm

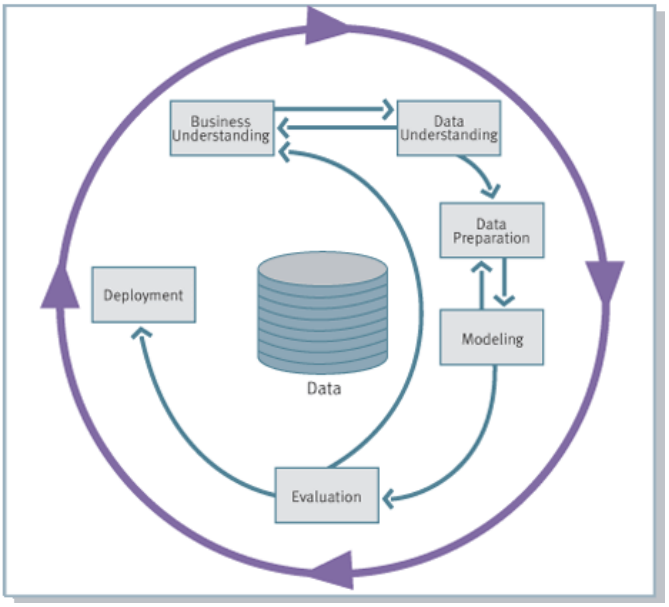


**Figure 2:** CRISP-DM process [8]

After the first version of CRISP-DM, DM applications have evolved. So the needs have changed: new techniques for data, increase the scale factor for data manipulation, real-time deployment, etc. To meet the changing needs of DM, CRISP-DM process will be improved in a new version of CRISP-DM 2.0 [19]. Normally this version should have appeared in 2007, but was not finalized yet.

Lately CRISP-DM entered in the shadow, after it was bought by IBM[14]: "CRISP is a great framework to communicate the process. Most people think it's all about the algorithms, but that ends up being a much smaller part of the total lifecycle. I'm disappointed that the crisp URL redirects to some generic IBM page, and not even to a page about CRISP. Anyone know what's going on?" (a LinkedIn[15] conversation on the CRISP-DM group).

CRISP-DM has been used in the following domains: medicine, marketing, sales, engineering, etc.

**Other industrial processes**:

- **SEMMA** [20] - the five steps of SEMMA are: *Sample*, *Explore*, *Modify*, *Model* and *Asses*. This process is included in SAS Enterprise Miner[16].
- **Cabena** [21] - has also five steps: *Business objectives determination*, *Data preparation*, *Data Mining*, *Analysis of results* and *Assimilation of knowledge*. This process is used more in the marketing and sales domain and represents one of the first processes that takes into account the business objectives, even though it is not very well documented.

## 2.2. Comparison between research and industry oriented processes

To summarize what has been presented so far, in this subsection we propose a comparison between the processes and methods mentioned in the previous sub-chapters. The following tables are based on the studies made by Kurgan et al. [22] and Marban et al. [21]. The comparison

[14] http://en.wikipedia.org/wiki/IBM

[15] http://www.linkedin.com/groupAnswers?viewQuestionAndAnswers=&discussionID=63862178&gid=1950547&trk=eml-anet_dig-b_pd-pmr-cn

[16] http://www.sas.com/technologies/analytics/datamining/miner/

is made according to the following criteria: application domain, the number of steps, related approach, software support, advantages and inconveniences. In table 1 and table 2 it can be seen that the processes steps are not quite equivalent between processes.

As table 1 indicates, there are several common features between these processes. Most processes follow the same sequence of steps and sometimes they use the same steps. Another important similarity between these processes are the iterative loops required between some of the steps. The process differentiates in the first step of "Business Understanding" (SEMMA and 5A's do not have it) and the final step of "Using discovered knowledge", which does not exist in all processes (SEMMA, 5A's, Anand and Buchner do not have it).

## 2.3. Discussion

Each presented process can be instantiated through DM systems. Piatetsky-Shapiro in [23] talks about several categories of DM systems.

The **first generation** of DM systems was primarily intended for data analyst experts in order to perform simple DM tasks on small size databases (reduces number of examples and attributes). The **second generation** tools (processes and DM tools were more and more complete) have been developed to cope with the complexity of the projects (volumes of data, heterogeneity of data, etc.)[32]. Even though these tools became more powerful, they are still oriented towards the expert and don't use the business knowledge.

Thus Fayyad [3] proposes the first process. This process is iterative, interactive, and it consists of several steps that run sequentially. Although important details were subsequently made with particular efforts to provide "standards" (SEMMA [24] - process associated with the SAS Enterprise Miner[17] and CRISP-DM[8] - associated with Clementine [26]), the process still remains iterative (the user is not guided, certain process tasks must be performed several times in order to obtain a satisfactory result or to abandon by lack of resources , see figure 1- the dotted arrows).

The DM tools that were developed in order to be compatible with the processes have the same defects. The race of having more and more features, increasing numbers of algorithms from one version to another, makes their use increasingly complex. Thus, despite their many qualities (automation of of some repetitive tasks, advanced results and visualization, support for interactive exploration of the model, support for comparing models, etc.), no DM tool (Weka [41], RapidMiner, Tanagra [25], SAS Enterprise Miner) offers full support to the user.

As highlighted R. Rakotomalala on his blog [27] "there are no good or bad data mining tools. There are data mining tools that meet the specifications or not".

The need for a third generation of DM tools became more and more stronger. We live in a world in which mining distributed data, understanding the client's need and offering viable solution became a "must". To respond to this insufficiency, later in this paper, the existent DM tools are analyzed — in order to identify the characteristics that a third generation process and tools need to meet, and at the end a prototype is proposed, as a solution . As a methodology is defined by the tasks that should be executed (the process) together with the performing methods for each task, it is considered to propose a solution that imitates the behavior of a methodology: a process representation together with an assistance for each task as a plug-in within a DM tool.

We can conclude this section with the following statement: "We emphasize that there is no universally "best" KDP model. Each of the models has its strong and weak points based on the application domain and on particular objectives." [11].

[17] http ://www.sas.com/

| | KDD (1996) | SEMMA (1996) | 5A's (1996) | Cabena et al. (1997) | Anand and Buchner(1998) | CRISP-DM (2000) | Cios et al. (2000) |
|---|---|---|---|---|---|---|---|
| **Domain** | Academic | Industrial | Academic | Industrial | Academic | Industrial | Academic |
| **Number of steps** | 9 | 5 | 5 | 5 | 8 | 6 | 6 |
| | 1. Learning application domain | | 1. Asses | 1.Business Objectives Determination | 1.Human resources identification | 1.Business understanding | 1.Understanding the problem's domain |
| | | | | | 2.Problem specification | | |
| | 2. Creating a target data set | 1. Sample | | 2. Data preparation | 3. Data prospecting | 2. Data understanding | 2. Understanding the data |
| | | | | | 4. Domain knowledge elicitation | | |
| | 3. Data cleaning and prepocessing | 2. Explore | 2. Acces | | 5. Methodology identification | 3. Data preparation | 3. Preparation of the data |
| | 4. Data reduction and projection | 3. Modify | | | 6. Data preprocessing | | |
| | 5. Choosing the function of Data Mining | 4. Model | 3. Analyse | 3. Data Mining | 7. Pattern discovery | 4. Modeling | 4. Data Mining |
| | 6. Choosing the Data Mining algorithm | | | | | | |
| | 7. Data Mining | | | | | | |
| | 8. Interpretation | 5. Asses | 4. Act | 4. Analysis of results | 8. Knowledge post-processing | 5. Evaluation | 5. Evaluation of the discovered knowledge |
| | 8. Using discovered knowledge | - | - | 5. Assimilation of knowledge | - | 6. Deployment | 6. Using the discovered knowledge |
| | | | 5. Automate | | | | |
| **Linked approaches** | KDD | KDD | - | KDD | KDD | CRISP-DM | CRISP-DM |
| **Software support** | MineSet, KDB2000, VidaMine | SAS Enterprise Miner | - | - | - | Clementine | Grid Miner-Core |

**Table 1:** Comparison between processes steps (after [6] and [7])

| | KDD (1996) | SEMMA (1996) | 5A's (1996) | Cabena et al. (1997) | Anand and Buchner(1998) | CRISP-DM (2000) | Cios et al. (2000) |
|---|---|---|---|---|---|---|---|
| **Advantages** | - The process is interactive and iterative<br>- can process large amounts of data<br>- detailed technique description on data analysis | - we can go back to the exploration step for further refining on the data | - automation of the Data Mining process<br>- the non-expert users can apply already defined models on new data | - the model is easy to understand by non-expert users | - offers a detailed analysis of the first steps of the process<br>- emphasizes the iterative nature of the process, where the experts check the knowledge obtained after the last step and decide whether to refine, to rerun a part or the entire process | - hierarchic model<br>- the sequence of steps is not compulsory<br>- the most used Data Mining process<br>- takes into consideration the business aspects<br>- it is closer to the real-life projects<br>- easy to understand<br>- well documented | - includes a description of the first steps of the process<br>- more feedback mechanisms<br>- we can go back from Data Preparation to Data Understanding<br>- emphasizes and describes the iterative and interactive nature of the process |
| **Inconveniences** | - omits the business step<br>- a lot of unnecessary loops between steps<br>- not too much information on the iterations between steps<br>- describes only the tasks and not how to execute them | - it doesn't include a special step for indicating how to use the obtained knowledge<br>- describes only the tasks and not how to execute them | - no alternatives for applying the built models or or for using the obtained knowledge<br>- describes only the tasks and not how to execute them | - not so much information on the iterative nature of the process<br>- no Data Understanding step<br>- describes only the tasks and not how to execute them | - no information on using the obtained knowledge<br>- describes only the tasks and not how to execute them | - limited information on the feedback loops<br>- needs to be updated<br>- describes only the tasks and not how to execute them | - oriented on academic research<br>- describes only the tasks and not how to execute them |

**Table 2:** Processes advantages and inconveniences (after [6] and [7])

# 3. A new approach for processes and tools in Data Mining

As the technology evolves, the produced data increases, which amplifies the need for a more complex data analysis [39]. Data analysis is the discipline that transforms the data into information [40], so we need a deeper knowledge about the continuous expanding set of existing algorithms.

There are several DM tools implemented to help the user in modeling the data, by offering a wide range of operators. The most popular DM tools are: Weka [41], Rapid Miner [42], KNIME [43], Clementine [44], etc. Each of these tools offer a large number of operators, but the users get confused having too many choices and no help. Having complex data, a large range of operators, is a challenge to follow all stages of the DM process and to build valid models. A model is a DM workflow, which means a sequence of operators (linear, acyclic or nested).

Further in this section it will be presented a comparison of different DM tools, to understand better their shortcomings and to define a list of characteristics of a "good" DM methodology and a "good" DM tool. In my opinion an assistant consists of two layers: a layer that is the process and knowledge representation and another layer that is represented by an advisor that helps the user during the manual construction of the workflow or proposes models during the automatic construction of the workflow.

Once the definition of an ideal assistance is established, some criteria to compare existing DM assistants can be defined. Firstly the criteria for comparing the process and the knowledge representation are defined, secondly the criteria for comparing different types of advisers and at the end some recommendations are made for a good assistant who gathers the advantages of existing ones.

## 3.1. Analyzing different types of Data Mining tools

Analyzing different types of Data Mining tools will help to establish the qualities and needs behind each of them. This will allow to map these strengths and needs into a "good" methodology, which will be dealing with the changes and the complexity of data. The analyzed DM tools are: Weka, Rapid Miner and KNIME. For each tool, workflows will be built for the "*iris*" database and it will be analyzed the ergonomics of these tools together with the offered support during the CRISP-DM steps (operators, information, assistance).

**Weka** "provides implementations of learning algorithms that you can easily apply to your dataset. It also includes a variety of tools for transforming datasets, such as the algorithms for discretization... You can preprocess a dataset, feed it into a learning scheme, and analyze the resulting classifier and its performance - all without writing any program code at all." [41]

**Rapid Miner** "is the world-leading open-source system for data and text mining. It is available as a stand-alone application for data analysis, within the powerful enterprise server setup RapidAnalytics, and finally as a DM engine which can be integrated into own products. By now, thousands of applications of RapidMiner in more than 40 countries give their users a competitive edge." [42]

**KNIME** "(Konstanz Information Miner) is a user-friendly and comprehensive open-source data integration, processing, analysis, and exploration platform. From day one, KNIME has been developed using rigorous software engineering practices and is currently being used actively by over 6,000 professionals all over the world, in both industry and academia." [18]

The results of the last KDnuggets poll[19] show that Rapid Miner is the most used DM tool (37.8%), KNIME and Weka are less used with a percentage of 19.2% and 14.3%, respectevly.

[18] http://www.knime.org/

[19] http://www.kdnuggets.com/polls/2010/data-mining-analytics-tools.html

Here it will be presented only the final results of the analysis. A more detailed analysis could be found in my last semester's work [45].

The grades are assigned on a scale from 0 to 10. The grades for each of the 7 attributes: Ergonomics, Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation and Documentation, were assigned based on their range of operators and the help that the user gets in choosing the operators. The results can be seen in figure 3.

