2009

An Enhanced Data Mining Life Cycle

Markus Hofmann
Technological University Dublin

Brendan Tierney
Technological University Dublin, brendan.tierney@tudublin.ie

Recommended Citation

M. Hofmann and B. Tierney, "An enhanced data mining life cycle," 2009 IEEE Symposium on Computational Intelligence and Data Mining, Nashville, TN, USA, 2009, pp. 109-117 , doi: 10.1109/CIDM.2009.4938637.

An Enhanced Data Mining Life Cycle

Markus Hofmann, Brendan Tierney

Abstract - **Data mining projects are complex and can have a high failure rate. In order to improve project management and success rates of such projects a life cycle is vital to the overall success of the project. This paper reports on a research project that was concerned with the life cycle development for data mining projects, its team members and their role. The paper provides a detailed view of the design and development of the data mining life cycle called DMLC. The life cycle aims to support all members of data mining project teams as well as IT managers and academic researchers and may improve project success rates and strategic decision support.**

An extensive analysis of eight life cycles leads to a list of advantages, disadvantages, and characteristics of the life cycles. This is extended and generates a conglomerate of several guidelines which serve as the foundation for the development of a new generic data mining life cycle. A detailed study of the human resources involved in a data mining project enhances the DMLC.

I. INTRODUCTION

Data mining is an interdisciplinary field [1], incorporating many different approaches, technologies, and methodologies to be able to generate and discover new and innovative knowledge. Data Mining is a non-trivial process [2] which has to be manageable in order to achieve the anticipated success in any data mining or knowledge discovery project. Due to the interdisciplinary nature and complexity of stages, processes, data and process flows during the progress of any data mining project the necessity for a comprehensive and complete data mining life cycle has been identified. The aim of the research project on which this paper reports was to develop a comprehensive new generic data mining life cycle called DMLC. The life cycle is a model that should assist all members of the data mining project team to succeed in their work and to deliver on time results that are within the predetermined budget and can be considered an extended version of the CRISP-DM life cycle.

The paper mainly focuses on the description of the life cycle's stages, processes, features, and all other parts that contribute to the comprehensiveness and integrity of the life cycle. Furthermore the introduction of human resources and their skill sets will be outlined in detail.

Markus Hofmann is a lecturer at the Institute of Technology Blanchardstown, Blanchardstown Road North, Dublin 15, Ireland. Phone: +353 (0)1 885 1553, markus.hofmann@itb.ie.
Brendan Tierney is a lecturer at the Dublin Institute of Technology in Kevin Street, Dublin 8, Ireland. brendan.tierney@dit.ie.

II. BACKGROUND

The following life cycles have been chosen as they represent the most authoritative, most cited, and most applied life cycles in both academia and industry:

- Knowledge Discovery in Databases Process [3, 4, and 5]
- Refined KDD Paradigm [6]
- Knowledge Discovery Life Cycle (KDLC) Model [7]
- Information Flow in a Data Mining Life Cycle [8 and 9]
- CRoss-Industry-Standard Process for Data Mining (CRISP-DM) [10]

The life cycles and their structure were explored, and the data or information flows are identified. The aim of the investigation stage of the project was to identify the advantages of stages or tasks within the life cycles, which would enhance the data mining project outcome. The authors also indicated the disadvantages of stages and tasks within the life cycle, which could affect the data mining project negatively.

Some of the life cycles only differ marginally whereas others differ considerably in structure and comprehensiveness. All but the CRISP-DM life cycle have been developed by academia. Although data mining has been a very evolving research area in recent years, there is still a need for a comprehensive and complete life cycle. The CRISP-DM reference model is the most complete and also the most widely accepted and applied data mining life cycle.

III. LIFE CYCLE ANALYSIS

The life cycles mentioned in the background section have been critically analyzed and compared with each other. The analysis of life cycles as outlined in [11] produced a list of shortened guidelines. These guidelines were used for the design of the enhanced life cycle:

- The number of processes is important to provide a comprehensive and detailed life cycle and should be above six;
- The life cycle should incorporate process, people and data issues in order to become equally process, people and data centric since all these aspects are considered to be important throughout data mining projects;
- A definite starting point is as important as a definite ending point of an individual cycle as well as numerous sub-cycles;
- Categorized processes add to the comprehensibility and task co-operation as well as distribution;
- The following processes have been identified as critical and are necessary to build a comprehensive life cycle: business

978-1-4244-2765-9/09/$25.00 ©2009 IEEE

understanding, data understanding, objective or hypotheses definition, selecting, sampling, data processing, transforming, data mining/modeling, evaluation, deployment, and post processing;

- In order to address the aforementioned feature of being people centric, personnel involved throughout the data mining project have to be identified and included in the new life cycle;
- The data sources have to be clearly identified;
- The issue of storing the newly gained information and knowledge is one of the most important issues of the data mining life cycle;
- The iteration of the life cycle should include the option to iterate through inner loops as well as to iterate the entire life cycle.

