SysPRE - Systematized Process for Requirements Engineering

Ana Neto[1(✉)], Duarte Pinto[1], and David Aveiro[1,2]

[1] Faculty of Exact Sciences and Engineering, University of Madeira,
Caminho da Penteada, 9020-105 Funchal, Portugal
ana.b.neto@gmail.com, duarte.pinto.oelabuma@gmail.com

[2] Madeira Interactive Technologies Institute, Caminho da Penteada, 9020-105 Funchal, Portugal
daveiro@uma.pt

Abstract. The domain of Knowledge Discovery (KD) and Data Mining (DM) is of growing importance in a time where more and more data is produced and knowledge is one of the most precious assets.

Having explored both the existing underlying theory, the results of the ongoing research in academia and the industry practices in the domain of KD and DM, it was found that this is a domain that still lacks some systematization.

It was also noticed that this systematization exists to a greater degree in the Software Engineering and Requirements Engineering domains, probably due to being more mature areas.

In this paper we propose SysPRE - Systematized Process for Requirements Engineering in KD projects to systematize the requirements engineering process for these projects so that the participation of enterprise stakeholders in the requirements engineering for KD projects can increase.

Keywords: Knowledge discovery · Data mining · Requirements engineering · DEMO

1 Introduction

Software development has been around for several decades now and discussion on its failures and successes has been strong.

It all started with the Standish Group's Chaos Report of 1994 [1] that stated that projects that did not meet customer satisfaction and/or went over time or budget in a significant way corresponded to 53%. It was a bit shocking to see a figure that amounted for over half of the projects and a discussion about a software crisis was started.

This report, however, was since then criticized for lack of peer review, for not having a complete description of the study design or of the project selecting criteria, and for defining successful and failed projects in a way that may bias the study [2, 3, 5].

Over 20 years later, the debate is still on, but there seems to be an agreement on the failure rate of software development projects having dropped [5–7]. Although the values do not coincide, they show a decrease tendency that may be significant if you take into account that projects are increasingly complex.

D. Aveiro et al. (Eds.): EEWC 2017, LNBIP 284, pp. 166–180, 2017.
DOI: 10.1007/978-3-319-57955-9_13

One of the areas of software development that has helped this increased success in software projects is Requirements Engineering (RE), following previous research such as [8, 9]. Furthermore, according to [6], one of the three main reasons for the positive development is that the communication of requirements has much improved. [10] makes an even stronger statement that "Meets user requirements" is the most important success criteria for both users (96%) and project managers (81%).

Knowledge discovery and data mining are much more recent areas than software development and less mature fields. For instance, if considering process model development to be a sign of maturity, it can be seen that the first process model for this area dates back to 1996 [11], while in software development, the well-known Waterfall model goes back to 1970 [12].

Nonetheless, it is indisputable that knowledge discovery and data mining are of growing importance in a time where more and more data is produced.

Data production numbers are, in fact, staggering, for example, 144.000 h of video are uploaded to YouTube per day [13], 182.900.000.000 emails are sent per day [14] and 1.000.000.000 pieces of content are shared on Facebook per day [15].

This results in massive amounts of data. Facebook has one of the largest data warehouses in the world, storing more than 300 petabytes [16].

With such a large production of data and in a time when knowledge is one of the most precious assets, it is no wonder that knowledge discovery and data mining are of increasing importance.

The road for knowledge discovery and data mining projects is to increase systematization, as the area becomes more main stream.

This seems important because the trends in this area, currently, are to have larger projects (with larger amounts of data involved) and, at the same time, to have the people involved in those same projects with lower technical skills and very little time to experiment with different approaches [17].

Within the knowledge discovery projects, the area of requirements engineering is the one that can reap more benefits thanks to a higher level of systematization.

Firstly, because requirements engineering is particularly neglected in this type of projects. Some authors even argue that this type of projects should be based on the available data and not on stakeholders' requirements [18].

Secondly, because, being a less mature field, less systematization efforts have been made so far and when they occur, the participation of enterprise stakeholders will be improved and facilitated and the area will follow software engineering in general, that has improved in terms of customer satisfaction and time and budget compliance.

For these reasons, the research question was "How can systematization be brought into Knowledge Discovery projects, in general, and into their Requirement Engineering phase, in particular, aiming at improvements in their success rate?"

The research started by analysing the Knowledge Discovery process through a systematic review of the state-of-the-art in academia and industry regarding knowledge discovery and data mining process models. To conclude this review a comparing of the main process models found was made.

Then the Requirements engineering area was analysed in a similar way followed byocusing on requirements engineering for KD projects. It was found that requirements engineering for KD is different. That is why it is claimed here that a requirements engineering for KD process model is needed and SysPRE, a Systematized Process for Requirements Engineering designed specifically for KD projects is proposed.

SysPRE, began from an initial textual description which was then formally specified as a DEMO ontology [45]. This formal specification was instantiated in two case studies so that trivial and non-trivial errors could be identified and the necessary adjustments made.