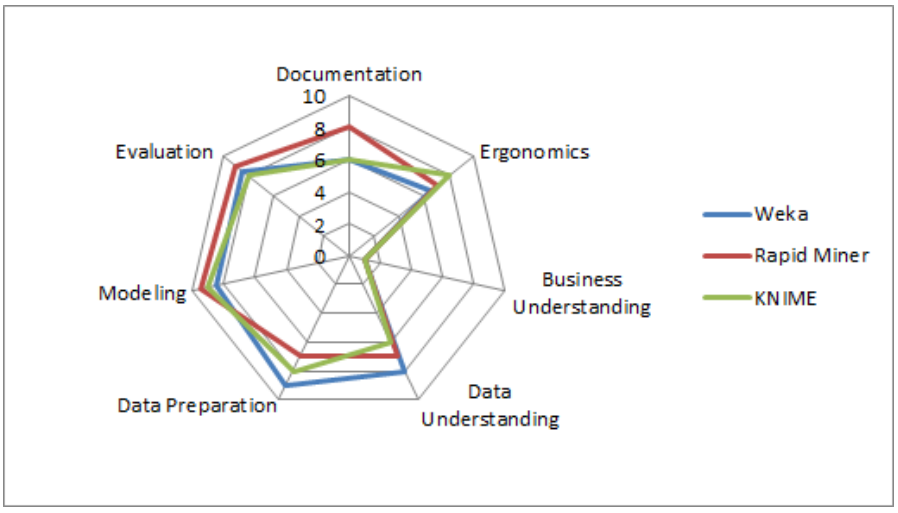


**Figure 3:** Tools Comparison

As shown in figure 3, it can concluded that most of the developer's attention is focused on the implementation of algorithms for the Data Preparation, Modeling and Evaluation. The user is left alone in the first steps, because for the moment, there is no assistance at the beginning of the project. The presented tools suppose that the user has already collected all the data and all he has to do is to model it, but without any assistance (only the minimum information about the operators) and without using the past cases. Lately, Rapid Miner has developed a plug-in, named PaREn[20], which offers support by automatically building a classification model for a given dataset. Even though, Rapid Miner offers a lot of operators, it's not so easy to use them. The wide range of operators and the complexity of data makes it almost impossible to find the optimal solution (the model that gives the best results). It offers only a valid workflow that is appropriate with the data features, but it's optimization should be done by an expert. The summary of the actual system's limits are:

- even though they offer to the user a wide range of operators, the built workflows have a limited number of operators.
- they do not offer results interpretation (for a specific type of user - beginner, expert, etc.)
- don't offer support for all steps of the modeled process
- don't take into consideration past similar cases

## 3.2. Characteristics of a "good" methodology and "good" Data Mining tools

Continuing the concept of third-generation DM systems [46], it is necessary to propose a methodology, instantiated with the help of DM tools. These tools are not only those that do not constrain

[20]http://rapid-i.com/content/view/240/1/lang,en/

the user, but also automate everything that can be automated and offers the needed assistance. The DM tools analysis that we have made, allows us to identify a number of necessary features for a "good" methodology and "good" DM tool [47].

- **Interactive, but not iterative** - the methodology must allow user's intervention at all times; ideally, it must minimize iteration between steps (see figure 1 which presents an iterative situation: there are iterations between steps and after each step the user moves along to the next or goes back to previous steps)[28].
- **Guidance for the user at all steps**, in order to make the best choice given a situation. The methodology must show at each step how to achieve the tasks and where to find the necessary information [29].
- **To offer to the user an intuitive environment** - the DM tool used to instantiate the methodology has to be implemented with a graphical interface, user-friendly, allowing to build manually the workflow or, to propose to the user some adapted workflows [30].
- **Automatic validation of the workflow** - the DM tool used to instantiate the methodology has to verify in each moment the workflow's semantic, to guide the user towards the construction of good workflows [31].
- If the computational method is not locally implemented, the software used to instantiate **the methodology has to be able of proposing externally solutions eventually in a transparent manner** [32].
- **To give the possibility of visualizing by the user in each step the workflow's execution progress** - very often the user is not informed by the workflow's execution progress and this is one of the very large data processing project's abandon problems (eventually the execution time estimation will be useful for informing the user by the finality of his project) [29].
- **To completely or partially reuse the built workflow** - the used modules for the construction of the workflows has to be loosely coupled and has to be reusable (for other modules or workflows) [33].
- **To share the knowledge (a meta-analyses platform)** - the built workflows has to have the possibility of being shared with the data mining community in order to help the research in this domain and reduce the execution time [34].
- **Encourage "Agile" environment and development** - software must stimulate the user's intuition and allow the integration of new solutions [35].

Once the characteristics of a good methodology and a "good" DM tool are decided, we arrive at the conclusion that different modules will need to be built. So, as it can be seen in figure 12 it will be needed: **Ontologies** for the representation of the DM process and the knowledge — for assuring the guidance, the validation of the workflows, **Workflow generator** for helping the user by proposing him off the shelf valid workflows, **CBR** for reusing the past experiences, sharing knowledge and **Meta leaning** module for learning from past experiences and avoiding the same mistakes.

In the following subsections we will focus on the process representation and the workflow generation.

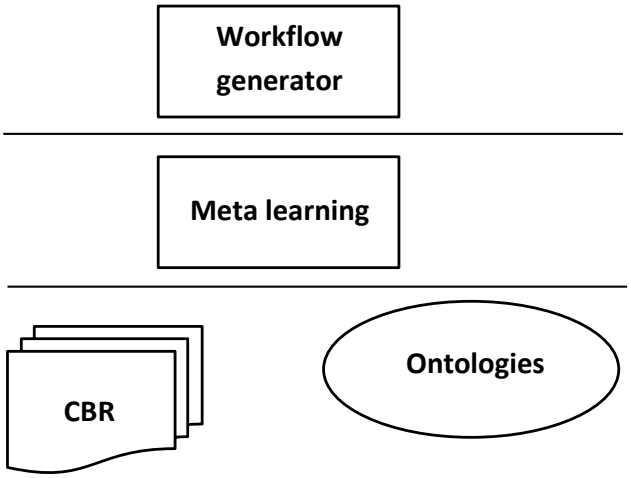


**Figure 4:** Needed modules for assuring the assistance

## 3.3. Analyzing different types of process and knowledge representation

This subsection aims to analyze several types of ontologies for the representation of the DM process and to propose some recommendations for a better representation. It has been chosen to analyze CRISP-DM as a DM process through ontology representation. Following the characteristics defined in the previous section, it will be defined firstly some criteria for comparing the different representations proposed in the DM domain. After, each of the analyzed ontologies will be presented, with their advantages and inconveniences, and, finally the drawn conclusions.

### 3.3.1. Defining the main criteria of comparison for different types of process and knowledge representation

Having as starting point the work from [39], in which the authors make a history of assistants, with their features, and to expand it, I begin my analysis from the process representation, because I believe that plays an important role in the assistant's activities.

The proposed criteria for analyzing the ontologies are:

- **Process criteria**

(a) **The Data mining process represented** - as CRISP-DM is the "defacto standard" for the DM processes, we want to offer assistance during its execution.

(b) **The modeled steps from the process** - it is wanted to cover all stages of the process.

- **Ontology criteria**

(c) **Completeness** – according to [48] refers to the coverage of the domain, in this case the modeled process. The idea is to have a process ontology that covers all the steps of the process (in length), and all the operators (in depth).

(d) **Extensibility** – refers to the facility of adding new classes to the ontology.

(e) **Modeling language** – in this case, it have been chosen to work with OWL-DL[21], because we want a "maximum expressiveness while retaining computational completeness".

[21] http://www.w3.org/TR/2004/REC-owl-features-20040210/#s1.2

(f) **Reasoner** - has to be chosen according to the modeling language. As it has been chosen OWL-DL for modeling the ontology, it can be used Pellet[22] as reasoner.

### 3.3.2. Comparison the different types of process and knowledge representation

It will be presented, further in this work, the chosen ontologies for analysis, together with their limits and advantages.

The first two ontologies, even though are developed by different teams, are part of the same european project e-LICO[23], in order to offer support for the creation of a framework for the development of complex processes for data analysis. Actually, DMWF and DMOP create DMO (Data Mining Ontology).

**DMWF** (Data Mining Ontology for Workflows) – this ontology is proposed in [36] and gathers rules for KDD domain in order to establish how to execute the steps from CRISP-DM **(a)**. The ontology contains informations on the handled objects (I/O Objects), on the meta data, on the operators applied on the data (the operators are enriched with conditions and effects over data) and the description of objectives. **Advantages**: the modeled process is CRISP-DM **(a)**; it integrates OWL-S to add operators accessible through web services, so it is an easy extensible ontology **(d)**; it is enriched with a wide range of operators **(c)**; the used modeling language is OWL-DL, enriched with conditions and effects on data **(e)**; for classifying the ontology it is used an in-house reasoner, a combination between Flora 2[24] and XSB[25]**(f)** (it is an advantage, because once that we have installed the eProPlan[26] plug-in, rules, conditions and effects can be written, that could be treated by this type of reasoner). One can modify the ontology in Protege[27], by including in the ontology tool the plug-in eProPlan, so it is available for analysis. **Inconveniences**: the most important limit is the fact that it doesn't treat all the steps from CRISP-DM process. The focus is on Data Preparation and Modeling **(b)**, so it doesn't offer help for the first two steps of CRISP-DM process.

**DMOP** [37] is an optimization ontology, which searches for the best algorithms and parameters, suitable for a certain task within the process, representing the main advantage. Another advantage is the fact that it is available for analyzing it. **Inconvenience**: for the moment, it is oriented towards classification, and it doesn't cover all the steps from CRISP-DM **(b)** and has a limited number of operators **(c)**. As it is complementary to DMWF ontology, we will refer later as DMO (DMWF+DMOP), which incorporates both of their advantages.

**KDDONTO** [31] provides a formalism for the algorithms from KDD domain. The ontology is used for finding the suitable algorithms for a certain situation. It is part of an ambitious project: a distributed and collaborative KDD environment, for accessing the web services (the algorithms). **Advantages**: it is easy extensible **(d)** through the possibility of accessing web services; it is modeled with OWL-DL language **(e)**; the used reasoner for classifying the ontology is Pellet **(f)**; available for analyzing it. **Inconveniences**: the followed DM process is generic (KDD) **(a)**; not all the steps of the process are covered (there are implemented 15 algorithms for classification , clustering and evaluation) **(b)**; neither we cannot talk about the completeness **(c)**.

**OntoDM** [38] is a generic ontology for describing the DM domain, by proposing a general framework for Data Mining with the help of ontologies (it describes the DM entities, the data, the main types, the tasks, the constraints, scenarios, etc.). **Advantages**: It assures the interoperability with other ontologies, so it is easy to extend it **(d)**; can store DM scenarios; available

[22] http://www.mindswap.org/2003/pellet/
[23] http://www.e-lico.eu/
[24] Flora 2
[25] XSB
[26] https://trac.ifi.uzh.ch/eproplan
[27] http://protege.stanford.edu/

| | **DMO** (DMWP+DMOP) | **KDDONTO** | **OntoDM** | **Zahova** | **Charest et al.** | **Choinski** |
|---|---|---|---|---|---|---|
| 1. *The DM process representation* | CRISP-DM and optimization ontology | KDD | generic DM process | CRISP-DM | CRISP-DM | CRISP-DM |
| 2. *Modeled steps* | Data Preparation Modeling, Evaluation | Data Preparation Modeling, Evaluation | ? | Data Preparation Modeling | Data Understanding Data Preparation Modeling | Data Understanding Data Preparation Modeling, Evaluation |
| 3. *Completeness* | >100 operations ++ | 15 alg. impl. 0 | - | - | ? | + |
| 4. *Extensibility* | through OWL-S ++ | web services ++ | easy to extend not OWL-S compatib. + | - | ? | + |
| 5. *Modeling language* | OWL-DL | OWL-DL | OWL-DL | OWL-DL | OWL-DL | OWL-DL |
| 6. *Reasoner* | XSB+Flora2 | Pellet | Pellet | ? | ? | ? |

**Table 3:** Summary table for process representation — "–" (very bad), "-" (bad), "0" (insufficient), "+" (good), "++" (very good) and "?" (undefined)

for analyzing it. **Inconveniences**: it is too generic, it can offer help to user, but inefficiently; it doesn't follow CRISP-DM process **(a)**; it is not complete **(c)** and it must be populated; not compatible with OWL-S[28].

**Zakova** (named after it's author) [33] wants to convert a KDD task into a planning task (to plan which is the algorithm that can be applied for this situation in order to produce a certain result). **Advantages**: the represented process is CRISP-DM **(a)**; it contains also an ontology for storing the workflows; the used modeling language is OWL-DL **(e)**; it is available for analyzing it. **Inconveniences**: it models only Data Preparation and Modeling **(b)**; not complete **(c)**, with limited number of operators.

**Charest** (named after it's author) [53] defines the used concepts: task, activity type, algorithm, etc. and it is linked to a CBR (Case Base Reasoning) which contains the detailed information about the data cleaning, model parameters, etc. (in brief, it is linked through SWRL[29] Rule Set, the problem to the solution). The ontology offers support for the non-experienced data miner, giving him recommendations during the DM tasks. **Advantages**: the represented process is CRISP-DM **(a)**; CBR can store the past cases, without a formal representation; the CBR can learn from past cases; the ontology is developed by using OWL-DL **(e)**. **Inconveniences**: it covers only Data Understanding, Data Preparation and Modeling **(b)**; sometimes the users have to fill in information for certain tasks; CBR offers support only for classification problems; the ontology is not available for analyzing it.

**Choinsky** (named after it's author) [51] proposes a framework for the KDD process by encouraging the collaboration between the domain experts and technology experts. It gathers more ontologies: Domain ontology, Corporate Data Model Ontology (CDMO), Business Ontology, Data Mining Ontology (for all the operators), and CRISP-DM ontology to structure the process (from the generic phase to the specific phase). **Advantages**: CRISP-DM ontology follows the structure of CRISP-DM process **(a)**; easy to extend **(d)**; as it is built on the [36] ontology (ontology for DM) and [49] ontology (ontology for grid programming), it gathers their advantages; it is a mixture between DMO ontology and CRISP-DM ontology. **Inconveniences**: the modeled steps are: Data Preprocessing, Modeling and Post processing **(b)**; the ontology is not available for analyzing it.

Table 3 summarizes the advantages and disadvantages of ontologies, by associating marks for each criteria. The marks are the following: "–" (very bad), "-" (bad), "0" (insufficient), "+" (good), "++" (very good) and "?" (undefined).