IV. DATA MINING LIFE CYCLE (DMLC)

The life cycle shown in Figure 1 represents a comprehensive approach to managing and optimizing data mining projects. The DMLC consists of 9 different processes which are part of 3 stages. The hypotheses/objectives preparation stage consists of the three processes: business understanding, data understanding and hypotheses/objectives definition. The second stage is called data preparation stage and represents the processes select/sample, pre-process and transformation. The final stage consists of the data mining process, the evaluation process and the deployment process. Two data stores are core to the life cycle: data warehouse/data mart and the information and knowledge repository (IKR). The outer circle of the diagram shows all the human resources or skill groups that have to be present throughout a large scale data mining project. This brief description of the life cycle will help to understand the following sections about human resource involvement in data mining projects. The placement of the symbols representing the human resources in the outer circle has no specific meaning.

The paper will focus on how this life cycle evolved and along what guidelines it was developed. Many features of the DMLC have been identified as advantages in other life. It should be noted that this paper describes and focuses on the development phases and the foundation of the life cycle segments and does not deal with the actual elements that have been implemented and displayed in Figure 1.

V. DEVELOPMENT PHASES OF THE DMLC

This core part of the paper deals with the development phases of the DMLC based upon features identified or not identified in the life cycles analyzed. Throughout the development phases this section will explain the various features, processes, or stages that will be introduced in detail. The section begins with the detailed foundations and guidelines of such a life cycle and is based on the list of guidelines identified and compiled in the section above.

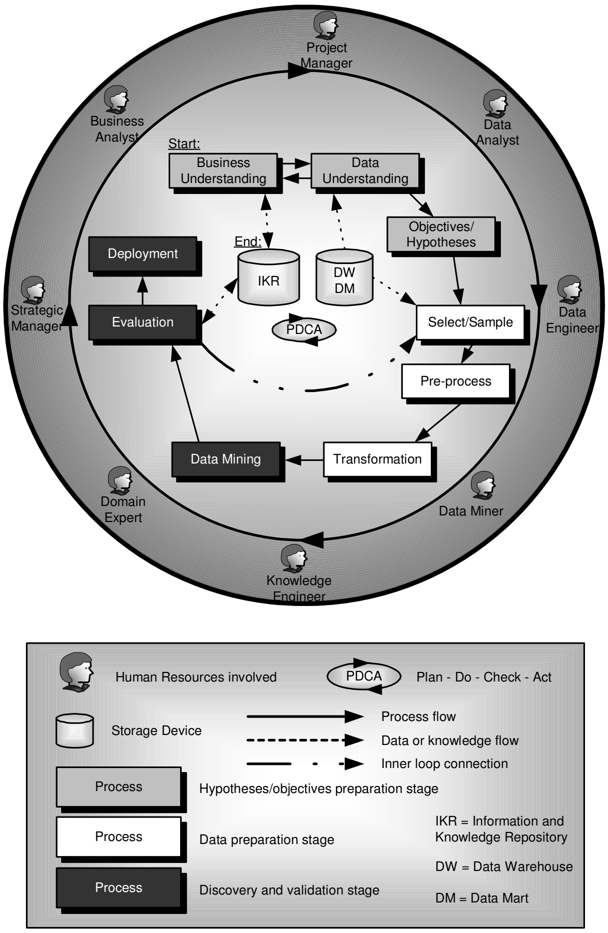

Figure 1: DMLC - Enhanced Data Mining Life Cycle

Three ways of identifying the points leading to the guidelines have been discovered:

- The present feature, process, or stage has been defined as strength or positive influence of a life cycle and therefore adds to the overall life cycle output. The emphasis of such an event would therefore be important and critical to incorporate into the new life cycle development;
- The feature, process, or stage has been defined as weakness or negative influence of a life cycle and therefore detracts from the overall life cycle output. This event needs to either improve the detraction in such a way that it becomes a strength or has to be left out entirely in order to improve the feature, process, or stage and therefore the overall life cycle;
- The life cycle was lacking in one area and the authors identified a missing feature, process, or stage, which also detracts from the overall life cycle output. This event leads to the necessity of adding the missing feature, process, or stage in order to improve the overall life cycle output.

The following sections will address the guidelines in greater detail.

A. Scope of the life cycle processes

After analyzing the existing life cycles it became obvious that the new life cycle has to have at least six processes in order to be detailed enough and to qualify as comprehensive and complete. The life cycles from [5] and [12] show the lack of completeness by only producing 3 phases. For example [5] introduces a process called 'post-processing'. The user is unable to determine what post-processing means and includes in this case. It is not mentioned that any processes or tasks should be included in the post-processing process such as evaluation, deployment or storage of the results.

The KDD Process [4] on the other hand introduces eight processes and produced in fact one of the more coherent life cycles. The amount of processes should be manageable and add to the project planning and execution rather than causing confusion and deflection. After limiting the lower borderline to 6 processes, the upper borderline is limited to approximately ten processes in order to keep the life cycle manageable and straightforward. This will not allow the life cycle to become too complex and also emphasizes the necessity to concentrate on processes rather than on activities. For example [6] include the process 'Clean data' in their life cycle, which is rather an activity than a process. The process should be named with a noun and not with an active verb since the final deliverable will be a life cycle and not a methodology.