SysPRE synthesises the knowledge obtained for the state-of-the-art reviews in a way that can be helpful for enterprises and other organizations with KD projects both for novice and expert users, with the hope of bringing improvements to the success rate of such projects.

2 Knowledge Discovery Process and Demo Specification

In this section the Knowledge Discovery Process (KDP) will be described as seen after analysing the existing process models listed in Sects. 2.1 and 2.2 with special detail with what regards Requirements Engineering within the KDP.

This specifically considers business KDPs, but this description would also be accurate for other types of organizations, namely governmental or non-profit.

2.1 Knowledge Discovery

The need for a process model stems from the fact that data mining is non-trivial. In 2006, Bernstein et al. referred that “there are many possible choices for each stage, and only some combinations are valid. Because of the large space and nontrivial interactions, both novices and data mining specialists need assistance” [19].

Still the need for a process model goes back to 1989, when it was first discussed during the IJCAI workshop on Knowledge Discovery in Databases (KDD) [20]. This was the original workshop that started the series of KDD workshops that, from 1995 onwards, grew into KDD conferences. Still, only in 1996 the first model was formally proposed.

This original KDD model consisted in nine steps: learning the application domain, understanding the domain and any relevant prior knowledge but also identifying the goal of the process; creating a target dataset; data cleaning and pre-processing; data reduction and projection; function of data mining selection (e.g., summarization, clustering); data mining algorithm(s) selection and specification of relevant parameters; data mining, which means the actual search for patterns; interpretation of the results; using discovered knowledge, which could be done in many ways, such as incorporating the knowledge into another system or simply generating a report of the findings.

From this model other models derived such as Ganesh et al. [21] and Adriaans and Zantinge [22] in 1996, Brachman and Anand [23] in 1997, Berry and Linoff [24], Cabena et al. [25], Knowledge Discovery Life Cycle (KDLC) model by Lee and Kerschberg [26] 1998 or Buchner et al. [27] in 1999.

The most widely used in the industry however was CRISP-DM [46]. Created in 1997 by a group of organizations involved in data mining (NCR, SPSS, Daimler-Chrysler and

OHRA). The first version was published in August 2000 [28]. Between 2006 and 2008 there were efforts to launch a second version of CRISP-DM, which was referred to as CRISP-DM 2.0, but no result was ever published.

The CRISP-DM model life cycle consists of six iterative steps: business understanding; data understanding; data preparation; modelling; evaluation; deploying.

To CRISP-DM many variations were proposed over the years, such as Rapid Collaborative Data Mining System (RAMSYS) model [31] in 2001, Data Mining for Industrial Engineering (DMIE) by Solarte [32] in 2002, Data Mining and Knowledge Discovery (DMKD) model by Cios and Kurgan [33] in 2005, Ontology Driven Knowledge Discovery (ODKD) by Gottgtroy [34] in 2007, Knowledge and Discovery and Communication Framework (KDCF) by Rennolls and AL-Shawabkeh [35] and ASD-DM by Alnoukari et al. [36] in 2008 or IKDDM by Osei-Bryson [37] in 2012.

Other models include Catalyst methodology in 2003 [30]. This methodology has two parts: business modelling and data mining. For each part, a detailed step-by-step methodology is suggested. Originally it was proposed both in printed form and online, and both formats followed a hyperlink structure.

Considering both parts of the methodology as a whole, we can say that it has six steps: business modelling; data preparation; tool selection; mining; refining; deploying.

What makes this methodology interesting is the level of detail that is includes in each step. It is very focused on what needs to be done and how it can be done. This is organized in what the author calls “boxes”. There are four types of “boxes”: Action Boxes, Discovery Boxes, Technique Boxes, and Example Boxes.

And finally SEMMA was created to be used is a specific application, SAS Enterprise Miner [29].

The acronym SEMMA stands for sample, explore, modify, model, assess, which are basically the five iterative steps proposed: sample, which consists of extracting sample data (optional step); explore, which means the exploring the data or the sample data in order to be able to simplify the model; modify, which can include any cleaning, preprocessing, reductions or projections deemed necessary; model, which is the actual search for patterns; assess, which is the evaluation and interpretation of the results.

SEMMA however is tied to the SAS Enterprise Miner tool and therefore overlooks any steps that are not related to the tool, namely any business understanding tasks.

2.2 Requirements Engineering

The IEEE Standard Glossary of Software Engineering Technology [38] defines a software requirement as:

1. A condition or capability needed by a user to solve a problem or achieve an objective.
2. A condition or capability that must be met or possessed by a system or system component to satisfy a contract, standard, specification, or other formally imposed document.
3. A documented representation of a condition or capability as in 1 or 2.

In short, a software requirement is something that we expect the software to meet.

In the studied methods there was a special focus on six, Waterfall by Winston Royce [12] in 1970, Spiral by Barry Boehm [39] in 1986, Rapid Application Development (RAD) by the New York Telephone Company in mid-1970s, becoming notorious in the early 90's by James Martin and his approach [40], Rational Unified Process (RUP) by the Rational Software Division of IBM [41], Agile proposed in 2001 in the Agile Manifesto [42] and Goal-Oriented Requirements Engineering (GORE).