**Recommendations**:

[28] http://www.w3.org/Submission/OWL-S/

[29] http://en.wikipedia.org/wiki/Semantic_Web_Rule_Language

As the ontologies are difficult to be built and maintained (they have to be build by an expert), it is better to have a starting off-the shelf ontology (could be one the above ontologies). After having made a detailed analysis of the ontologies presented above, it can be observed the complexity of DMO ([36], [37]), the team behind this project (e-LICO), which developed plug-ins for Rapid Miner. The main advantages are the multitude of operators and the facility to add other operators through web services. DMOP ontology, built for the optimization of workflows, is currently oriented towards classification. A major limitation of this approach is the lack of a mechanism for learning from past experiences. And like most of the presented ontologies, it lacks of information about the first and second step of the CRISP-DM (Business Understanding and Data Understanding).

Another interesting ontology is the one proposed by Choinski in [51], which consists of several ontologies (Data Mining ontology — algorithms for taxonomy, CRISP-DM ontology to structure the CRISP-DM process , the Domain Ontology, the Business Ontology) and interprets the results to provide them to the users according to their type (business user, technical user, etc.). The DM ontology combines the ontology for grid programming with the ontology of [36], but the main problem is that is not public. Another ontology that can be taken into account is the ontology of Charest and Delisle [53], also not public, which in addition to the ontology of [36] (DMO) offers the possibility of learning from previous cases (ontology + CBR).

My recommendation is to choose among these three ontologies and ideally combine their concepts in order to minimize their limitations. As the last two are not public, it should be considered DMO ontology as a starting point and to complete it with the other two presented (Choinsky et Charest). It is preferred the structure of Choinski's ontology, built in accordance with the four abstraction levels of CRISP-DM, and which displays the same results in different ways, depending on the type of user, while keeping the multitude of operators of DMO ontology, finding the optimal parameters with DMOP, and the opportunity to learn from previous experience as in Charest's ontology.

## 3.4. Analyzing different types of advisers

The second part of the analysis should offer a more complex vision on the help offered to the user, taking into account also the possibility to learn from past experiences. In order to do that, the focus will be on the presented ontologies (the ontologies are the basis of the assistant) and on the DM tools analyzed in Subsection 3.1. The DM assistants must meet certain characteristics. It is difficult to specify a detailed architecture for the assistant, but it can be expressed its main characteristics in terms of quality of the results (if the assistance is specific to the needs, the quality of the results is very good too) and the time spent to perform specific tasks of the DM process is minimal. The users should be assisted during the manual construction of the workflow (during the process) or to automatically propose them several workflows that meet their needs.

### 3.4.1. Defining the main criteria of comparison for different types of advisers

In order to analyze the existing solutions, some criteria for comparing them will be defined and some recommendations for a proper assistant will be made. Two categories of criteria are considered: *Workflow criteria* and *Environment criteria*.

- **Workflow criteria**

(a) **Automatic workflow generation and classification** - the system should automatically generate workflows if the user wants, based on the defined objectives and on the data features.

(b) **Types of generated workflows** - the workflows can be linear (no recommended), acyclic or nested (as in Rapid Miner).

(c) **Workflow validation** - the validity of workflow is assured through the logic of the ontology.

(d) **Optimal workflows** - the workflow will give the best result. For example, for a prediction case, for a given database, the chain of operators within the workflow will be enriched with the appropriate parameters that will give optimal results.

(e) **Possibility to store or to reuse the workflows** - in order to avoid the same mistakes over and over again, the user can use the past cases to build the workflow. The system offers the possibility to store the cases that matter.

- **Environment criteria**

(f) **Intuitive environment** - intuitive for all types of users (experts, non-experienced, etc.)

(g) **Interactive system** - a system that don't take the decisions itself, but is influenced by the user's actions. The user can have the possibility to follow at each step the execution status of the workflow.

(h) **Results interpretation** - the same results should be presented differently for users from different domains (using the appropriate vocabulary of the domain).

(i) **Guidance during manual construction of the workflow** - it is related to Interactive system criteria (g). At each moment, the user can see/ask the system what should do next or how to do it.

(j) **Parallel execution of workflows** - the generated workflows could be executed in parallel.

(k) **Software integration** - refers to the way the assistant is integrated within the DM tool.

### 3.4.2. Comparison the different types of advisers

Further in this work, there are presented several ways of assuring the assistance, together with their limits and advantages.

In the article proposed by **Kietz et al.** [36] (**eProPlan**), the assistance is provided through an IDA-API (Intelligent Discovery Assistance - Application Programming Interface). The workflows are generated through an HTN (Hierarchical Task Network) planner based on the DMWF ontology having the objectives and the characteristics of the data. The basis on the planner is the ontology, which defines its behavior. The process CRISP-DM was modeled through the decomposition Task/Method. It's behavior can be changed once that the objectives are changed. The planning and sharing the workflows can be now possible through Rapid Miner. **Advantages**: the planner can generate workflows and will classify them **(a)**; the generated workflows are nested **(b)**; the validation is assured through the DMWF ontology (conditions/effects) **(c)**; as it is a version integrated into Rapid Miner, it quite assures the interactivity with the system **(g)**, **(k)**; the user is somehow guided during the manual construction of the workflow (he can see in real time the errors of interconnection for operators and proposes the user solutions) **(i)**. **Inconveniences**: the generated workflows use a limited number of operators and have to be optimized by an expert **(d)**; the generated workflows don't use past cases, but the workflow can be stored and shared by using myExperiment[30] environment, a collaborative platform; the environment is

[30] http://www.myexperiment.org/

not the most intuitive — the wide range of operators intimidates the user **(f)**; the results are not interpreted **(h)**; the execution in parallel of several workflows cannot be parallelized **(j)**.

The solution proposed by [37] is oriented towards finding ways to provide a better assistance to the data miner. The ontology proposed by Hilario **DMOP**, wants to optimize the process of DM through meta-learning: the selection of algorithms and models, order workflows and optimize workflows DM. **Advantages**: the main advantage is that it targets the workflow's optimization by selecting the right algorithms with the right parameters. **Inconveniences**: it is oriented on classification task and it is not included in a DM tool **(k)**.

The assistance proposed by [31] has as basis the **KDDONTO** ontology, and it generates valid workflows in three phases: define the objectives (each workflow is built with a purpose); by following the process and having the data features, the workflows are generated (the process of workflow generation is iterative - having a task, it goes back and searches for the compatible algorithm); and order the workflows (the user can indicate some constraints to limit the number of generated workflows). **Advantages**: The main advantage is the automation of valid workflow generation from objectives and data features **(a)**, **(c)**; the generated workflows are acyclic. **Inconveniences**: As it is not included in a DM tool, all the environment criteria are not respected; the workflows are not optimal **(d)**; and they cannot be stored nor reused **(e)**.

**Zakova**'s assistant [33] automatically generates workflows by using the Fast-Forward algorithm (FF) (it checks the neighboring states and verifies the possibles algorithms preconditions). The generated workflows are presented to the user visually. **Advantages**: the valid workflows are automatically generated **(a)**, **(c)**; their type is acyclic **(b)**; it is integrated in Orange[31] DM tool. **Inconveniences**: no optimal workflows **(d)**; no possibility to store or the workflows **(e)**; as it generate's the workflows thought the FF algorithm, the user do not interact with it, so no interactive environment **(g)**; no results interpretation **(h)** no guidance during the manual construction of the workflow **(i)**.

**Charest and Delisle** [56] propose a hybrid assistant by mixing the CBR with a process ontology. The majority of assistants store cases, but leave the user to find and reuse them. It is not enough just to determine the right algorithm for a certain case, but to reuse the past cases and to learn from them. It is not included in a DM tool, it has only just an interface that defines the problem and finds recommendation for certain steps of CRISP-DM process. **Advantages**: As it has an intuitive interface with which the user interacts, the user is helped during the manual construction of the workflow, but by using at the same time a DM tool for the operators **(g)**, **(f)**; it uses CBR for finding solutions in the past experiences, in order to build valid and optimal workflows **(c)**, **(d)**, **(e)**. **Inconveniences**: it doesn't offer result interpretations, but recommendation **(h)**; it is not included in a DM software; it is not available for testing.

**Choinski**'s assistant [51] it is a web application that integrates the operators from Weka. It is based on the ontologies described in the previous subsection, interprets the system's knowledge differently according to the user type. The main difference from the previous assistant is the fact that offers the possibility to generate automatically or create manually the workflows. The workflow composition is similar with the one described by [36], but the ontologies have different structure. It is the only one that mentions Business Understanding or Data Understanding steps from CRISP-DM process, but leaves the user to fill in the business and the DM information, without indications. **Advantages**: automatically generates valid workflows **(a)**, **(c)**; the workflows are automatically stored within the CBR, and has mechanism for searching similar projects and reuse **(e)**; the user interacts with the interface in the first two steps of the process CRISP-DM **(g)**. **Inconveniences**: the generated workflows are linear **(b)**; the workflow can be optimized later by experts **(d)**; offer no support for result interpretation **(h)**; no parallel workflow

[31] http://orange.biolab.si/

| Workflow criteria | DMO | KDDONTO | Zahova | Charest et al. | Choinski |
|---|---|---|---|---|---|
| 1. *Autom. workflow generation & classif.* | ++ | ++ | ++ | – | ++ |
| 2. *Types of generated workflow* | nested | cyclic | cyclic | ? | linear |
| 3. *Workflow validation* | ++ | ++ | ++ | help the user to build workflow – | ++ |
| 4. *Optimal workflow* | DV: Modeling & Evaluation + | - | - | – | - |
| 5. *Possibility to store or to reuse workflow* | – | – | – | CBR + | OCBR ++ |
| 6. *Intuitive environment* | DM tool inv. + | ? | DM tool inv. 0 | site web + | web app.: integrates Weka's ops. ? |
| 7. *Interactive system* | + | ? | - | + | + |
| 8. *Result interpretation* | – | ? | - | – | – |
| 9. *Guidance during manual construction of workfl.* | + | ? | - | autom. gener. workflow + | – |
| 10. *Parallel exec. of the workflow* | – | ? | | ? | – |
| 11. *Software integration* | Rapid Miner ++ | ? | Orange + | ? | Weka ++ |

**Table 4:** Summary table of ways of assisting the user — "–" (very bad), "-" (bad), "0" (insufficient), "+" (good), "++" (very good) and "?" (undefined)

execution **(j)**; don't offer the possibility to create workflows manually **(i)**.

The table 4 summarizes the strengths and weaknesses, for each criteria, of each intelligent assistant.

**Recommendations**:

As it can be seen in the previous table 4, if one considers that the criteria have equal weights, the assistants that meet most of the features are those of [36] (eProPlan) (already included in a free software), and Choinski's assistant [51]. Choinski's assistant offers the possibility to reuse the previous cases (CBR). The main inconvenience of Choinski's assistant is that is not public for testing, while eProPlan is free to use it. So the final choice will be made out of those two options, perhaps to unify their advantages in order to have an intelligent assistant.

## 3.5. Discussion

The recommendations made in this section for the process representation and for a more appropriate way of assisting the user during the DM process, serve for proposing later in the paper a prototype for a general architecture of a DM assistant, starting with the first two steps of CRISP-DM process.

# 4. Our proposal for a Data Mining adviser

This section presents a solution for assisting the user during the DM process. Having defined a KDD process, which is the equivalent "what to do" with a DM project, we want to add "how to do" to the existent tasks in the process. We have chosen CRISP-DM as a KDD process mostly because is the "*defacto standard*" and is the most used DM process. The added assistant to the KDD process will transform the process into a methodology which will help the user to carry out the list of tasks (briefly: *process* = list of tasks; *methodology* = list of tasks + methods to carry out each task).

In order to understand the shortcomings of the KDD processes, in [54] Marban et al. shows that the history of the development of KDD and Software Engineering(SE) is quite similar. Both of them were oriented towards the implementation of algorithms at the beginning. After the Software Crisis [32], because of the 70% failure rate for the software projects [33] (due to the complexity of actual projects, the increasing power of the machines, etc.), SE oriented its attention towards the development of standard processes for software projects. This approach has improved a lot the development of software and lowered the failure rate of projects of about 15-20%[34].

Same as SE, Data Mining oriented it's research towards algorithm implementation, rather than methodology development. This can explain the high failure rate for the DM projects: more than half of all DM projects fail[35]. As we can see in figure 5 there are a lot of steps that are not included in the KDD process, but which are very important to SE, so one possible solution is to offer assistance during the missing steps (or steps that are superficially treated) from the KDD process. The assistance will be offered through the DM tools which allow to follow a specific DM process.

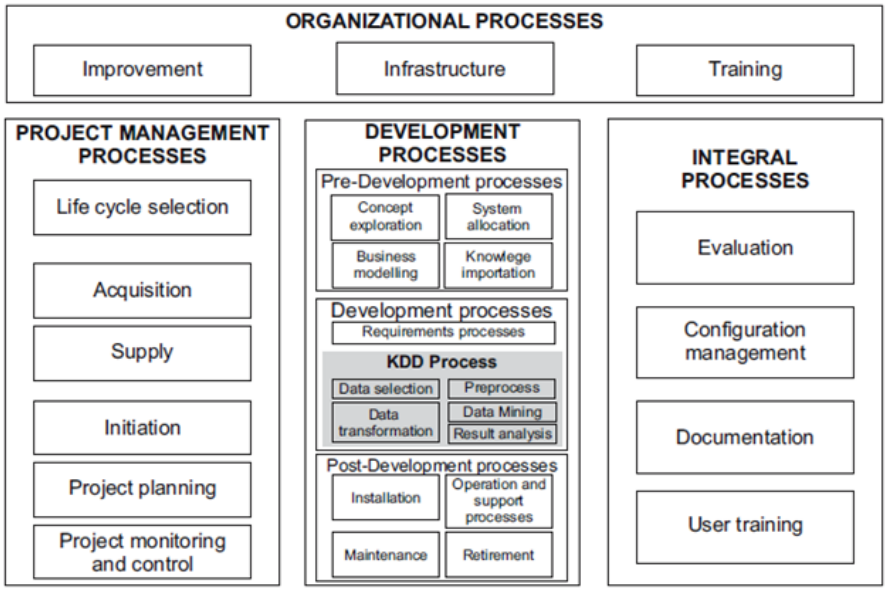


**Figure 5:** Software Engineering vs. Knowledge Discovery in Databases [54]

Starting from the existing concepts in SE[36], the idea is to help the user during the first two steps

[32] http://en.wikipedia.org/wiki/Software_crisis
[33] http://blogs.msdn.com/b/karchworld_identity/archive/2011/04/04/the-software-crisis-how-methodologies-evolved-from-the-influence-of-rework.aspx
[34] http ://www.galorath.com/wp/software-project-failure-costs-billions-better-estimation-planning-canhelp. php
[35] http://www.analyticbridge.com/main/search/search?q=Search+AnalyticBridge&page=4
[36] http://agilemanifesto.org/principles.html

in the CRISP-DM process: Business Understanding and Data Understanding. At the beginning it will be presented a general architecture that constitutes the base of the system and in which it will be included the assistant that will help the user during the CRISP-DM process (the focus is only for the first two steps).