B. Process, people and data issues

There were three orientations detected after analyzing the life cycles. Some of the life cycles were data centric, some process centric and others tried to incorporate the people issues and requirements. Only [8] and[9] used all three aspects to improve their life cycle. Using all three aspects of the data mining project will enhance the life cycle due to greater comprehensiveness and completeness. Therefore the life cycle should incorporate process, people and data issues in order to become equally process, people and data orientated.[2], [8] and [9] mention human resources as critical source in their models and only the latter two authors cover all three issues, even if only marginally. The new life cycle has to identify clearly what data sources are used, what data flows exist, what processes are necessary to generate the data flow, and furthermore, who are the people responsible for carrying out these processes. The life cycle has to consider all three factors equally in order to provide a platform for successful data mining project planning and implementation.

C. The Processes

This section will describe the various processes defined for the DMLC. The processes are split into three stages each containing three processes:

- Objectives/Hypotheses preparation stage;
- Data preparation stage;
- Discovery and validation stage.

Business Understanding, Data Understanding and Objectives/Hypothesis build the 'Objectives/Hypotheses preparation stage'. It is, for example, not possible to carry out the business understanding analysis if no initial thought or draft of a hypothesis or objective is in place. And, vice versa, it is not possible to define the final objectives or hypotheses without carrying out the business understanding or data understanding process. It is therefore important that these three processes are approached simultaneously, going back and forth until the objectives or the hypotheses are set. It is further noteworthy to mention that the first two stages can take up to 80 percent of the project time [5]. Although [13] and [14] do not give a percentage, both agree that the first two stages take up most of the workload in order to obtain a valid and beneficial data mining result.

The various processes and their integration into the life cycle as well as the human resources that are responsible to create the process deliverable are described in the following paragraphs. Again, this is not a methodology and therefore the deliverable of this paper is not a complete list of activities or tasks that have to be carried out but rather a guideline for what role the process plays in the overall life cycle.

Business Understanding

This process is one of two fundamental processes of the DMLC. Not only because it is the starting point but also because of its influence and importance with regard to the overall data mining project. This process is necessary to set an objective or to formulate the hypothesis. In order to do so, the user has to understand how the business works, what the overall objectives of the business are and of course, what the rules of the business are. The following activities are necessary to cover the business understanding process:

- Determine the Business Objectives – Research the background, the business objectives, and business success criteria [10].
- Assess the overall situation – research the requirements, assumptions, and constraints, estimate the risk and contingencies and based on that research create a cost and benefit analysis [10].
- Determine basic business rules that are necessary for the data mining project and store them in the Information and Knowledge Repository (IKR).

The sharing of knowledge is important to achieve consistency and correctness in the anticipated data mining results [14]. It is also important to mention that there should be a distinct difference between the IKR and any metadata related to the semantics of the data kept in the data warehouse/mart. The IKR is simply responsible for holding business rules and previously achieved data mining results [7] whereas the usual metadata is focused on the semantics of the data [15]. This differentiation is necessary to ensure integrity and reduce redundancy of these two storage devices.

The business analyst holds the main responsibility of this process. However, the data analyst, domain expert, and the strategic manager assist throughout this process. The IKR is used to access previously defined business rules or previously gained information or knowledge.

Data Understanding

Business understanding is the second of two fundamental processes in the DMLC. Data understanding is necessary to create a good platform for the data mining project. The following issues have to be addressed to fulfill the requirements of this process, assuming that the initial data collection has already taken place:

- Describe data [10] and store this information in the metadata repository of the data source (data warehouse or data mart) [15]. This contributes to the general data understanding and the semantics of the individual attributes themselves as well as their values and restrictions. It is, according to [13], important to describe tables and their relation to each other. Without metadata the workload of the data mining project increases considerably [13].
- The second important activity is to explore data [10] in order to more completely understand the contents of the data source and the data itself. Missing values, outliers, or skewed distributions can affect the data mining model and distort its results considerably [3, 15]. This is usually carried out with a small sample and not with the entire data set [3].
- To know the volume of the data such as number of records or number of attributes is helpful to find the right sample size for the later processes [3]. Information about the initial data type is also necessary.
- Verification of the data quality is vital [13] to determine what steps have to be carried out in the data preparation stage. If there is an issue with quality the user has to analyze how distorting this error is in order to take the appropriate action and whether this error can be removed or not.

The data analyst holds the main responsibility of this process. However, the business analyst and the data engineer assist throughout this process. The data warehouse/mart is used to access data and metadata. Again, this process should be carried out in concurrence with the remaining two phases of this stage. Input from and to the business understanding processes is vital.

Objectives or Hypotheses Definition

The last process of this stage is the objectives/hypotheses definition process where the knowledge and outputs from the two previous processes are brought together and the final objectives or hypotheses of the data mining project can be defined. [10] warns of the confusion between business objectives and data mining objectives. A data mining result can lead towards the business objective, but cannot be the same. For example:

- Business objective: e.g. To better co-ordinate linked bus journeys
- Data mining objective: e.g. Predict how many passengers will be in need of linked bus journeys of a certain route, considering journey behaviors and patterns over the last 4 years and demographic information (age, salary, and route).
- Data mining hypothesis: e.g. Knowing the number of required linked trips on a certain route would improve planning and scheduling processes.

This process can be further improved by stating the data mining success criteria [10]. A success criterion could contain, for example, the level of predictive accuracy, the level of significance, the p-value or the power of the results. It is believed that an estimation of the result prior to the data mining execution could eliminate or identify impossible results and therefore decrease mistakes within or throughout the data mining life cycle.