2.3 PIF and CAP Analysis

To the KDP a Performa-Informa-Forma (PIF) analysis and a Coordination-Actors-Production (CAP) analysis were made with the goal to gain insight to what concepts and activities are important in the KD process. Namely, in terms of activities, the Performa items are the truly relevant ones and will later be the transactions of the DEMO specification of SysPRE.

Most of the Performa-Informa-Forma is being omitted remaining only the Performa items in italic. The Coordination-Actors-Production analysis was done simultaneously by enclosing a piece of text indicating an actor role between the brackets "[" and "]". Transaction's id (for instance T01) are also marked next to Performa items.

The knowledge discovery process begins {T01} when the [business analyst] *realizes that there is a business problem or opportunity* {T02} in which Knowledge Discovery and Data Mining might be helpful. More commonly, the [business analyst] starts with a question and needs certain information relevant to the decision he must make.

He or she starts by trying to *learn* {T03} as much as possible about the business and the application domain. He will *identify* the [stakeholders] {T04}. He will try to *understand what issues are important* for the [stakeholders] {T05}. The five core issues are [30]: product (goods or services, tangible or intangible); place; price; time; quantity.

The [business analyst] will *classify the knowledge discovery process* as {T06}:

- Demand driven - process is aimed to fulfil the information requirements of the users
- Data driven - process is aimed to discover the best use to the specific existing data
- Exploratory - process is designed to find how KD and DM in general can offer value within that specific business

He will try to discover any relevant prior knowledge, namely the currently existing solutions for the problem, and *identify the goal* for the project {T05}.

If it is an exploratory process, the [business analyst] will *identify several possible goals* {T05} and *review his stakeholders' identification* {T04} for each one (including the core issues that each one might be concerned with {T05}).

Since starting the project might have costs, the [business analyst] might have to ask for *approval* {T14} for the data mining project to the [business manager]. The [business manager] might *ask* {T13} the [project manager] for a cost and resources estimation so that he can *decide on the approval* {T14}. The [project manager] *will create the cost, time and resources estimates or a project plan* {T13}, if necessary. The [project manager] will hand these to the [business manager]. The [business manager] will *decide to go ahead or not* {T14}, that is, he will decide on the feasibility of the KD project. If the decision is to go ahead, the [project manager] might have to get the resources (human or otherwise) that are necessary and that were not available in the beginning.

If it is a demand driven project, the [business analyst] will then begin *eliciting specific requirements* {T07}. If it is a data driven project, the [business analyst] will then proceed by asking the [data analyst] to perform the data analysis. A hybrid approach is also possible, in which both will happen in parallel. For the *requirements elicitation* {T07}, the [business analyst] will *choose the elicitation techniques* {T08}, which might be one or more. He will execute them and document the resulting requirements from each technique at what is judged to be an appropriate level of detail. These requirements will be mostly information demand requirements, that is, requirements that describe why and how the [stakeholders] need specific information. The [business analyst] will also *elicit non-functional requirements* {T07}, and for that he will be particularly concerned with the delivery mechanism (how will the results be physically made available to the [end user]? What tools will the [user] employ to view it?), the format (will the [user] view the results in reports, dashboards, or other formats?) and the degree of interaction needed (to what extent must the [user] be able to manipulate the results following delivery?).

A detailed analysis of the requirements will be done by the [business analyst]. The [business analyst] and the multiple [stakeholders] will *negotiate* to:

- *Decide which requirements are accepted* {T09} (which, in fact, is the same as deciding the system boundaries or scope)
- *Do a triage and prioritization of the requirements* {T10}
- *Assess requirements risks* {T11}

The [business analyst] will *validate* {T12}, that is, check for completeness and for consistency the resulting requirements.

The *triage and prioritization* {T10} should be done after the *validation* {T12}, as the validation {T12} process might result in adding, changing or removing some requirements.

The [business analyst] will also need data, so he will ask the [data analyst]. Again, note that in a demand driven project this request will normally happen after the requirement elicitation {T07}, but in a supply driven project the data gathering that we will describe next will happen before the *requirement elicitation* {T07}. The [data analyst] will *look for the raw data* {T15} to use for the project. The data might come from databases, internal or external, or from other sources. It might also need still to be collected for this specific purpose. The [data analyst] will need to *select the data* {T16} and *decide if and when the data might need to be combined* {T17}. If the [data analyst] considers the data to be too large for an initial analysis, he might *consider using a sample* {T17} of the data.

The [data analyst] will also try to understand the data. To begin with, if the data was already available at the beginning of the project, the [data analyst] should find the business motivation to collect and store the data in the first place, as it might provide some insights. From the data understanding he might *suggest a possible hypotheses or objective* {T18} to the [business analyst]. He might also *identify constrains* {T19} that arise from the data, so he will inform the [business analyst] of the detected constrains.