## 4.1. Architecture

Figure 6 presents a general view of the system, in which the central actor is the user. Later in the section each module will be presented with its functionality.

The user interacts with the assistant all the time through the DM tool's interface, in which it is included as a plug-in. He can have access at different types of knowledge depending on his status.

The modules can be divided into: **Knowledge Modules** (Process Knowledge, Data Knowledge, Enterprise Knowledge and Domain Knowledge), **Metadata** (meta-data, process meta-data, business meta-data), **CBR** (Case Base Reasoner), **Plan generator** and **Plan classifier**.

***1. Knowledge Modules***

**Process Knowledge** groups the knowledge from CRISP-DM, Data Mining operations, optimization and knowledge rules. *CRISP-DM process* is represented as an ontology as it was discussed in the previous section and the knowledge is obtained by asking SPARQL[37] queries. *Data Mining Ontology* is based on [49] and gathers methods and algorithms for the majority of CRISP-DM phases. The *Optimization ontology* is described in [52] and it gives the right/optimal algorithms and their parameters. *Knowledge rules* for determining the solutions for certain problems. It makes the link between the CBR and the process ontology. Figure 7 is an overview of the Process Knowledge's modules.

**Data Knowledge** gathers *Enterprise Data* (data sources, recipients, data features, etc.), *Transactional Data Ontology* (relational schema), *Customizing Data Ontology* (from ERP system), based on the work of [50]. This module is built from the existing Data Warehouse. The user is allowed to access certain information based on their status. Figure 8 gives an overview of the main components of this module.

**Enterprise Knowledge** makes part of a substantial work in the past years, dealing with the modeling of business processes. The research was oriented especially towards [50] in which it is explained how all the organizational information are obtained and how they are grouped. Mainly, this module contains: *Organization and Resources Ontology* (*Organizational Ontology* — terminology, resources, organization's structure; *Business Organization Ontology* — tasks, roles, etc.; *Business Resource Ontology* — business resources), *Business Function Ontology* (for modeling different types of business and activities), *Strategy Ontology* (for modeling the strategy of the enterprise, which is the business goal of the organization, the sub-goals of a particular goal, etc.), *Business logic* (business rules and constraints), *Provisioning and Consumption Ontology* (for having an evidence of all costs). As described in [50], this information exists already in the ERP[38] system. So, only a mapping between the ontologies and the ERP structure is needed. It's not the focus of this work to enter more into details with business processes. It is wanted only to justify obtaining and handling business data. In the figure 9 it can be seen the main components of enterprise knowledge module.

**Domain Knowledge** offers support information about different domains : IT, marketing, fi-

[37] http://en.wikipedia.org/wiki/SPARQL
[38] http://en.wikipedia.org/wiki/Enterprise_resource_planning

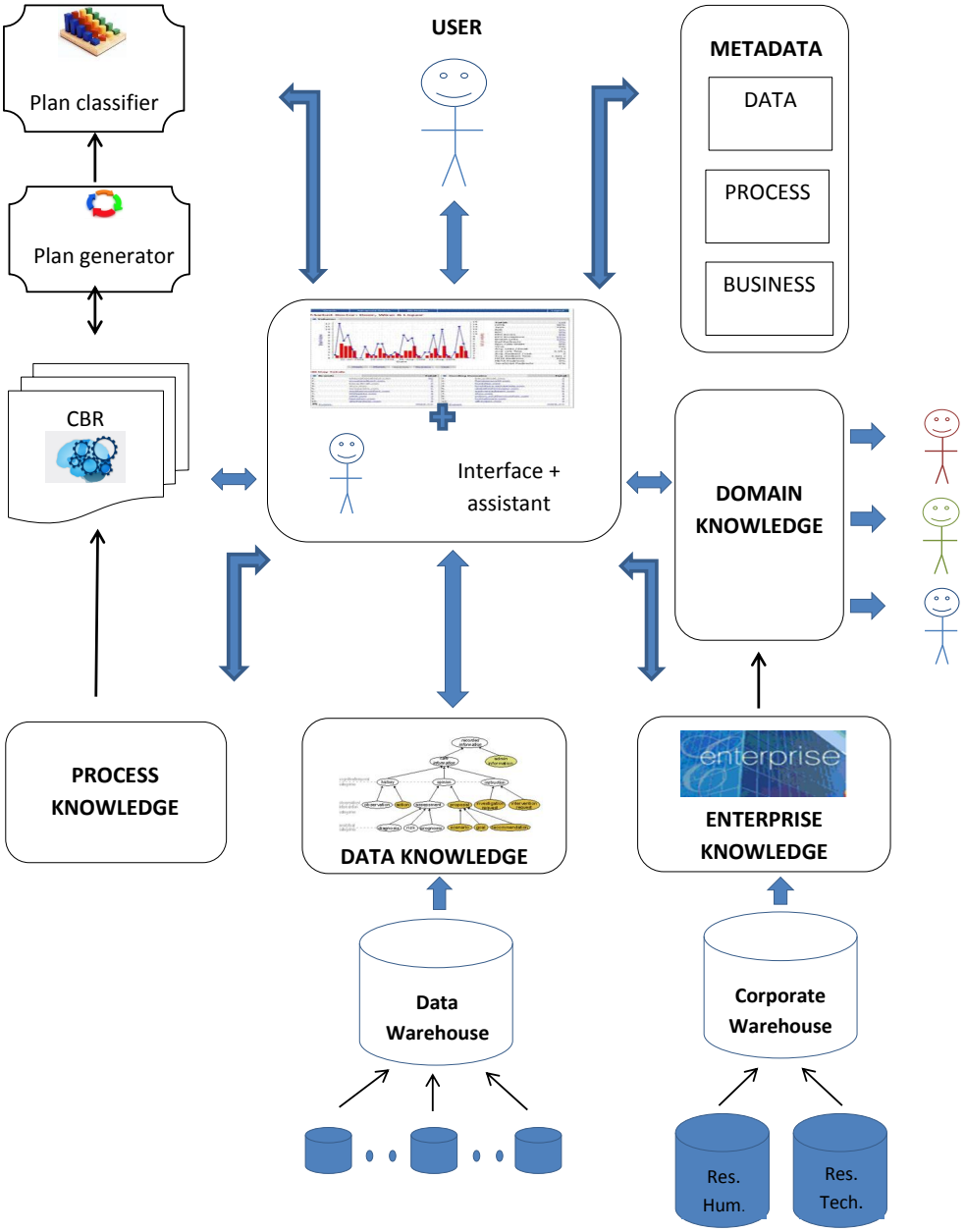


**Figure 6:** General Architecture

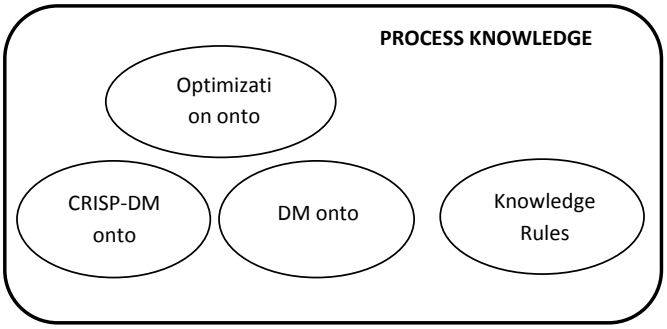


**Figure 7:** Process Knowledge

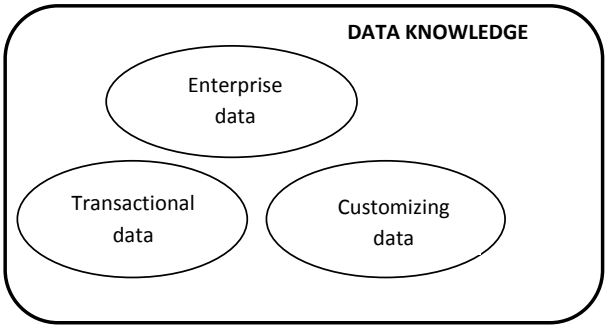


**Figure 8:** Data Knowledge

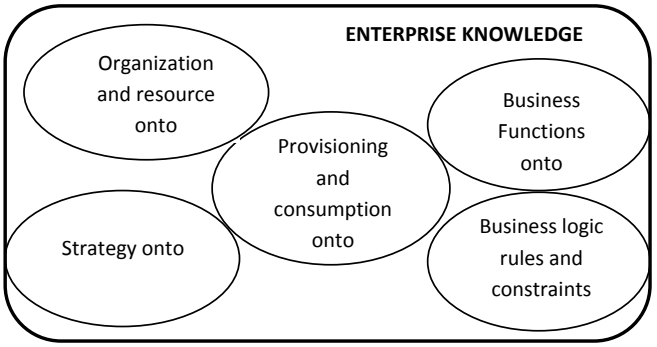


**Figure 9:** Enterprise Knowledge

nancial, etc, as mentioned in [50] and [51]. This can be used in order to show and interpret the results for different types of users using their vocabulary: IT specialists, marketing specialists, financial specialists, etc. The implementation of this module is not the focus of this internship.

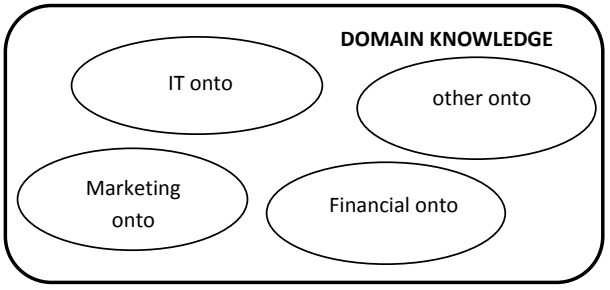


**Figure 10:** Domain Knowledge

***2. Metadata*** contains the metadata of the data, the process and the business are being modeled. The metadata for the data used by the process provides user with information about the characteristics of the data, statistics (mean, variance, minimum, maximum, etc.). The metadata for the process contains information about the current phase, task, activity and it's the information used for guiding the user during its work. Business metadata gives information about the current domain in which the modeled process is included.

***3. Case Base Reasoner*** stores the best models or the ones that the user wants, following the CRISP-DM structure. The stored models can be reused entirely, or for a certain phase, task, situation, etc. (e.g. for a certain problem determined similar with another case, we will use the same operators for the Data Preparation phase.). The Case Base paradigm has been studied more recently, and a good example for our situation is [53], in which CBR is composed of four

knowledge containers: *Vocabulary Knowledge* (the basic elements for representing the knowledge: predicates, operations, functions, etc.), *Similarity Measure Knowledge* (how to determine the similarity between the cases, for the moment, we can be restrained in searching similar cases by using key words), *Adaptation Knowledge* (how to adapt the found solution in the current situation) and *Case Base Knowledge* (stores cases for a certain domain). One important aspect with which a CBR can be enriched is the meta-learning module in order to learn from past cases (by learning problem-solution as we can see in the image 11). For the moment, the CBR will be implemented later, in future works because is not part of the current internship.

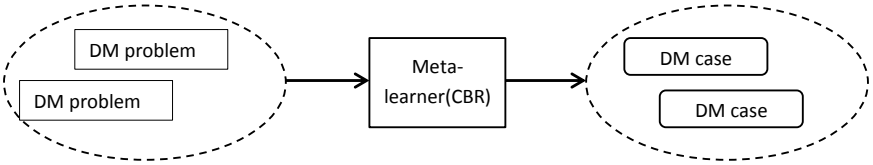


**Figure 11:** Data Mining meta-learning problem[53]

***4. Plan generator***

Having as basis the process knowledge (the process ontologies and the knowledge rules), from the user's objectives, data feature and the organization's objectives, the plan generator will build workflows as described in the previous section.

***5. Plan classifier***

Once the workflows were generated, the user can select certain criteria (the execution speed, the precision, etc.) to classify them or to eliminate the ones that are not important. More details about the plan classifier were presented in the previous section.

***6. Interface***

The interface provides an intuitive environment for the data miners. They can have access to a multitude of operators, data, will be guided in their choices, can build valid plans by themselves or they can be automatically generated.

In order to integrate more the first two steps of CRISP-DM in the organizational projects, the focus of the thesis work is to propose a solution to offer assistance to the data miner during these first two steps of the CRISP-DM process: Business Understanding and Data Understanding. The architecture presented in this section gives all the information needed for building such an assistant.

## 4.2. Exemplification

To offer a better understanding of the architecture, there are proposed some examples in which we identify the inputs, constraints and the corresponding outputs as a result. As it can be observed from the scientific literature, that there is no formal representation for Business Understanding or Data Understanding. Through these examples it will be offered a conceptual representation of these two first steps of CRISP-DM process. This conceptualization will be the first step towards a formalization of the BU and the DU steps.