To define an objective or a hypothesis is only mentioned by [6] and [7]. [10] mentions that determining the data mining goals as part of their business understanding process. The determination of, what [10] call ‘data mining goals’ is before the data understanding process and also not an individual process. To set the definition of the objectives or hypotheses as an individual process is considered critical and vital. [7] states that the generation of a hypothesis is important but this gets lost in the unstructured way the life cycle is presented. Only [6] add the process of defining an objective to the life cycle. They further say that the objectives used in conjunction with the business understanding, can successfully reveal and discover new business insights. This was also confirmed by Turban *et al.* [14].

This process concludes the cooperation of the following human resources: business analyst, data analyst, domain expert, data engineer, and strategic manager. Both data stores are accessed to increase data, information, and/or knowledge transparency.

Select and Sample

The goal and final deliverable of this process is to identify and select the target data, which is the data needed to carry out the data mining algorithm [3]. Selection and sampling data from a larger data source such as data warehouse or data mart is responsible for creating the required target data. Data selection includes manual and automatic selection of records, selection of attributes or features, as well as reducing the number of values, for example, by applying discretization techniques [10, 16 and 17]. The selection and sampling process often has to be iterated numerous times in order to finally select a suitable and correct data sample [9]. The selected data set/sample will then be further processed and finally used for modeling and the main analysis work. [17] warns that the selection criterion not only includes relevance to the data mining goal, but also technical constraints, such as limits on data volume or data types.

The selection process covers the selection of attributes as well as records from one or more tables [18]. The decision on what data is used for the analysis is based on a number of criteria, including relevance to the data mining objective/hypotheses, data quality, and technical constraints [10].

Pre-Process Data

After selecting and sampling the suitable data set further steps are necessary to adjust the data to comply with the general and specific needs of the data mining model or algorithm. Although the data obtained from the data warehouse (or similar) has already undergone steps such as data migration, data scrubbing, and/or data auditing [5], it still needs to be further processed to guarantee the most suitable and error free data set [4, 5, 8 and 9]. To improve the quality of the data it has to be cleansed from noise and outliers. Outliers are attribute values that did not come from the assumed population of data [14]. This can include a non-numeric value when the user is only expecting numeric values or data items that fall outside the boundaries set by most of the other data items in the data set [19]. Data can be referred to as noisy when it includes errors such as many missing or incorrect values or when there are irrelevant attributes [4].

These constraints have to be removed in order to maximize the data quality and therefore the quality of the result. More ambitious techniques such as the estimation of missing data by modeling can further contribute to the improvement of data quality [17, 20]. A strategy or procedure for handling missing data items has to be created [3].

Transform Data

In the transformation process, which follows the pre-processing phase, the data is constructed, integrated and formatted [10, 17] to finally comply with all data mining model requirements. It has to be transformed to fit the essential normalization rules implied by the model [6]. Transformation is the data reduction and projection – using dimensionality reduction or transformation techniques to condense the effective number of variables under consideration or to discover invariant representation of the data [3]. This could simply mean to eliminate unwanted or highly correlated fields so the results will be valid.

The last three processes are usually carried out by the data engineer and data miner who obtains assistance from the data analyst. The required data is accessed from the data warehouse/mart.

Data Mining

After all the data preparation has been concluded, the appropriate data mining model can be applied. During this process the information or knowledge is extracted from the main data set. The following list includes the most common types of information obtained by data mining algorithms:

- Classification: gather the significant characteristics of a certain entity or group (e.g. customers who changed their route of travelling)
- Clustering: recognizes entities or groups that share a particular characteristic (This differs from classification in the sense that no predefined characteristic is given)
- Association: identifies relationships between events that occur at one time
- Sequencing: similar to association, except that the relationship is present over a period of time

Although the type of information that is aimed for is already known from a very early stage of the life cycle, it becomes crucial in supporting the decision what data mining model (e.g. decision trees, neural networks, memory based reasoning, etc.) should be used to carry out the analysis task. Most problem types can be solved by applying different or a combination of data mining models [2, 14, and 15].

The data miner's main responsibility is to generate a data mining model and to run this model in order to produce the anticipated analysis result. The expertise of the domain expert supports this task.

Evaluation

The evaluation process is aimed at validating the result and putting it into context of the initial objectives or hypotheses. Many authors claim that the data mining process can produce an unlimited number of patterns hidden in the data and the evaluation process is solely responsible for selecting the useful results [3, 4, 22]. Due to the fact that the DMLC has a clearly defined objective or hypothesis, it should not be necessary to choose from different results since the entire data mining algorithm is based on the objective or hypothesis. [10] and [23] say that the evaluation process is mainly responsible for validation of the model and the results. Obviously one must keep the business rules and data mining objective or hypothesis in mind. This process also includes the overall review process and determines further steps such as reiteration (inner loop) or deployment. Each correct result is stored in the IKR, regardless of whether the outcome is relevant to the current business situation or not. If it is a correct result and therefore reflects a business rule, limitation, or boundary, it has to be stored in order to include it in any future data mining project. [10] states that the evaluation should include the search for incidentally generated results which could be beneficial for the business understanding process or could trigger another data mining project. It is important that the data mining result is understood.