Since the raw data might be incomplete, noisy or inconsistent, the [data engineer] will perform *data cleaning, pre-processing and transformation* {T17}. This might include filling missing values, normalization, discretization, reduction, projection or other techniques. The data cleaning, pre-processing and transformation is guided by the data itself and also by what data mining techniques are going to be used on the data. The [data miner] *selects the tool* {T20} to be used (for the same project, more than one tool might be used). For selecting the tool he will start by *identifying possible tools* {T28} and *decide on how he will compare them* {T20}, specifying the evaluation criteria that are important and how the evaluation will be performed (for instance, he might decide to run a specific algorithm using all the tools and a sample of the data). He will then proceed with the evaluation and *choose the tool* {T20} (or tools). The [data miner] also *selects the data mining technique* {T21} (e.g., summarization, classification, regression, clustering) and the *specific algorithms* {T22}. For the same project, more than one tool might be used, as well as more than one data mining technique and one algorithm.

Some authors believe the choice of data mining technique can be simplified *to four decisions* {T21}.

The [data miner] will entail the prepared data to the tool and be responsible for the *generation of the model* {T24}. This means he will have, for instance, to *decide on the appropriate parameters* {T23}.

After the actual data mining has occurred and the KD results are available, both the [domain expert] and the [strategic manager] will *analyse the results* {T25}.

The [domain expert] *analyses* {T25} the data mining result, in the sense that he *evaluates how the results fit his domain knowledge* {T25}, possibly resulting in the need for refining what was done previously through:

- *Creating new questions or hypothesis* {T18} for the [business analyst]
- *Pointing the need for new or more data* {T15} for the [data analyst]
- *Indicating the need to use a different function* {T21} *or algorithm* {T22} *or simply to adjust parameters* {T23} to the [data miner]

The [strategic manager] *interprets and evaluates* {T25} the data mining result, in the sense that he *evaluates how these results are relevant to or have an impact* {T25} on the current or future business situation.

The [knowledge engineer] will use the analysis results from the [domain expert] and the [strategic manager] and make sure the discovered knowledge is used. He will *specify* {T26} how the knowledge discovery result should be deployed, for instance he can decide that an annual report should be produced for the senior management. The knowledge discovery result will then be deployed to the [end users] as planned.

2.4 Transaction Result Table

From the Performa-Informa-Forma analysis and Coordination-Actors-Production analysis the Transaction Result Table (TRT) was the following.

This table shows the transactions (that correspond to the main tasks of the process) and the result types corresponding to each transaction. In the result types, we can see (between square brackets) the main concept that is being created or whose state is being changed.

The last transactions (T28 to T31) refer to the specification of an elicitation technique for requirements or, regarding the data mining stage, the specification of a tool, data

Table 1. Transaction Result Table

Transaction		Result type	
Id	Name	Id	Description
T01	Knowledge Discovery	R01	[knowledge discovery process] was realized
T02	Problem/Opportunity identification	R02	[problem/opportunity] was identified
T03	Problem/Opportunity analysis	R03	[problem/opportunity] was analysed
T04	Stakeholder identification	R04	[stakeholder] was identified
T05	Goal/core issue identification	R05	[goal/core issue] was identified
T06	Process classification	R06	[knowledge discovery process] was classified
T07	Requirement elicitation	R07	[requirement] was elicited
T08	Choice of elicitation technique	R08	[elicitation technique] was chosen
T09	Decision of scope	R09	decision on whether the [requirement] is in scope was made
T10	Requirement prioritization	R10	priority of [requirement] was defined
T11	Assessment of requirement risks	R11	risks of [requirement] were assessed
T12	Requirement validation	R12	[requirement] was validated
T13	Cost and resources estimation	R13	[cost and resources] were estimated
T14	Go-no-go Decision	R14	go-no-go decision of [knowledge discovery process] was made
T15	Data source identification	R15	[data source] was identified
T16	Data selection	R16	[data] was selected
T17	Data preparation	R17	[data] was prepared
T18	Hypothesis creation	R18	[hypothesis] was created
T19	Data constrain identification	R19	[data constrain] was identified
T20	Choice of tool	R20	tool was chosen for [result]
T21	Choice of data mining technique	R21	data mining technique was chosen for [result]
T22	Choice of algorithm	R22	algorithm was chosen for [result]
T23	Choice of data mining parameter	R23	data mining parameter was chosen for [result]
T24	Result obtention	R24	[result] was obtained
T25	Result analysis	R25	[result] was analysed
T26	Deployment specification	R26	[deployment] was specified
T27	KD area artefact management	R27	[KD area artefact] was managed
T28	Elicitation technique specification	R28	[elicitation technique] was specified
T29	Tool specification	R29	[tool] was specified
T30	Data mining technique specification	R30	[data mining technique] was specified
T31	Algorithm specification	R31	[algorithm] was specified
T32	Data mining parameter specification	R32	[data mining parameter] was specified

mining technique, algorithm or data mining parameter that was previously unknown to the system. This is necessary as the knowledge discovery and data mining area is very dynamic and it is very likely that new tools, data mining techniques, algorithms or data mining parameters need to be considered.