For the task Determine Business Objectives we will have the following:

```
> Task: Determine Business Objectives
> Input: CBR, EnterpriseKnowledge, ProcessKnowledge
> subtask := currentTask.verifyAccess(user)
> Constraint: if (user doesn't have access to the enterprise
```

```
                data AND no enterprise knowledge changes)
            then begin
            while(similar cases in CBR)
            begin
               *refine NewEnterpriseKnowledge := currentTask.
                nextSubtask(ProcessKnowledge).hasInput(
                user information, EnterpriseKnowledge, CBR)
             end while
               *define BusinessOjbectives := currentTask.
                nextSubtask(ProcessKnowledge).hasInput(
                NewEnterpriseKnowledge)
            end if
            else begin
               *define BusinessOjbectives := currentTask.
                nextSubtask(ProcessKnowledge).hasInput(
                EnterpriseKnowledge)
            end else
>*identify BusinessSuccessCriteria := currentTask.
  nextSubtask(ProcessKnowledge).hasInput(BusinessObjectives,
  EterpriseKnowledge)
>*new CBR case := add.(new CBR(BusinessObjctives,
  BusinessSuccessCriteria))
> Output: BusinessObjectives.
```

It will be taken the example with the determination of the Business Objectives. First, the user should consult the background information, that could be found in the Enterprise Knowledge modules. As not all the users have access to all the organization's information, the first subtask is to verify the user's access to this kind of data. If the user doesn't have access, the NewEnterpriseKnowledge will be created, based on the past similar cases. The user can define now the business objectives. Otherwise, if the user has access to the Enterprise Knowledge, the business objectives will be created having as inputs this knowledge and the CBR. The new subtask of determining the Business Success Criteria is dictated by the Process Knowledge. At the end of this task, the CBR will be updated.

For the Data Understanding step we will have the following:

```
> Task: Data Understanding
> Input: BU, CBR, ProcessKnowledge
> subtask := currentTask.verifyData(BU.BusinessObjectives,
  BU.DMObjectives)
> Constraint:
   if (subtask not verified)
     then begin
        return to BU
     end if
     else begin
      while(similar cases in CBR)
       begin
        *refine initialDataCollectionReport :=
```

```
        currentTask.nextSubtask(ProcessKnowledge).hasInput(BU, CBR)
        *refine dataDescriptionReport :=
        currentTask.nextSubtask(ProcessKnowledge).hasInput(BU, CBR)
        *refine dataExplorationReport :=
        currentTask.nextSubtask(ProcessKnowledge).hasInput(BU, CBR)
        *refine dataQualityReport :=
        currentTask.nextSubtask(ProcessKnowledge).hasInput(BU, CBR)
       end while
        create_updateMetadata :=
        currentTask.nextSubtask(ProcessKnowledge).hasInput(Metadata,
        DataKnowledge)
     end else
>*new CBR case := add.(new CBR(initialDataCollectionReport,
  dataDescriptionReport, dataExplorationReport, dataQualityReport))
> Output: initialDataCollectionReport, dataDescriptionReport,
  dataExplorationReport, dataQualityReport
```

Data Understanding task has as inputs the pas similar cases (CBR), BU and ProcessKnowledge. If the outputs from BU step do not satisfy the business objectives, BU should be redone. If they are satisfied, the DU tasks begin their execution. The tasks are refined until the client's needs are accomplished. At the end, CBR is updated and the outputs are generated.

The information from CBR, Enterprise Knowledge, Process Knowledge could be obtained through simple requests on the knowledge. In the following examples, it will be provided some requests on the proposed ontologies in the system, and show how the information can be retrieved (I don't dispose for the moment of the ontologies, they have to be built by an expert following the indications proposed in the architecture and Section 3.4). The queries will follow the SPARQL grammar.

***Query 1**: Show me all the project names, alphabetically ordered, that contain the keyword "finance".*

```
> PREFIX cbr: <http://something/URI/CBR#>
.....
> SELECT ?x
> WHERE {
> ?proj cbr:name ?x
> ?proj cbr:keyword "finance"
> }
> ORDER BY ?x
```

***Query 2**: Show me all the Vinho Verde enterprise's goals.*

```
> PREFIX entr: <http://something/URI/Enterprise#>
.....
> SELECT ?goal
> WHERE {
> ?goal entr:name "Vinho Verde"
> }
```

***Query 3**: Show me the next subtask(s) and the methods for handling it from previous cases for the step Business Understanding, knowing that the actual task is "Task1" and actual subtask is "Subtask1".*

```
> PREFIX crisp: <http://something/URI/CRISP#>
> cbr: <http://something/URI/CBR#>
.....
> SELECT ?subtask ?method
> WHERE {
> ?subtask crisp:isSubtaskOf Task1
> ?subtask crisp:isNextAfter Subtask1
> ?method cbr:handlesSubtask ?subtask
> }
```

## 4.3. Implementation

The proposed solution is considered to be a plug-in for DM tools. As it can be seen in the figure 12, the assistant connects the knowledge (Process Knowledge, Data Knowledge, CBR, Enterprise Knowledge) with the DM tools (DM operators). So, the architecture presented in the figure 6 can be summarized to the architecture from the figure 12.

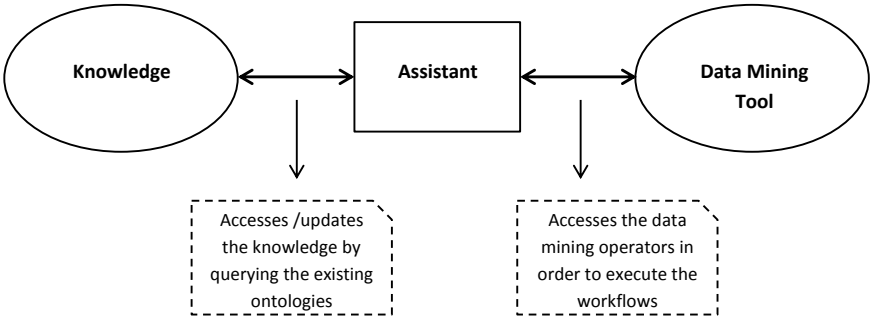


**Figure 12:** Simplified Architecture

It's not the focus of the thesis to structure the knowledge into ontologies, but to propose recommendations for this, to implement the assistant by connecting it to a DM tool and to the knowledge modules, and finally to present it as use cases.

The implementation of the assistant is a work in progress. For the moment, there are implemented the connection with the knowledge models and queries to the included ontologies can be made, in order to guide the user, by proposing him appropriate advices. The next step is to include the DM tool and to have access at all the operators included in the tool.

TobBraid Composer[39] (developed under the Eclipse framework) was chosen as framework for the development of the assistant and Java as programming language, because the most used DM tools were implemented in this language. The operators from Weka environment ware chosen as DM operators. For the knowledge representation it will be used the implementation of the ontologies in OWL-DL[40] (Description Logic) as "maximum expressiveness while retaining computational completeness" is required. TopBraid Composer has the advantage of including in the same framework different perspectives: *Java* — in which it can be implemented as in Eclipse framework, *TopBraid* — which allows to build visually OWL-DL ontologies with the possibility to include the access to web services[41] through OWL-S[42] (see the recommendations in Section 3) and *Plug-in development*. It can switched between them keeping the same window opened.

[39]http://www.topquadrant.com/products/TB_Composer.html
[40]http://www.w3.org/TR/2004/REC-owl-features-20040210/#s1.2
[41]http://en.wikipedia.org/wiki/Web_service
[42]http://en.wikipedia.org/wiki/OWL-S

As OWL-DL contains logic, it is needed to include a reasoner. It has been chosen, in this case — as we are using OWL-DL for describing ontologies, Pellet reasoner[43], an open-source Java based reasoner. The application contains also Jena framework for querying the ontology. The queries respect the SPARQL[44] grammar, as the ontology follows the OWL-DL structure.

The current implemented functionalities are:

- **Load Ontology** — the user has the possibility of one or more ontologies. This will allow to interact with the knowledge modules (see the Subsection 4.1). The ontology could be loaded from a local file (`loadOntologyFromfile()`) or from a URI address (`loadOntologyFromURI()`).
- **Remove Ontology** — allows the user the possibility to remove a certain ontology with which the assistant interacts.
- **Classify Ontology** — the loaded ontology could be classified using the Pellet reasoner.
- **Create Query** — the user has the possibility to create his own SPARQL queries.
- **Execute Query** — the user can execute his own queries or the predefined ones.

Further implementations:

- **Generate advice** — having to execute a certain task, and having the results of the query on the existing knowledge, the system has to be able to generate appropriate advices.
- **Assistant's interface** — the system should be able to guide the user through the manual construction of the workflow or to generate automatically appropriate workflows to a certain situation. As we want to build a plug-in for DM tools, it will be used their interface as a support for the offered help.
- **DM tool integration** - building the connection to a DM tool, in order to access their operators. As a starting point Weka will be used.

The current focus is in offering a conceptualization for a methodology (in our case process + assistant), to build the "how to" around the process in order to help the user during the CRISP-DM process (the proposed architecture), to identify the steps from CRISP-DM that are less treated (most of the focus in the scientific literature is in offering help during the Modeling, the Data Preparation or Evaluation steps), to offer a conceptualization for Business Understanding and Data Understanding, to justify it through a case study and, finally to start the development.

## 4.4. Discussion

As presented in the previous section, Kietz et al. in [56], Charest et Delisle in [55], Choinski et al. in [51] present solutions for DM assistants, but they don't justify the data and information acquisition (the informations obtained from IS and ERP) and they do not offer assistance in the first two steps of CRISP-DM process. This solution, instead, tries to justify it's integration as a plug-in between the organization and the DM tool, being built in conformance with the characteristics defined in the previous section 3.4. The interconnection with the organization's IS in the both directions: the assistant gets the information from the IS and the obtained results or the new informations obtained update the knowledge from the IS. In addition, this contribution

[43] http://www.mindswap.org/2003/pellet/
[44] http://fr.wikipedia.org/wiki/SPARQL

helps the user during the first two steps of the CRISP-DM process, steps that are usually superficially treated. The usefulness of the presented solution will be proved with the help of a case study during the next section.

# 5. Case study

In order to prove the usefulness of this solution, it is proposed to have a look at the example presented in this section. This study case is taken from CrowdAnalytix[45] contest. The challenge of this year is to help the wine company "Vinho Verde" from the north of Portugal, to improve the process of wine certification and quality assessment. In order to apply the proposed solution to this example, I will go further and try to identify the root of the problem. I will go mainly through the first two steps of CRISP-DM process on which it will be added the assistance proposed in Section 4, in order to refine the process adding the part that is missing: "**how to carry out**" the different tasks proposed by this process.

Portugal has become one of the top ten countries that export quality wine [57] with a market share of 3.17% in 2005[57]. Exports of Vinho Verde increased with 36% from 1997 to 2007 [58] and with 14% in 2009 [59]. To systematically increase the market and to export more wine, Vinho Verde needs wine certification in order to avoid the falsification of the brand and a better quality assessement in order to improve wine production. Therefore, having business information about the situation, will help to set down more clearly the business and DM objectives.

For more informations about the contest and Vinho Verde, please check the site[57].

## 5.1. Motivation

The fact that this example is presented in this year's contest, is one of the reasons for which it was chosen for study. Another important reason is that we believe that improving the first two steps of the DM process, it can be improved the final result of the project. I consider, as it was said in the previous sections, that the first two steps are influencing the trajectory of the project. Bad decisions in this part of the project will produce unfavorable results in the next step, which will require different types of interventions: try to find new solutions, go back to previous steps in the process and try to apply new solutions, etc. Either way, it will make the organization to spend more time and invest more money for the success of the project.

## 5.2. Exemplification

For a better clarification of the solution proposed in Section 4, the example described in [57] will be adapted it for this case. As the system is an ongoing work, the case study presented here will help to strengthen its validity and usefulness.

The case study is organized as follows: in 5.2.1 it will be made an introduction of the proposed system and afterwards the logic diagrams representing Business Understanding in 5.2.2 and Data Understanding in 5.2.3 are followed for exemplification.

### 5.2.1. Overview

As it was presented in the previous sections, the advisers proposed by Choinski et al. in [51], Kietz et al. in [56], Charest et Delisle in [55] don't guide the user for the first two steps of a DM process and I try to extend their work in order to offer guidance from the beginning of the process. Taking the example described in 5, we'll try to go through all the tasks proposed by the process (for the first two steps of the process CRISP-DM).

***Situation*** : As described in [62], the wine industry represents 15% of the total Portugal's production, and only 10% of the produced wine is exported. The wine company Vinho Verde wants to expand the market with more that 5%, so the wine company decides to invest more into

[45] http://www.crowdanalytix.com/

technology and to develop new strategies for improving their wine quality, their production and to certificate their wine.

***Moving to action*** : Before turning this problem into a DM problem, the user will try to formalize all the information he has with the help of the DM assistant integrated within a DM tool.

Once it was opened or created a new project, the assistant is ready to use. Let's say that the project's name is "Wine project - Vinho Verde". The user can now be guided through all the stages during the project. The guidance is in conformance with the figure 24 described later in the section. For offering help to the users, it is proposed in the figures 24 and 25 the logic diagrams that represent the flow of tasks and actions to be executed during the Business Understanding and Data Understanding steps. Further, the flow of tasks and actions will be presented on smaller parts (zoomed) in order to offer a more detailed view.

### 5.2.2. Business Understanding

***Business Objectives***

As each project can be identified at the beginning with key words, we can access the history of the other projects to ease our work. So for example, in this case, if we type "wine" we'll obtain as a result a list of all projects that contain the word "wine". In our case, let's suppose that we find a project named "Wine certification - Vinho Verde".

*1. Background information*

As it was said before in the previous section, the system developed it is a plug-in that will make the link between the enterprise's (with the help of Enterprise Knowledge) Information System (IS) and the DM tool. If it can't be accessed some of the business information (security problems, non-existence of such information, etc.), it's up to the user to introduce some of the background knowledge about the organization's business situation at the beginning of the project, having as suggestions the previous cases (CBR).

The figures 13 and 14 show how to view the background information of *"Vinho Verde"* with the possibility to add new information that can be used in the next stages.