The domain expert, the business analyst, and the strategic manager ensure that the evaluation of the results are carried out. The knowledge engineer assisted by the domain expert is responsible for the storage of the data mining and evaluation results.

Deployment

The deployment process covers the deployment activities in the event that the data mining result added to the business understanding and discovered new knowledge. Only [6] and [10] cite the deployment stage as a necessity of a data mining life cycle. [21] and [24] say that Return on Investment (ROI) of data warehouses cannot be achieved without data mining and successful data mining cannot be attained without correct deployment. Data mining deployment enables communication about the knowledge and experience gained from data mining projects to the human resources interested in the new findings. Deployment of data mining results provides employees the information they need, in a form they can use, where they need it and exactly when they need it [24].

After deploying the results it is important to analyze the impact of the deployment. Only [6] indicated this aspect in their life cycle. The analysis of that deployment would give an indication whether to introduce the findings of the data mining result on a larger scale or whether to abort the deployment. An example of this is the typical beer and nappies scenario. The result is that every Friday many customers tend to buy beer along with nappies. What the analyst does not know is what to do with that result. Should the beer be stocked beside the nappies or should it be kept on the opposite side of the supermarket hoping that the customer purchases more goods on the way to the beer or the nappies? Looking at the beer/nappies example, two possibilities arise. One, knowledge about a similar situation already exists and can be used to apply the necessary changes or, two, the result has to be deployed and the life cycle has to be repeated comprising the changes provoked by the impact of the deployment. By deploying the data mining results in a selective distribution, further knowledge can be gained which is filtered and analyzed through a new data mining project. If for example 10 supermarkets locate the beer and nappies next to each other and a further 10 stores locate them as far from each other as possible the result generated through a new data mining project will reveal which of the options was more profitable and therefore will be deployed on a larger scale.

The deployment process is usually under the control and supervision of the strategic manager and the domain expert.

D. Human Resources

Only three [7, 8, and 9] of eight analyzed life cycles consider human resources as vital and include or mention them in their life cycle description or diagrams. The following human resources were mentioned: Data Miner [8], Data Analyst [8], Knowledge Engineer [9], and Domain Experts [7].

After consulting various general systems development literatures [25 and 26], indications that data mining projects require more qualified and specialized personnel covering the various aspects and skill requirements arising during a data mining project became more obvious. The following human resources have been identified as necessary to approach and implement a data mining project successfully. All identified human resources have been taken from [7, 8, 9, 25, 26, and 27]: Business Analyst, Data Analyst, Data Engineer, Domain Expert, Data Miner, Knowledge Engineer, Strategic Manager, Project Manager.

The following points will describe why these skill groups are considered as crucial for the success of a data mining project and what their tasks throughout the project are:

- The business analyst is responsible to understand the business aspects and plays the major role in generating the hypotheses or objectives in cooperation with the data analyst, domain expert, and the strategic manager. The business analyst also evaluates the data mining result and qualifies its relevance to and impacts on the current or future business situation.
- A data analyst is someone who analyses database requirements and designs and constructs corresponding databases [26]. The data analyst is responsible for the data understanding process and plays the other major role in building the hypotheses or objectives in cooperation with the business analyst, domain expert, and strategic manager.
- Data engineers are computer professionals trained to elicit knowledge from domain experts [25] and their responsibilities are based on a specialized body of knowledge [26]. They are responsible for the entire data preparation stage. The inputs to the hypotheses or objectives. The data analyst will help and guide the data engineer through this task.
- A domain expert is a subject matter expert with relevant background, experience, or expertise in specific subjects. These can include individuals in research, academia, government, industry, or non-profit institutions [28]. In many cases these 'experts' will be locally active individuals with indigenous knowledge of their communities, environment and cultures [28 and 29]. The domain expert works closely with the knowledge engineer.
- The data miner is responsible for the generation of the data mining algorithm or model. It is also necessary to forward all requirements to the data engineer. Briefly, a data miner needs to know more than just database technology or statistics to perform efficiently. Data mining requires knowledge in the areas such as database/data warehousing, domain expertise, statistics, and business processes [30 and 31].
- The knowledge engineer plays a critical role in ensuring that knowledge is not only obtained, but also represented and structured for optimal use and re-use. The knowledge engineer is further responsible for storing, extracting and maintaining information and knowledge kept on the IKR. He/she contributes indirectly to the hypotheses or objectives generation and to the evaluation process.
- The strategic manager or planner is mostly concerned with crystallizing the strategic issues that ought to be addressed [32]. [32] further add the responsibilities of providing data, conducting studies of industry and competitive conditions and to developing assessments of strategic performance. The strategic manager is mainly concerned with the evaluation and deployment of the results but also plays a minor role during the definition stage of the hypotheses or objectives. Knowing the business and its parameters extensively, the impact of the results and their deployment has to be evaluated by the strategic manager or planner.
- The project manager's responsibility is to define, plan, direct, monitor, organize, staff, and control a project to develop an acceptable system [25, 26] and to execute a data mining project within the prearranged budget and time. He/she also supervises and controls all processes and interactions between the other human resources. The tasks comply with the general tasks of a project manager.