T27 is the transaction that manages all this. The elicitation techniques, tools, data mining techniques, algorithms and data mining parameters are referred to as artefacts in the context of T27 (KD area artefact management) (Table 1).

2.5 Object Fact Diagram

Due to space constrains the DEMO's Actor Transaction Diagram and the Process step diagram are omitted in this paper.

We then specified the DEMO's Object Fact Diagram (OFD).

In this diagram it can be seen the classes that correspond to the main concepts identified in the DEMO transactions of the Transaction Result Table, as well as other related classes, the fact types that are associated with each class and the cardinalities and dependence laws.

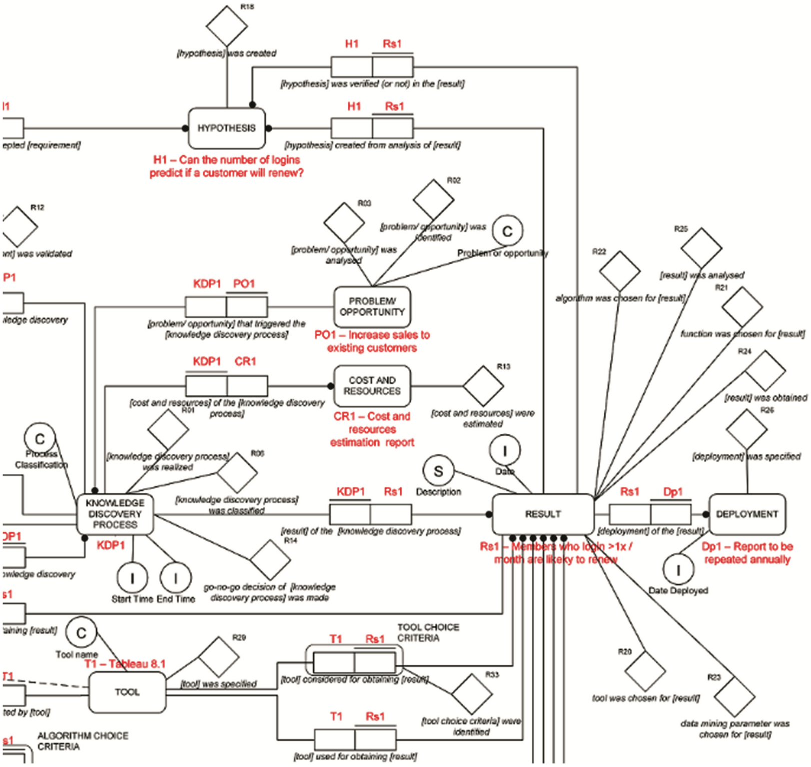

R18
[hypothesis] was created
H1
Rs1
[hypothesis] was verified (or not) in the [result]
H1
epted [requirement]
HYPOTHESIS
H1
Rs1
[hypothesis] created from analysis of [result]
H1 – Can the number of logins predict if a customer will renew?
R12
ent] was validated
R03
R02
[problem/ opportunity] was analysed
[problem/ opportunity] was identified
C
Problem or opportunity
R22
R25
[result] was analysed
algorithm was chosen for [result]
R21
function was chosen for [result]
P1
KDP1
PO1
PROBLEM/ OPPORTUNITY
wledge discovery
[problem/ opportunity] that triggered the [knowledge discovery process]
PO1 – Increase sales to existing customers
KDP1
CR1
COST AND RESOURCES
R13
R24
[result] was obtained
[cost and resources] of the [knowledge discovery process]
R01
CR1 – Cost and resources estimation report
[cost and resources] were estimated
R26
C
Process Classification
[knowledge discovery process] was realized
R06
I
Date
[deployment] was specified
S
Description
KDP1
Rs1
Rs1
Dp1
KNOWLEDGE DISCOVERY PROCESS
[knowledge discovery process] was classified
RESULT
DEPLOYMENT
DP1
KDP1
[result] of the [knowledge discovery process]
R14
[deployment] of the [result]
Rs1 – Members who login >1x / month are likeky to renew
Dp1 – Report to be repeated annually
owledge discovery
I
Start Time
I
End Time
go-no-go decision of [knowledge discovery process] was made
I
Date Deployed
s1
aining [result]
C
Tool name
T1 – Tableau 8.1
R29
TOOL CHOICE CRITERIA
T1
Rs1
R20
T1
TOOL
[tool] was specified
[tool] considered for obtaining [result]
R33
tool was chosen for [result]
R23
ted by [tool]
T1
Rs1
[tool choice criteria] were identified
data mining parameter was chosen for [result]
s1
ALGORITHM CHOICE CRITERIA
[tool] used for obtaining [result]