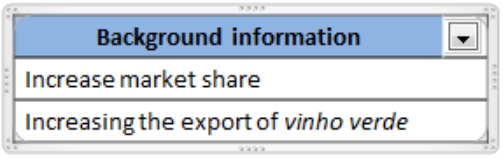


**Figure 13:** Viewing the backgroung information of Vinho Verde

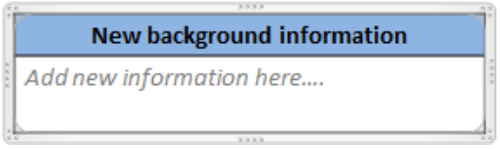


**Figure 14:** Adding new background information

*2. Determine the business objectives*

Having the background information about the strategies and the objectives of the enterprise from Enterprise Knowledge (Strategy) and the precedent cases from CBR, the user can determine and fill in the business objectives for the project. For this case, the determined business objectives are showed in the figure 15.

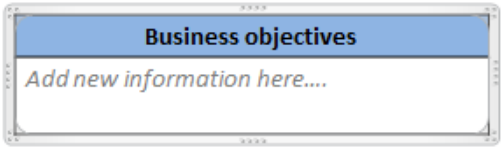


**Figure 15:** Business Objectives

*3. Determine business success criteria*

A well-formulated business objective is a measurable one. It will be used the principle GQM (Goal, Question, Metric) as in [60] and [61], for defining the measure for the business objective. In our case, increasing the export percentage with 5% respects the metric.

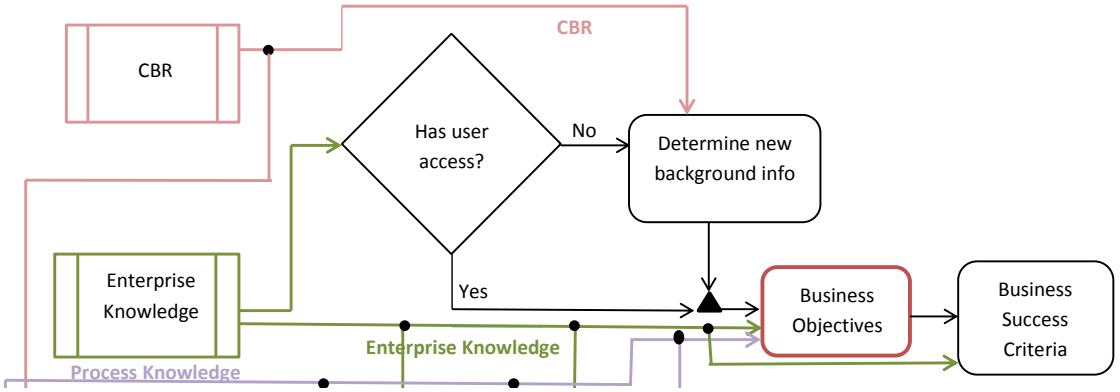


**Figure 16:** Determine Business Objectives

Figure 16 synthesizes the task of determining the business objectives. After this task, CRB will be updated, by modifying the existing case, or by creating a new case. This situation is not depicted in Figure 16 as the figure is too detailed.

***Assess situation***

*4. Resource inventory*

Using the modules past cases (CBR - Case Base Reasoning) and the Enterprise Knowledge (Organizational and Resources, Business Functions) the list of human and technical resources will be automatically generated (the human resources will be distributed in the corresponding business group: IT, Financial, Marketing, etc.). For example in this case, for the human resources will be retrieved a list of employees and experts along with their role, qualification, availability, income, history etc., with all the useful informations for the project.

*5. Determine the useful data*

One of the important tasks of this step is to determine which are the required data to model in the next stages. Normally the data is spread in the IS and the expert has to determine which ones have significance for this project. The data in the proposed system are located in the Data Knowledge and we can obtain them only through simple requests asked by the expert. Based of the composition of the wine and in conformance with CrowdAnalytix, the needed information is as follows: type, fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol content. The class attribute has to be the quality of the wine.

*6. Requirements, assumptions and constraints*

The user has to determine the requirements, the assumptions and the constraints for the project, by filling in the form presented in figure 17. All the requirements, assumptions and constraints should be consulted with business personnel, and in addition to [60] our solution offer

the possibility to consult similar past cases (CBR), Enterprise Knowledge (Business logic rules and constraints, Strategy, Provisioning and Consumption and Organizational and Resources) to help the user in building and refining these lists.

An example of constraint for "Vinho Verde" could be the amount of money that can be spent on the project. Supposing that all the departments of the enterprise should reduce their costs with 10%, the amount of money for the project will be reduced by 10%. As a direct consequence is the reduction of working time for the project or/and the reduction of employees allocated for working on it.

| Wine project - *vinho verde* | | |
|---|---|---|
| **Requirements** | **Assumptions** | **Constraints** |
| *Add new information here....* | *Add new information here....* | *Add new information here....* |

**Figure 17:** Requirements, Assumptions and Constraints Form

*7. Risks and contingencies*

The constraints, the requirements and the assumptions defined before, CBR and the data informations will help the user to define the risks and the contingencies for the entire project, which will lead to determine ways of controlling the risks. The new information obtained will probably generate new knowledge and the module Knowledge Rules will be updated.

A contingency risk for the project is the weather, which can influence the quality of wine and implicitly, the export. Bad weather, means lower quality of wine, and as a consequence a reduction of the amount of exported wine and losing the invested money. The risks can be easily added, by filling in the form presented in the figure 18.

| Wine project - *vinho verde* | | |
|---|---|---|
| **Risk** | **Impact** | **Solution** |
| *Add new information here....* | *Add new information here....* | *Add new information here....* |

**Figure 18:** Risks and contingecies

*8. Terminology*

In this step there are defined the DM (Data Mining) terminology and a glossary for business terms in order to create/modify Business Metadata. We can obtain most of the terminology from the Enterprise Knowledge module, Domain Ontologies and CBR which can be entirely or partially reused for this step. The problem still remains for the new added terms, because this cannot be done automatically by the user. The terms can be introduced (name of the term, definition, type, etc.) by the user, but the ontologies (the knowledge domain) will be updated afterwards by an expert. In the figure 19 we can see a short example of how to introduce a new term for this project (the definition of the term was taken from [46]).

*9. Costs and benefits*

At this stage the costs and the benefits of the project need to be defined: costs of data collection (the types of wine, their composition, their quality, all the attributes that are necessary for building a good model), costs for implementing the solution (personnel: persons that taste the wine, developers, etc., resources techniques), risk costs (to estimate the loss for the risks) and to estimate the benefits of the success of the project. The information about the terms, costs,

[46] http://www.powerhomebiz.com/Glossary/glossary-A.htm

| Wine project - *vinho verde* | | |
|---|---|---|
| **Terminology** | | |
| **Name** | **Definition** | **Type** |
| Amortization | To liquidate on an installment basis; the process of gradually paying off a liability over a period of time | Business Term |
| *Add new information here....* | *Add new information here....* | *Add new information here....* |

**Figure 19:** Terminology

organization, risks, inventory, etc; can be retrieved from the Enterprise Knowledge, as can be seen on the figure 20. The task chaining is dictated by the process Process Knowledge — the rules, the ontologies, etc.

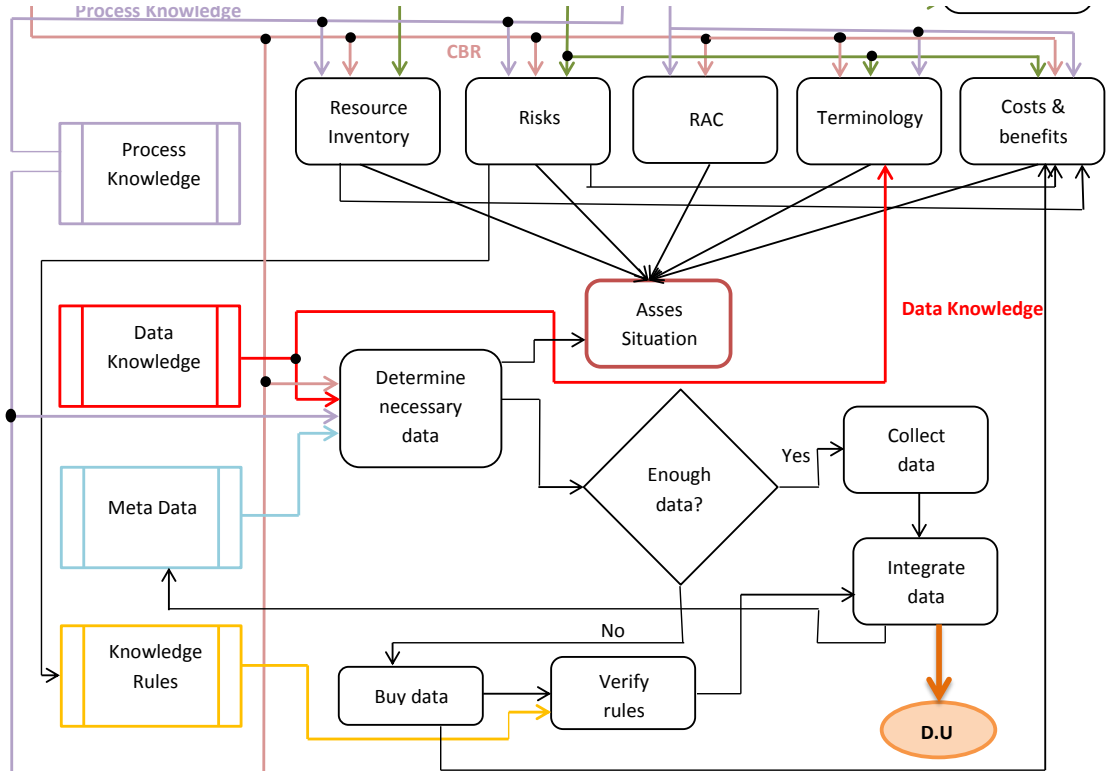


**Figure 20:** Assess Situation

In the figure 20 it can be seen all the subtasks, how they are chained and interconnected with the elements of the system. After all the data is obtained, the module Metadata will be updated with the new information and the data will be examined more in the next phase, Data Understanding. At the end of this task, CRB will be updated, by modifying the existing case, or by creating a new case. In figure 20 it is not represented this situation, because the figure it is already too detailed. The terms acronyms can be consulted in Appendix A.

***Determine the Data Mining objectives***

*10. Data mining objectives*

The main objective is to translate the business objectives into DM objectives. As usual, we

have the possibility to look at the past similar cases. It is considered that in the CBR we have as DM objectives the cases presented in [62], for instance "classifying 3 sensory attributes" and "using the mineral characteristics to discriminate 54 samples into red wine classes". The results obtained for theses situations are 94%, respectively 95%. The DM objective is presented in the figure 21.

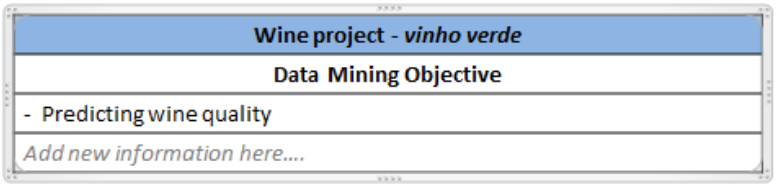


**Figure 21:** DM Objective

*11. Data Mining success criteria*

The accuracy was chosen as a DM success criteria.

*12. Determine the data mining techniques*

Having the DM knowledge, the DM objectives, the DM criteria and the past similar cases (CBR) it can be established which data mining techniques to use. In [62] (and implicitly in our CBR) they used classification models, such as NN (Neural Networks) and SVN (Support Vector Machines). The used models with their parameters can be seen in CBR. As demanded in the CrowdAnalytix[47], it is needed to propose a predictive model that can identify the overall quality of the wine.

The figure 22 shows the chaining of activities for the determination of DM objectives. The results obtained for this step will be needed in the Modeling phase. At the end of this task, CRB will be updated, by modifying the existing case, or by creating a new case. On the figure 22 it is not represented this situation, because the figure it is already too detailed.

***Produce the project plan***

The final output for this phase is the generation of the working plan and throughly documenting all the business problems that were discussed before (see figure 23).

The figure 24 will synthesize the all the activities and the actors involved in this process and their interactions.

[47] http://www.crowdanalytix.com/contests/cheers-predict-wine-quality/

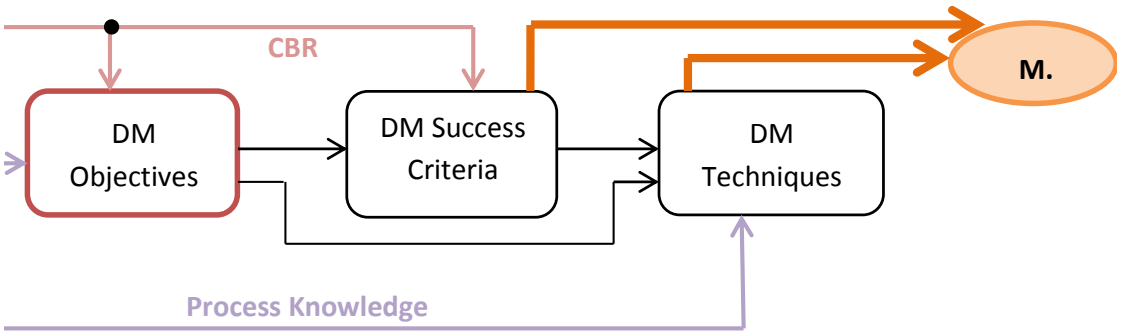


**Figure 22:** Determine Data Mining Objectives

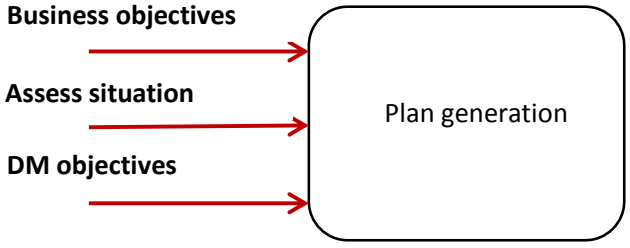


**Figure 23:** Plan Generation

### 5.2.3. Data Understanding

Once it was collected data from the previous phase and with the help of CBR, a deeper analysis of all the data is made. It will be started by doing the **initial data report**, in this case by explaining how the data was obtained. As we simulate the case of a wine enterprise, by analyzing the quality of their wine, it wasn't necessary the external data (all the data were stocked in tables in the data warehouse). Having all the information the user needs, the user can automatically do the **data report** (describe the format of the data - *.xls in our case, the number of examples - 4548 for our case, etc.). The user can easily **explore the data** with the help of the DM tool in which the assistant is integrated. The user can visualize the data, make queries, etc. The key attribute for this case is the quality of the wine, which is important for the construction of the prediction model. Once done the data quality report, all the meta data and data will be updated and all the informations will be stocked in CBR for the current case. All the activities presented here are a preparatory stage for Data Preparation phase.