E. Data Source

The data source(s) has to be clearly identified. Many authors are unclear what the actual data source for a data mining project should be. Some say that any data store is sufficient to carry out the analysis [3, 7], others do not even mention this issue throughout their life cycle [6, 10]. On the other hand, many authors explicitly state the necessity to have a data warehouse [5, 8, 9].

We believe that any semi-structured source of data is more or less sufficient to carry out data mining operations [33]. However, to ensure and enhance the level of quality in the anticipated results the data should be pre-processed and put into structured form, which usually complies with the characteristics of at least a small data mart or a data warehouse [34]. Data warehouses and data marts are inevitable data sources for larger data mining projects (Inmon, 1996; Kimball *et al.*, 1998; Turban and Aronson, 2001).

The issue of storing the newly gained information and knowledge is one of the most important issues of the data mining life cycle. An information knowledge repository (IKR) is a collection of both internal and external knowledge and is generally stored on a relational database in a way that enables efficient storing [14]. [14] further state that the scope of the knowledge repository mainly depends on the type of knowledge that has to be stored. The three basic types of repositories found in practice [35]:

- External knowledge, such as competitive intelligence, which generally needs explanations and interpretation.
- Structured internal knowledge, such as research reports, presentations, and marketing materials, which is mostly explicit knowledge with some tacit knowledge.
- Informal internal knowledge, such as discussion databases, help desk repositories, and shared information databases.

The importance of creating and maintaining an information or knowledge repository has been identified. The term metadata for the repository is purposely circumvented since metadata is already present and used in relation to the data source. Metadata is according to [15] all of the information in the data warehouse environment that is not the actual data itself. Metadata describes the data being investigated on the semantic, structural, statistical, and physical level in order to support tasks such as data validation and imputation, selection and application of data mining methods, and interpretation of the results [16]. Only [7] mention the usage of an information and knowledge repository and insist on storing the newly gained information and knowledge.

F. Starting and Ending Point

A definite starting point is as important as a definite ending point of an individual cycle as well as numerous cycles (when iteration becomes necessary). Although most of the analyzed life cycles have a starting point as well as an ending point, it is often not clear where the cycle actually begins and even more importantly when and where the end of the cycle has been reached. [7] have neither, which leads to more confusion than actual guidance throughout the data mining project. The only three life cycles where the ending point is chosen are [4, 5, 10]. However, in all three cases the ending point is defined as a process, which automatically means that data, information, or knowledge are getting lost. The authors believe that the ending point has to be a data store in order to store the outcome of the last process and to build an improved platform for the processes that access this data store throughout the life cycle. More important than simply having a starting and ending point is the placement of the two such important milestones. Where should the project begin and where does the life cycle finally end or lead back into the starting point? This is a very important question. Two different ways have been analyzed:

- The life cycle starts with the data source which then has to be further processed [3, 5, 8, 9]
- The starting point is equivalent to the first process of the life cycle [6, 10]

The authors believe that the first process should indicate the starting point since the data source (as the word source indicates) should only be used as the basis of the life cycle – the initial input. The data source should therefore be in a stage where data can be extracted, fulfilling the requirements of the initial data mining processes. After determining the starting point (the first process), the life cycle has to clearly indicate where the loop is closed or where the life cycle finally ends.

G. Iteration

The iteration of the life cycle should include the option to iterate through inner loops as well as to iterate the entire life cycle. Inner loops are necessary to get a data mining result correct without going through processes that remain unchanged. For example the data selection or data mining algorithm may change throughout a project whereas the hypothesis or objective remain the same. Often more than one iteration is necessary to achieve the anticipated result and to prove or disprove the hypothesis or objective [36]. The feature 'forward skipping' is when certain processes can be skipped throughout the data mining projects. This has been defined by the authors as a disadvantage and will therefore not be supported in the new DMLC. A sequential flow will be present at any given time in order to contribute to the transparency and comprehensiveness of the life cycle.

H. Quality

The quality of the life cycle and its outcome is seen as one of the most important objectives. Most authors forget about quality assurance and therefore minimize the usefulness of the result or outcome. According to [37] quality is the *"…extent to which an industry-defined set of desirable features are incorporated into a product…"*. To introduce quality into a life cycle is rather difficult since the quality of the outcome is determined by the handling of the activities within the processes. However, a good data and process flow between the various processes controlled by the suitable skill group can considerably improve the quality of the outcome. Since all processes are dependent on their deliverables, the quality of these deliverables is crucial to the quality and therefore the success of the project. In order to ensure the quality of the deliverables a methodology was introduced that ensures the successful completion of each process and thus the complete data mining life cycle. The methodology is a variation of the Plan Do Check Act cycle (PDCA Cycle) which is also known as the Deming Cycle [38]. Although this control circle is usually used on a larger scope such as entire projects rather than individual processes, the PDCA methodology is very suitable to control each of the processes.