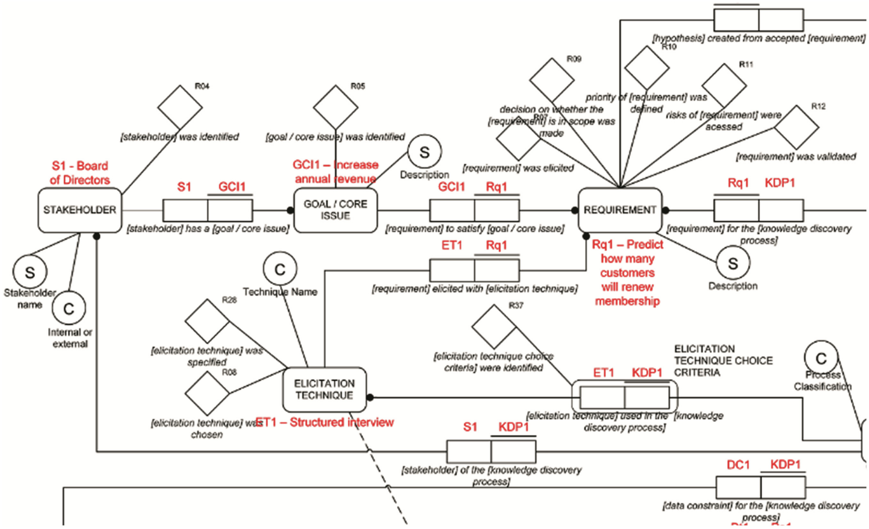
Fig. 1. Object Fact Diagram (Part 1)

In the image marked in red are comments of an instantiation of each class derived from a concrete case of a real organization, so that the interpretation of the diagram is easier (Fig. 1).

The main class of this OFD is the KNOWLEDGE DISCOVERY PROCESS (KDP), related to the main transaction T01. Each instance of this class will specify a particular KDP. Most of the classes that follow (in all caps text) are self-explanatory, so will presented as the example is described.

Instances of the class PROBLEM/OPPORTUNITY specify a problem or an opportunity that triggered the KDP. Let's say that a company wants to increase its sales to existing customers. The company we are considering, sells memberships, so basically they'll want to increase the percentage of customers that renew their memberships. This is the problem/opportunity.

One STAKEHOLDER is the Board of Directors. This particular stakeholder had a GOAL/CORE ISSUE: they want to increase the annual revenue. Using one or more ELICITATION TECHNIQUES, a REQUIREMENT to satisfy the above GOAL/CORE ISSUE was elicited: Predict how many customers will renew. One possible ELICITATION TECHNIQUE is a structured interview, but many others were possible (Fig. 2).



R04
[stakeholder] was identified
R05
[goal / core issue] was identified
S1 - Board of Directors
STAKEHOLDER
S1
GCI1
[stakeholder] has a [goal / core issue]
GCI1 – Increase annual revenue
GOAL / CORE ISSUE
S
Description
S
Stakeholder name
C
Internal or external
C
Technique Name
R28
[elicitation technique] was specified
R08
[elicitation technique] was chosen
ELICITATION TECHNIQUE
ET1 – Structured interview
[hypothesis] created from accepted [requirement]
R09
R10
R11
R12
R07
priority of [requirement] was defined
decision on whether the [requirement] is in scope was made
risks of [requirement] were acessed
[requirement] was elicited
[requirement] was validated
GCI1
Rq1
[requirement] to satisfy [goal / core issue]
REQUIREMENT
Rq1
KDP1
[requirement] for the [knowledge discovery process]
ET1
Rq1
[requirement] elicited with [elicitation technique]
Rq1 – Predict how many customers will renew membership
S
Description
R37
[elicitation technique choice criteria] were identified
ET1
KDP1
ELICITATION TECHNIQUE CHOICE CRITERIA
C
Process Classification
[elicitation technique] used in the [knowledge discovery process]
S1
KDP1
[stakeholder] of the [knowledge discovery process]
DC1
KDP1
[data constraint] for the [knowledge discovery process]

Fig. 2. Object Fact Diagram (Part 2)

Normally several STAKEHOLDERS will be identified (T04), each with one or more GOAL/CORE ISSUE from which several REQUIREMENTS will stem and be elicited (T05).

From the accepted REQUIREMENTS, we will then proceed to create a HYPOTHESIS that can be tested in a KDP. For this case, one of the tested HYPOTHESIS was if the number of logins can be used to predict if a customer will renew. The link between HYPOTHESIS and REQUIREMENTS is important for traceability.

In the end, the RESULT of the KDP will either confirm this hypothesis or not. For the KDP, there needs to be an estimation of COST AND RESOURCES, so that a Go-no-go decision (T14) can take place.

If the KDP proceeds, instances of classes corresponding to the DATA SOURCE (from which DATA will be selected and prepared), the data mining TOOL (in this case, Tableau 8.1), the type of DATA MINING TECHNIQUE (in this case, classification) and the ALGORITHM (in this case, AdaBoost) will be used to obtain a particular RESULT. The ALGORITHM might require a DATA MINING PARAMETER (or more) to be set. In this case we could change the value for a_t weight, but did not.