Figure 25 gives a schematic view of the DU and how it interacts with the system's modules. At the end of this task, CBR will be updated, by modifying the existing case, or by creating a new case (not presented on the figure 25).

### 5.2.4. Discussion

In this section it was exemplified how the system proposed can be used in an organization's project, even though not all the activities have been treated. The idea was to understand how the system works across an organization.

In [60] Sharma proposes a similar description of the CRISP-DM project, but without including the different parts of the organization's IS which are important to prove the way this modules interact with the assistant. This solution shows what parts of the organization's system are implicated in the chain of activities, how the assistant integrates and guide the user.

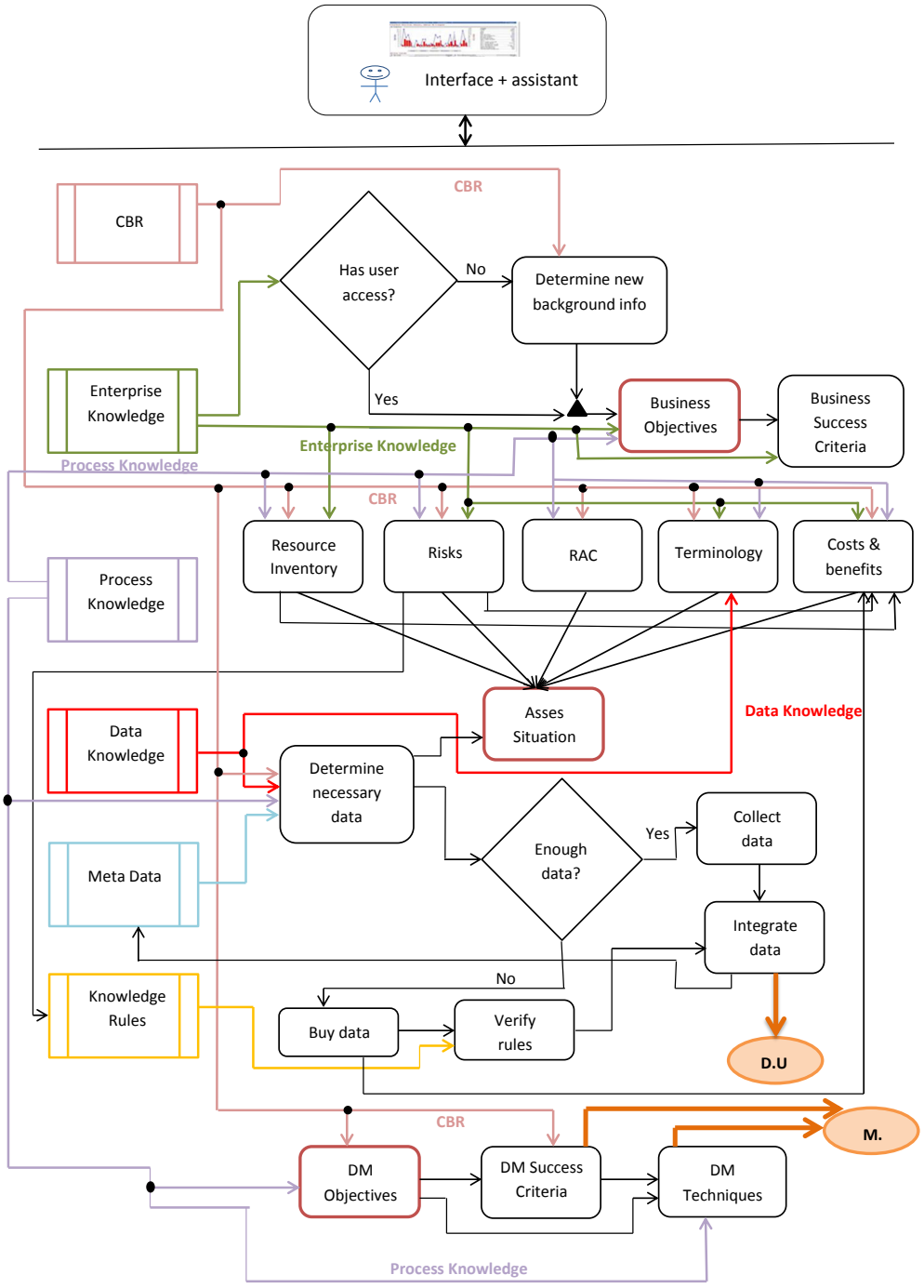


**Figure 24:** Business Understanding

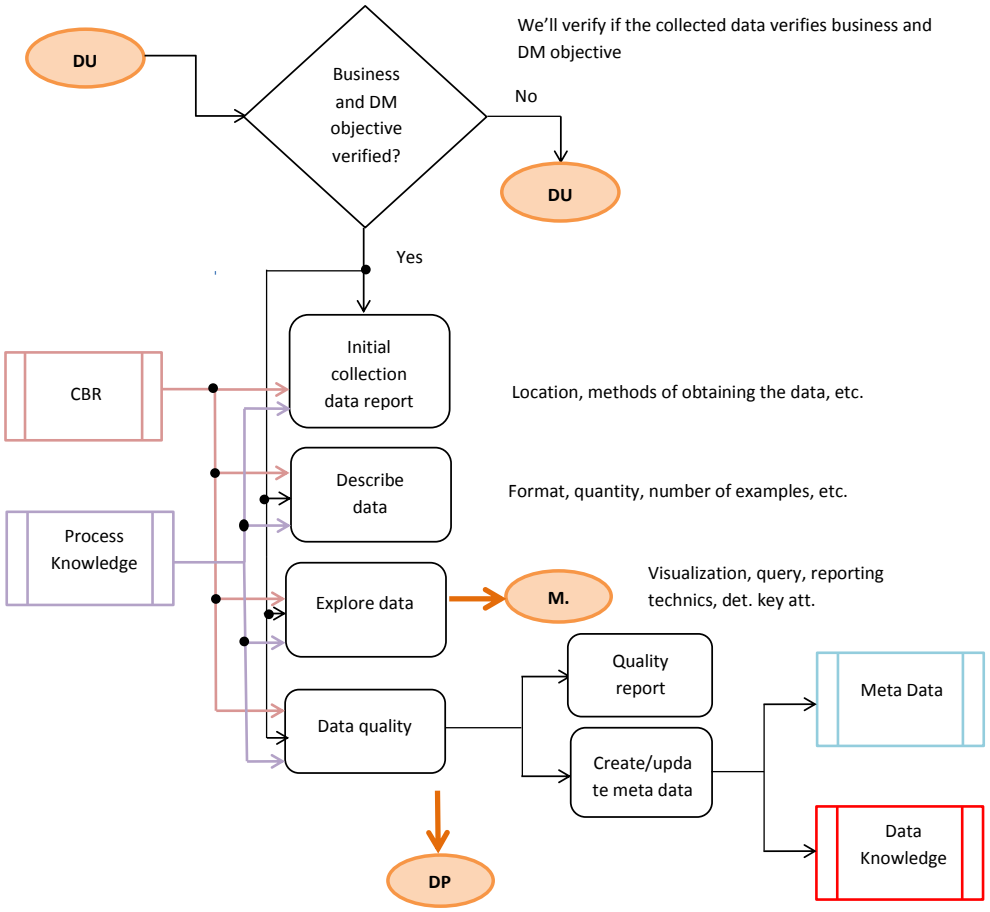


**Figure 25:** Data Understanding

# 6. Conclusions and future works

As it can be observed, because of the lately computer science progress (high-tech computers, high-speed processors, increased memory, etc.), the amounts of data invades our daily life (internet, extranet, locally, etc.). Of course, none of this data could be modeled (pattern extraction, interpreting the results) without the help of DM field. As Cios said: "Finding a good model of the data, which at the same time is easy to understand, is at the heart of DM" [11].

On one hand, DM became very mature because of the multitude of implemented algorithms. In the last decades the focus was oriented towards the development of new, more complex algorithms, in order to handle the new type and dimensions of the data. Only DM tool Weka includes more than 500 algorithms in its framework. Paradoxically, the user is confused by having this wide range of operators, without any real assistance.

On the other hand, DM can not handle the exponential growth of data, and cannot meet the client needs and objectives. This can explain the high failure rate for the DM projects: more than half of the projects end by being abandoned[48].

[48] http://www.analyticbridge.com/main/search/search?q=Search+AnalyticBridge&page=4

## 6.1. Contributions

Having these inputs, the current master thesis aims to propose a prototype for user assistance, targeting the first two steps of CRISP-DM process. In order to arrive to a solution, the following steps were realized:

**First**, the objective of my work was to make a state of the art of existing processes in DM and to analyze the second generation of DM tools (by instantiating the processes through these systems), in order to determine their shortcomings.

**Secondly**, the analysis continued at a different level. It was defined first a prototype of a "good" methodology. It is considered that a methodology is the mixture between a **process** (because it gives the tasks that should be performed) and **assistance** (which will add the complementary information - how to perform the proposed tasks). It was chosen CRISP-DM as a process to be represented, and were defined criteria for comparing ways of process representation. After, taking into account the assistant, there were defined criteria for comparing ways of offering assistance during the DM process. At the end of each comparison, recommendations were made, in order to define a skeleton for the future prototype and to make the passage towards the third DM systems.

**Third**, as a prototype, was proposed a general architecture as a basis for a DM assistant. The considered assistant should be developed as a plug-in for DM tools which allows to use the organizational and business knowledge. Based on organizational studies, there were identified the implicated modules in the architecture and the interaction between them. As the current solutions do not take into consideration helping the user during the first two steps of the CRISP-DM process (Business Understanding and Data Understanding) my work was concentrated in offering a conceptualized vision of the first to steps of CRISP-DM, which constitutes a contribution towards their formal representation. The development of the plug-in was started, by implementing the interconnections with the knowledge (organizational ontologies, process ontologies, business ontologies, etc.) and the DM tool (Weka operators). The main problem encountered in this phase was the fact that interesting existing solutions for process representation (as Choinski's or Charest's) were not available for test and the referenced articles and reports towards a more detailed representation were not found.

**Fourth**, the usefulness of the proposed solution was evaluated through a case study on the Vinho Verde organization, by following the chain of tasks of CRISP-DM process and making use of the modules integrated within the architecture. The case study aims to show how the system interacts with the organization's knowledge and with the end-user.

## 6.2. Future works

As for the moment, building the knowledge didn't make part of the thesis work, it should be interesting to build the process ontologies and populate them by following the given recommendations and to include the indicated modules within the system.

The proposed assistance should be extended to the other steps of CRISP-DM process, in order to improve the quality of the results.

It would be interesting to implement the CBR and initialize it with remarkable cases. Also, the engine for searching pas similar cases, should be also developed. The construction itself could be the subject of another thesis.

The proposed solution imitates the comportment of a methodology (instance of a process, gives the tasks that should be executed and indicates how to execute them), by putting together the process representation and the method for assisting the user during the DM project. The solution is integrated within a DM tool. As the user will be assisted from the first steps of the

project and be advised in all the tasks on "how" they should be executed, the number of projects that meet the client's needs increases. So, the number of abandoned projects will be reduced and through this thesis, the questions ask in the Section 1 could be answered.

The validation of the solution should be done by exemplifying it through a case study for a real organization which will adopt the proposed system. The organization will dispose of all the modules presented in this paper, so CBR will contain previous projects with remarkable solutions. As the end-user for the current project will be guided through the entire process by being advised and by presenting him previous solutions adapted for this case, the results should be at least as satisfying as the previous cases.

As this thesis has big goals, the list of possible future works is not yet complete. The researches led our way from a general topic "Methodologies and processes in KDD" to more punctual objectives: enriching DM processes in order to reduce the fail rate in DM projects. As Neil Armstrong said: "Research is creating new knowledge", this research paper will lead towards new horizons of possible solutions.

# References

[1] Mateyaschuk J. : *The 1999 National IT Salary Survey: Pay up*, Information Week, 1999
http://www.informationweek.com/731/salsurvey.htm.

[2] *Emerging Technologies That Will Change the World*, Technology review, Published by MIT, 2001.

[3] Fayyad U., Piatetsky-Shapiro G., Smyth P., Uthurusamy R. : *Advances in Knowledge Discovery and Data Mining*, MIT Press, 1996.

[4] Parsaye K., Chignell M., Khoshaan S., Wong H.,: *ntelligent Databases ; Object-Oriented, Deductive Hypermedia Technologies*, John Wiley & Sons, 1989.

[5] Fayyad U.M., Piatetsky-Shapiro G., Smyth P. : *From data mining to knowledge discovery : an overview*, The MIT Press, vol. Advances in knowledge discovery and data mining, 1996.

[6] Mariscal G., Marban O., Fernandez C. : *A survey of data mining and knowledge discovery process models and methodologies*, The Knowledge Engineering Review, vol. 25, no. 2, 2010.

[7] Kurgan Lukasz A., Musilek P. : *A survey of Knowledge Discovery and Data Mining process model*, The Knowledge Engineering Review, vol. 21, no. 1, 2006.

[8] Chapman P., Clinton J., Kerber R., Khabaza T., Reinartz T., Shearer C., Wirth R. : *CRISP-DM 1.0. Step-by-step data mining guide*, NCR Systems Engineering Copenhagen (USA and Denmark), DaimlerChrysler AG (Germany), SPSS Inc. (USA) and OHRA Verzekeringen en Bank Groep B.V (The Netherlands), 2001.