Other overall objectives in creating the life cycle are to fulfill the following characteristics:

- Industry neutral [10]: The aim is to develop a generic life cycle that applies to all industries. This is seen as an advantage since the life cycle has to incorporate issues applying to all industries whereas an industry specific life cycle might not include aspects necessary for data mining projects of neighboring industries.
- Application neutral [10]: The aim is to create a generic life cycle that applies to all applications used for the execution of the data mining project. The same advantage applies as mentioned above: since the life cycle aims to be application neutral, it has to incorporate all issues no matter what application(s) are used to carry out the project.
- Tool neutral [10]: The aim is to create a tool neutral life cycle that incorporates all issues irrespective of the tools used to execute the data mining project.
- Easy to comprehend: The comprehensibility of the life cycle is vital. The user should be able to apply the life cycle to the data mining problem type without having to do much background reading. Again, the life cycle must not be seen as a methodology and does not act as a detailed step-by-step guide to carry out a data mining project. It is rather to set the cornerstones and the data or process flows between these. Even users who have a general IT understanding but not a specific knowledge about data mining should be able to comprehend the life cycle.

VI. CONCLUSION

The exploration and classification of eight life cycles built the foundation of the Data Mining Life Cycle (DMLC). The life cycles have been qualified for their usage and suitability for implementing data mining projects. This analysis led to the guidelines for the design and development of the DMLC. The advantages, disadvantages and characteristics of each life cycle contributed significantly to the overall outcome.

The definition and visualization of the involvement of human resources and their skill groups in a data mining project contributed to the completeness and comprehensiveness of the DMLC. The DMLC aims to be process, people and data focused. The development phases of a generic data mining life cycle were the core part of this paper. The final version of the DMLC was built on various guidelines that were derived from an extensive life cycle analysis. Characteristics, advantages and disadvantages were the foundations of these guidelines.

The data aspect introduced the Information and Knowledge Repository (IKR) which stores all analysis findings as well as business rules and requirements. The data flow as well as the data content is different to that seen by other authors. Only [7] introduced a repository which is, however, mainly focused on database and data warehouse information and not on the actual discovered knowledge and therefore the business knowledge. The life cycle is built and based on data warehouses or data marts, which ensure better data integrity and quality than OLTP systems.

The processes as well as the process flows are more comprehensive and complete than generally displayed in other life cycles as identified throughout the paper. This is in accordance with the PDCA methodology, which ensures the quality of each process and therefore the overall quality from a process point of view.

To see the data mining life cycle from a human resource point of view contributes considerably to the DMLC as well as to the present body of knowledge. The listing of all skill groups involved in a general data mining project indicates the multi-disciplinary nature of such projects.

The main contribution however is the product of these three aspects – the complete life cycle, which shows the interactions of the three aspects as well as all other features. The DMLC consists of nine different processes, various iteration possibilities, two different data sources, eight different skill groups, and a methodology that ensures the successful completion of each life cycle (PDCA).

It is believed that the life cycle can contribute significantly to data mining projects when applied correctly. Project management is a key success factor to ensure the expected outcome with a predefined budget. The DMLC and its underlying methodology focuses on all areas arising throughout a data mining project and can contribute to managing complexity, is able to improve project outcome and is therefore an information technology tool that can be used successfully to improve strategic management.

REFERENCES

[1] Simoudis, E. (1998) Discovering Data Mining, foreword in Cabena, P et al., Upper Saddle River, Prentice Hall PTR, New Jersey, USA.

[2] Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (1996b) 'From Data Mining to Knowledge Discovery: An Overview', Advances in Knowledge Discovery and Data Mining, AAAI Press, USA.

[3] Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (1996a) 'Knowledge Discovery and Data Mining: Towards a Unifying Framework', Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96), AAAI Press, USA.

[4] Collier, K. et al. (1998) 'A Perspective on Data Mining', Centre for Data Insight, Northern Arizona University, USA, pp 2-4.

[5] Feldens, M. (1998) 'Towards a Methodology for the Discovery of Useful Knowledge – Combining Data Mining, Data Warehousing and Visualization', CNPq/Protem-cc Fase III (SIDI Project), Universidade Federal do Rio Grande do Sul, Brazil.CRISP-DM (2000) – 'CRISP-DM 1.0 – Step by Step data mining guide', CRISP-DM Consortium.

[6] Collier, K. et al. (1998) 'A Perspective on Data Mining', Centre for Data Insight, Northern Arizona University, USA, pp 4-6.

[7] Lee S. W. and Kerschberg, L. (1998) 'A Methodology and Life Cycle Model for Data Mining and Knowledge Discovery in Precision Agriculture', Conference Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics, IEEE Computer Society Press, pp. 2882-2887, San Diego, CA, USA.

[8] Ganesh, M. et al. (1996) 'Visual Data Mining: Framework and Algorithm Development', Department of Computing and Information Sciences, University of Minnesota, MN, USA.

[9] Kopanakis, I. and Theodoulidis, B. (1999) 'Visual Data Mining & Modeling Techniques', Centre of Research in Information Management (CRIM), Department of Computation, University of Manchester Institute of Science and Technology, UK.

[10] CRISP-DM (2000) – 'CRISP-DM 1.0 – Step by Step data mining guide', CRISP-DM Consortium.

[11] M. Hofmann and B.Tierney. Development Phases of a Generic Data Mining Life Cycle (DMLC). International Conference on Software Engineering Theory and Practice. Proceedings, Orlando, USA, 2007.

[12] Simoff, S. and Maher, M. L. (1998) 'Analyzing Participation in Collaborative Design Environments', Key Centre of Design Computing, University of Sydney, NSW, Australia.

[13] Inmon, W. H. (2002) Building the Data Warehouse (3rd edn), John Wiley & Sons, Inc., New York, USA.