The KD AREA ARTEFACT is a generalization that includes ELICITATION TECHNIQUE, TOOL, DATA MINING TECHNIQUE, ALGORITHM and DATA MINING PARAMETER. The management of these artefacts (T27) involves specifying an artefact that was previously unknown to the system whenever needed (T28, T29, T30, T31, T32). The can then be chosen for use (T08, T20, T21, T22, T23) using ELICITATION TECHNIQUE CHOICE CRITERIA, TOOL CHOICE CRITERIA, DATA MINING TECHNIQUE CHOICE CRITERIA, ALGORITHM CHOICE CRITERIA or DATA MINING PARAMETER CHOICE CRITERIA respectively. It is important that the choice criteria are all documented, which is why all these classes appear (Fig. 3).

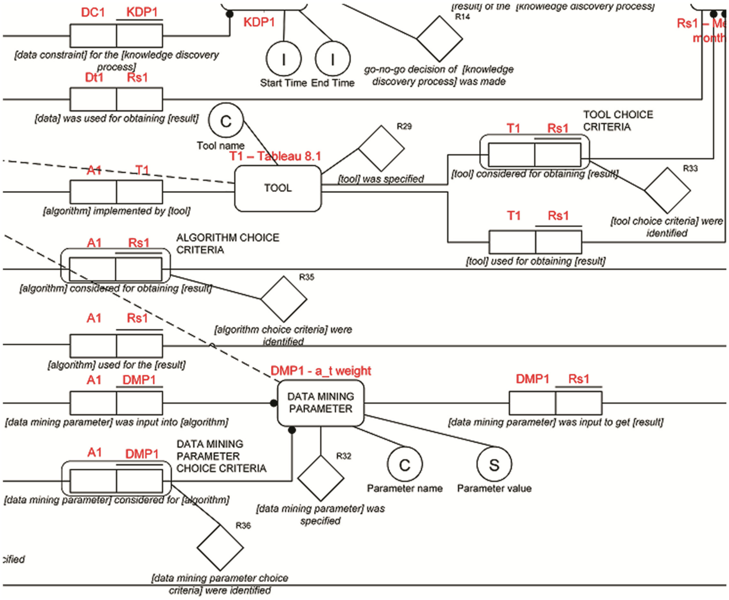

DC1 KDP1
[data constraint] for the [knowledge discovery process]
KDP1
I
I
Start Time End Time
[result] of the [knowledge discovery process]
R14
go-no-go decision of [knowledge discovery process] was made
Rs1 – Me month
Dt1 Rs1
[data] was used for obtaining [result]
C
Tool name
T1 – Tableau 8.1
R29
TOOL CHOICE CRITERIA
T1 Rs1
A1 T1
TOOL
[tool] was specified
[tool] considered for obtaining [result]
R33
[algorithm] implemented by [tool]
T1 Rs1
[tool choice criteria] were identified
A1 Rs1
ALGORITHM CHOICE CRITERIA
[tool] used for obtaining [result]
[algorithm] considered for obtaining [result]
R35
A1 Rs1
[algorithm choice criteria] were identified
[algorithm] used for the [result]
DMP1 - a_t weight
A1 DMP1
DATA MINING PARAMETER
DMP1 Rs1
[data mining parameter] was input into [algorithm]
[data mining parameter] was input to get [result]
A1 DMP1
DATA MINING PARAMETER CHOICE CRITERIA
R32
C
S
Parameter name
Parameter value
[data mining parameter] considered for [algorithm]
[data mining parameter] was specified
R36
cified
[data mining parameter choice criteria] were identified

Fig. 3. Object Fact Diagram (Part 3)

From the DATA might result some kind of DATA CONSTRAINT. In this case, it was very noticeable that the customer age was not available. The identified DATA CONSTRAINTS affected the KDP.

As mentioned, the execution of a particular algorithm with particular parameters and applied to a particular data, in the context of a KDP will produce a particular RESULT

- for example, a classification model or a set of association rules. For the case study at hand, we found that the members who login more than once per month are more likely to renew.

The RESULT will be target of an analysis (T25). From such analysis the conclusion might be that new hypothesis needs to be formulated and/or new data, tools, data mining techniques or specific algorithms applied so that refined or alternative results are found. If none of this is necessary, the DEPLOYMENT of a RESULT can also be specified. For example, in this case it was decided that an annual report with the obtained result was to be produced (Fig. 4).