[9] Simon P. : *Tech Transformation: What's Holding You Back?*, Business Finance, http://businessfinancemag.com, 2011.

[10] Yang Q., Wu X. : *10 Challenging problems in Data Mining research*, Journal of Information Technology and Decision Making, World Scientific Publishing Company, vol. 5, 2006.

[11] Cios K.J., Pedrycz W., Swiniarski R.W., Kurgan L.A : *Data Mining: A Knowledge Discovery Approach*, Springer, 2007.

[12] Fayyad U. M., Piatetsky-Shapiro G., Smyth P. : *From data mining to knowledge discovery: an overview*, Book: Advances in knowledge discovery and data mining , American Association for Artificial Intelligence Menlo Park, CA, USA, 1996.

[13] Wikipedia: *Exploration de données*
http://fr.wikipedia.org/wiki/Exploration_de_donn%C3%A9es consulted on: 06/08/2011.

[14] Maimon O., Rokach L. : *The Data Mining and Knowledge Discovery Handbook*, Springer, Tel Aviv , 2005.

[15] Brunk C., James K., Kohavi Mountain Viez R. : *MineSet : An integrated System for Data Mining*, KDD-97 Proceedings, AAAI, 1997.

[16] Anand S.S., Patrick A.R., Hughes J.G., Bell D.A. : *A Data Mining Methodology for Cross Sales*, Knowledge-Based Systems, vol. 10, 1998.

[17] SPSS Website
http://www.spss.com consulted on: 21/09/2010.

[18] Kurgan L., Cios K., Tadeusiewicz R., Ogiela M., Goodenday L. : *Knowledge discovery approach to automated cardiac SPECT diagnosis*, Artificial Intelligence in Medicine, vol. 23, no. 2, 2001.

[19] Daimler-Chrysler Project Overview, *CRISP-DM*
http ://www.crisp-dm.org/Overview/index.html consulted on 07/09/2010, 1996

[20] *Data Mining and the Case for Sampling*. SAS,
http ://sce.uhcl.edu/boetticher/ML_DataMining/SAS-SEMMA.pdf, consulted on 07/09/2010.

[21] Marban O., Mariscal G., Fernandez C : *A survey of data mining and knowledge discovery process models and methodologies*, The Knowledge Engineering Review, Cambridge University Press, Cambridge, vol. 25, 2010.

[22] Kurgan L. A., Musilek P. : *A survey of Knowledge Discovery and Data Mining process model*, The Knowledge Engineering Review, Cambridge University Press, Cambridge, vol. 21, no. 1, 2006.

[23] Piatetsky-Shapiro G. : *Data mining and knowledge discovery: The third generation*, Foundation of Intelligent Systems in the 10th International Symposium, 1997.

[24] SAS Enterprise Miner, *SEMMA*,
http://www.sas.com/offices/europe/uk/technologies/analytics/datamining/miner/semma.html consulted on 07/09/2010.

[25] Rakotomalala R. : *TANAGRA : un logiciel gratuit pour l'enseignement et la recherche*, Actes de EGC'2005, RNTI-E-3, vol. 2, 697-702, 2005.

[26] Khabaza T., Shearer C. : *Data Mining with Clementine*, IEEE Colloquium on Knowledge Discovery in Databaes, IEEE Digest vol. 21, London, 1995.

[27] Rokotomalala, R. : *K-Means - Comparaison de logiciels,* http ://tutoriels-data-mining.blogspot.com/search/label/Classification-Clustering, 2008, consulted on 21/01/2011.

[28] Hidalgo M., Menasalvas E., Eibe S. : *Definition of a Metadata Schema for Describing Data Preparation Tasks*, ECML Workshop on third generation data mining: Towards service-oriented knowledge discovery (SoKD-2009), Bled, Slovenia, 2009.

[29] Kietz J. U., Serban F., Bernstein A., Fischer S. : *Towards cooperative planning of data mining workflows*, ECML Workshop on third generation data mining: Towards service-oriented knowledge discovery (SoKD- 2009), Bled, Slovenia, 2009.

[30] Vanschoren J., Blockeel H. : *Stand on the Shoulders of Giants: Towards a Portal for Collaborative Experimentation in Data Mining*, ECML Workshop on third generation data mining: Towards serviceoriented knowledge discovery (SoKD-2009),Bled, Slovenia, 2009.

[31] Diamantini C., Potena D. , Stort E. : *KDDONTO: An Ontology for Discovery and Composition of KDD Algorithms*, ECML Workshop on third generation data mining: Towards service-oriented knowledge discovery (SoKD-2009), Bled, Slovenia, 2009.

[32] Podpecan V., Lavrac N., Kok J. N., de Bruin J. : *Preface*, ECML Workshop on third generation data mining: Towards service-oriented knowledge discovery (SoKD-2009), Bled, Slovenia, 2009.

[33] Zakova M., Podpecan V., Zelezny F., Lavrac N. : *Advancing Data Mining Workflow Construction: A Framework and Cases using the Orange Toolkit*, ECML Workshop on third generation data mining: Towards service oriented knowledge discovery (SoKD-2009),Bled, Slovenia, 2009.

[34] Vanschoren J., Blockeel H., Pfahringer B., Holmes G. : *Organizing the world's machine learning information*, Communications in Computer and Information Science vol. 2008, no. 17, 693-708, 2008.

[35] Grigoriev P. A., Yevtushenko S. A. : *Elements of an Agile Discovery Envoronment*, Darmstadt: Springer-Verlag, 2003.

[36] Kietz J.-U., Serban F., Bernstein A. : *eProPlan : A Tool to Model Automatic Generation of Data Mining Workflows*, In : ECML Workshop on third generation data mining: Towards service-oriented knowledge discovery (SoKD-2010), Barcelona, Spain, 2010.

[37] Hilario M., Kalousis A., Nguyen P., Woznica A. : *A Data Mining Ontology for Algorithm Selection and Meta-Mining*, ECML Workshop on third generation data mining: Towards service-oriented knowledge discovery (SoKD-2009), Bled, Slovenia, 2009.

[38] Panov P., Dzeroski S., Soldatova L. N.: *OntoDM: an ontology of data mining*, Proceeding on ICDM workshops 2008, Los Alamitos (California), Washington, Tokyo: IEEE Computer Society Conference Publishing Services, 752-760, 2008.

[39] Serban F., Kietz J. U., Bernstein A. : *An overview of intelligent data assistants for data analysis*, 3rd Planning to Learn Workshop (WS9) at ECAI'10, 7-14, Lisbon, Portugal, 2010.

[40] Hand D. J. : *Intelligent data analysis: issues and opportunities*, Advances in Intellligent Data Analysis. Reasoning about Data: Second International Symposium, Lecture Notes in Computer Science, vol. 1280, 1-14, 1997.

[41] Witten I. H., Frank E., Hall M. A. : *Data Mining : Practical Machine Learning Tools with Java implementations*, Morgan Kaufmann, 2011.

[42] *Rapid Miner* :
http://rapid-i.com/content/view/186/196/, consulted on 01/08/2011.

[43] Berthold M. R., Cebron N., Dill F., Di Fatta G., Gabriel T. R., Georg F., Meinl T., Ohl P., Sieb C., Wiswedel B. : *Knime: The Konstanz Information Miner*, http://www.knime.org/files/knime_whitepaper.pdf, 2006.

[44] Khabaza T., Shearer C. : *Data Mining with Clementine*, IEEE Colloquium on Knowledge Discovery in Databaes, IEEE Digest, vol. 21, London, 1995.

[45] Oprean C. : *Méthodologies et processus en Knowledge Discovery in Databases*, rapport pour le semestre 1 du 3ème année à Télécom Bretagne, 2011.

[46] Kurgan L., Cios K. : *Trends in Data Mining and Knowledge Discovery*, Knowledge Discovery in Advanced Information Systems, Pal N.R., Jain, L.C. and Teoderesku, N. (Eds.), Springer, 2002.

[47] Oprean C. : *Etat de l'art sur les aspects méthodologiques et processus en Knowledge Discovery in Databases*, rapport d'étude bibliographique du master de recherche en informatique de Télécom Bretagne/Univ.Rennes I, 2011.

[48] Tankeleviciene L., Damaseviciu R. : *Characteristics of Domain Ontologies for Web Based Learining and their Aoolication for Quality Evaluation*, Informatics in Education, vol. 8, no. 1, 2009.

[49] Cantarro M., Comito C. : *A data mining ontology for grid programming*, Proceedings of the 1st International Workshop on Semantics in Peer-to-Peer and Grid Computing, 2003.

[50] Hepp M., Roman D. : *An ontology framework for semantic Business Process Management*, Proceedings of Wirtschaftsinformatik, Karlruhe, 147-155, 2009.

[51] Choinski M., Chudziak J. A. : *Ontological Learning Assistant for Knowledge Discovery and Data Mining*, Proceedings of the International Multiconference on Computer Science and Informaton Technology, Karlruhe, Germany, 2007.

[52] Hilario M., Kalousis A., Nguyen P., Woznica A. : *A Data Mining Ontology for Algorithm Selection and Meta-Mining*, ECML Workshop on third generation data mining: Towards service-oriented knowledge discovery (SoKD-2009), Bled, Slovenia, 2009.

[53] Charest M., Delisle D., Cervantes O., Shen Y. : *Bridging the gap between data mining and decision support: A case-based reasoning and ontology approach*, Intelligent Data Analysis, vol. 12, no. 2, 211-236, 2008.

[54] Marban O., Segovia J., Menesalvas E., Fernandez-Baizan C. : *Toward data mining engineering : A software engineering approach*, Information Systems, vol. 34, no. 1, 87-107, 2009.

[55] Charest M., Delisle S. : *Ontology-guided intelligent data mining assistance: Combining declarative and procedural knowledge*, Proceedings of Artificial Intelligence and Soft Computing, 9-14, 2006.

[56] Kietz J.-U., Serban F., Bernstein A., Fischer S. : *Data Mining Workflow Templates for Intelligent Discovery Assistance and Auto-Experimentation*, ECML Workshop on third generation data mining: Towards service-oriented knowledge discovery (SoKD-2010), Barcelona, Spain, 2010.

[57] Wikipedia: *Portuguese wine*, http://en.wikipedia.org/wiki/Portuguese_wine, consulted on 15/07/2011.

[58] CVRVV: *Portuguese Wine - Vinho Verde. Comissao de Viticultura da Regiao dos Vinhos Verdes (CVRVV)*, http://www.vinhoverde.pt, consulted on 15/07/2011.

[59] Blogspot: *Portuguese Wine Experience*, http://portuguese-wine.blogspot.com/2010/07/portuguese-vinho-verdes-exports.html, consulted on 15/07/2011.

[60] Sharma S., Osei-Bryson K.-M. : *Toward an integrated knowledge discovery and data mining process model*, Knowledge Engineering Review Journal - KER, vol. 25, no. 1, 2010.

[61] Basili V. R., Weiss D. M.,: *A methodology for collecting valid software engineering data*,IEEE Transactions on Software Engineering, vol. 10, no. 6, 728-738, 1986.

[62] Cortez P., Cerdeira A., Almeida f., Matos T., Jantzem H., Reis J. : *Modeling wine preferences by data mining from physicochemical properties*, Decision Support Systems Journal, vol. 47, no. 4, 2009.

# A. List of Terms and Abreviations

**Activity** = "the work of a group or organization to achieve an aim" [49]
**BU** = **B**usiness **U**nderstanding (the firs step of CRISP-DM process)
**CBR** = **C**ase **B**ase **R**easoning
**CRISP-DM** = **CR**oss-**I**ndustry **S**tandard **P**rocess for **D**ata **M**ining
**Data Mining** (DM) = In this paper this word is used with the meaning: "methods and techniques for extracting knowledge from large amounts of data", as KDD. The term could also mean the modeling step from a KDD process.
**Data miner** = the DM user
**DMO** = **D**ata **M**ining **O**ntology
**DMOP** = **D**ata **M**ining **O**ptimization **O**ntology
**DMWF** = **D**ata **M**ining **W**ork**F**low Ontology
**DP** = **D**ata **P**reprocessing
**DU** = **D**ata **U**nderstanding
**ERP** = **E**nterprise **R**esource **P**lanning
**FF** = **F**ast **F**orward
**HTN** = **H**ierarchy **T**ask **P**lanning
**ICT** = **I**nformation **C**ommunication and **T**echnology
**IDA-API** = **I**ntelligent **D**iscovery **A**ssistance - **A**pplication **P**rogramming **I**nterface
**IS** = **I**nformation **S**ystem
**IT** = **I**nformation **T**echnology
**KDD** = **K**nowledge **D**iscovery in **D**atabases
**KDDONTO** = **KDD ONTO**logy
**KDP** = Knowledge Discovery Process
**M.** = **M**odeling
**Method** = set of procedures used for accomplish a task
**Methodology** = instance of a process, gives the tasks to be executed and the methods for indicating how these tasks will be executed
**Modeling** = step from the CRISP-DM process, that succeeds the Data Preprocessing step.
**NN** = Neural Network
**Ontology** = "An ontology renders shared vocabulary and taxonomy, which models a domain - that is, the definition of objects and/or concepts, and their properties and relations" [50]
**OWL-DL** = **O**ntology **W**eb **L**anguage **D**escription **L**ogic
**OWL-S** = **O**ntology **W**eb **L**anguage for **S**ervices
**Phase** = each distinct step from a process
**Process** = sequence of tasks executed for arriving to result
**RAC** = **R**equirements, **A**ssumptions, **C**onstraints
**RDF** = **R**esource **D**escription **F**ramework
**SE** = **S**oftware **E**ngineering
**SPARQL** = **S**imple **P**rotocol and **RDF** **Q**uery **L**anguage
**User** = in this paper, the term is used as a DM end-user (data miner)
**Workflow** = chain of operators in DM

[49] http://dictionary.cambridge.org/dictionary/british/activity_2
[50] http://en.wikipedia.org/wiki/Ontology_(information_science)