[14] Turban, E. and Aronson, J. (2001) Decision Support Systems and Intelligent Systems (6th edn), Prentice-Hall, New Jersey, USA.

[15] Kimball, R., Reeves, L., Ross, M. and Thornthwaite, W. (1998) The Data Warehouse Life-cycle Toolkit, John Wiley & Sons, Inc., New York, USA.

[16] Klösgen, W. (2002b) 'Types and Forms of Data', In Klösgen, W. and Zytkow, J. M. (eds), Handbook of Data Mining and Knowledge Discovery, Oxford University Press, New York, USA, pp. 33-44.

[17] Reinartz, T. (2002) 'Stages of the Discovery Process', In Klösgen, W. and Zytkow, J. M. (eds), Handbook of Data Mining and Knowledge Discovery, Oxford University Press, New York, USA, pp. 185-192.

[18] Date, C. J. (2000) An Introduction to Database Systems (7th edn), Addison Wesley Longman, Inc., USA.

[19] TWO CROWS (1999) 'Introduction to Data Mining and Knowledge Discovery' (3rd edn), Two Crows Corporation, (online) (cited 20 September 2008). Available from <URL: http://www.twocrows.com/glossary.htm>.

[20] Bloom, T. (2002) 'Data Warehousing – Data Cleaning and Loading', In Klösgen, W. and Zytkow, J. M. (eds), Handbook of Data Mining and Knowledge Discovery, Oxford University Press, New York, USA, pp. 193-204.

[21] Noonan, J. (2000) 'Data Mining Strategies', DM Review, published in July 2000 (online) (cited 27 September 2008). Available from <URL:http://www.dmreview.com/master.cfm?NavID=198&EdID=2367>.

[22] Hsu, W. et al. (1998) 'High-Performance Commercial Data Mining: A Multistrategy Machine Learning Application', Automated Learning Group, National Centre for Supercomputing Applications (NCSA), University of llinois, USA.

[23] SHEARER, C. (2000) 'The CRISP-DM Model: The New Blueprint for Data Mining', Journal of Data Warehousing, Volume 5, Number 4, p13.

[24] Battaglia, M. (2001) 'Taking Data Mining to the Next Level', DM Direct, published in June 2001 (online) (cited 27 September 2008). Available from <URL:http://www.dmreview.com/master.cfm?NavID=198&EdID=3530>.

[25] Hoffer, J. A., George, J. F. and Valacich, J. S. (2001) Modern Systems Analysisand Design (3rd edn), Prentice Hall College Div,

Washington, USA.

[26] Whitten, J. L, Bentley, L. D. and Dittman, K. C. (2001) Systems Analysis and Design Methods (5th edn), International Edition, McGraw Hill/Irwin, New York, USA.

[27] Berry, M. J. and Linoff, G. (1997) Data Mining Techniques – For Marketing, Sales, and Customer Support, John Wiley & Sons, Inc., New York, USA.

[28] Wong, P. Y. (1999) 'A Scalable Framework for Collaborating Web Clearinghouses', National University of Singapore.

[29] Knublauch, H. (2002) 'Extreme Programming of Knowledge-Based Systems', Research Institute for Applied Knowledge Processing (FAW), Conference Proceedings from eXtreme Programming and Agile Processes in Software Engineering (XP2002), Alghero, Sardinia, Italy.

[30] Luan, J. (2002) 'Data Mining and Knowledge Management in Higher Education', Cabrillo College, Presentation at AIR Forum, Toronto, Canada.

[31] MITKAS, P. et al. (2002) 'An Agent Framework for Dynamic Agent Retraining: Agent Academy', Conference proceedings, eBusiness and eWork 2002 12th annual conference and exhibition', Prague, Czech Republic.

[32] Thompson, A. A. and Strickland, A. J. (2001) Strategic Management – Concepts and Cases (12th edn), International Edition, Irwin/McGraw-Hill, USA.

[33] Hofmann, M. and Tierney, B. (2003) 'The Involvement of Human Resources in Large Scale Data Mining Projects', International Symposium on Information and Communication Technologies, Dublin, Ireland.

[34] Hofmann, M., O'Mahony, M. and Tierney, B. (2003) 'A Framework to Utilise Urban Bus Data for Advanced Data Analysis', 10th World Congress on Intelligent Transport Systems, Madrid, Spain.

[35] Davenport, T. H. and Prusak, L. (1998) Working knowledge: How Organizations Manage What They Know, Harvard Business School Press, Boston, USA.

[36] Klösgen, W. (2002a) 'Data Mining Tasks and Methods – Change Analysis', In Klösgen, W. and Zytkow, J. M. (eds), Handbook of Data Mining and Knowledge Discovery, Oxford University Press, New York, USA, pp. 361-364.

[37] Fitzpatrick, R. (1996) 'Software Quality: Definitions and Strategic Issues', Staffordshire University, School of Computing Report.

[38] Deming, W. E. (2000) Out of Crisis, Massachusetts Institute of Technology, Centre for Advanced Engineering Study, MIT Press, Cambridge, MA, USA.

[39] ISO (1998) 'Information technology – Guide for ISO/IEC 12207 (Software Life Cycle Processes)', Technical Report, ISO/IEC TR 15271.