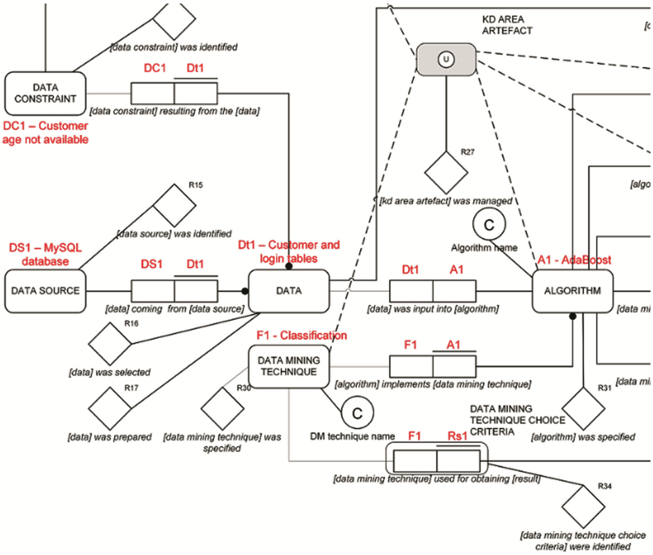

KD AREA
ARTEFACT
U
R27
[kd area artefact] was managed
[data constraint] was identified
DC1
Dt1
DATA CONSTRAINT
[data constraint] resulting from the [data]
DC1 – Customer age not available
R15
[data source] was identified
DS1 – MySQL database
DS1
Dt1
DATA SOURCE
[data] coming from [data source]
Dt1 – Customer and login tables
DATA
C
Algorithm name
A1 - AdaBoost
Dt1
A1
[data] was input into [algorithm]
ALGORITHM
[algo
[data mi
R16
[data] was selected
F1 - Classification
DATA MINING TECHNIQUE
F1
A1
[algorithm] implements [data mining technique]
R31
[data mi
R17
[data] was prepared
R30
[data mining technique] was specified
C
DM technique name
F1
Rs1
DATA MINING TECHNIQUE CHOICE CRITERIA
[algorithm] was specified
[data mining technique] used for obtaining [result]
R34
[data mining technique choice criteria] were identified

Fig. 4. Object Fact Diagram (Part 4)

3 Discussion and Conclusion

Other efforts have been made regarding knowledge discovery ontologies such as OntoDM [43] or Knowledge Discovery Ontology [44], but focus in great detail in the knowledge discovery process itself and don't show any particular insight regarding its surroundings, like the business side information.

This DEMO based ontology gives several interesting insights. Thanks to the specified classes, for a particular problem/opportunity we can keep a record of detailed and important information of a respective KDP. Namely keep a consistent and integrated record of important business side information like the stakeholders, requirements, hypothesis and costs; and also of the technical side like tools, sources and algorithms used. The class RESULT is pivotal in the sense that each instance will include not only the patterns obtained using the data mining technique, but also an analysis of the results which may lead to the formulation of new hypothesis and requirements on the business side.

Having SysPRE, an ontology that represents both the KD work in general and the RE for KD work specifically can help technical roles not lose track of the big picture while working on the task at hand. Also, since it is understandable not only by the technical roles involved, but also by other stakeholders, SysPRE can foster a more effective dialogue between them.

This ontology can encourage knowledge reuse of the KD process or RE KD process itself in a consistent and integrated fashion because it enables keeping a record of iterations and refinements of a particular process in a highly structured way. This way, it's hoped to make enterprises become aware of their own KD process and RE process in the KD projects, but also to improve such processes in reality, namely in terms of the success rate. In other words, this can help the lessons learned from the past be reused to improve the present.

The main contribution of this paper is to provide a systematization that can be applied to KD projects in general and to the requirements engineering process in such processes in particular.

Having a short, plain text description of a generic KD process with emphasis on RE that was proposed after doing a thorough literature review can be useful for novices in the area, both in the research and in the industry communities.

Having the SysPRE formal ontology can be helpful within the organization using them because it can:

- Enable keeping a record of iterations and refinements of a particular process in a highly structured way.
- Make enterprises (and specifically decision makers within the enterprise) become aware of their own KD process and RE process in the KD projects.
- Assist enterprises that want to improve their own KD process and RE process in the KD projects.
- Help each technical role involved keep an eye on the big picture while working on whatever task they are working on at that specific moment.

Having the SysPRE formal ontology can also be helpful for the communication between the organization and other stakeholders because, despite being formal, they are understandable and do sum up a lot of information in a graphical way.

Acknowledgments. This work was partially funded by FCT/MCTES LARSyS (UID/EEA/50009/2013 (2015-2017)).

References

1. The Standish Group, "1994 CHAOS Report," (1994)
2. Glass, R.L.: IT Failure Rates-70% or 10–15%? IEEE Softw. **22**(3), 110–112 (2005)
3. Jørgensen, M., Moløkken-Østvold, K.: How large are software cost overruns? A review of the 1994 CHAOS report. Inf. Softw. Technol. **48**(4), 297–301 (2006)

4. Glass, R.L.: The Standish report: does it really describe a software crisis? ACM Commun. **49**(8), 15–16 (2006)
5. Eveleens, J., Verhoef, C.: The Rise and fall of the Chaos report figures. IEEE Softw. **27**(1), 30–36 (2010)
6. Pohl, K.: Requirements Engineering: Fundamentals, Principles, and Techniques. Springer, Heidelberg (2010)
7. El Emam, K., Koru, A.G.: A replicated survey of IT software project failures. IEEE Softw. **25**(5), 84–90 (2008)
8. Atkins, C.: An Investigation of the Impact of Requirements Engineering Skills on Project Success. East Tennessee State University (2013)
9. Paiva, A., Varajão, J., Dominguez, C.: Principais aspectos na avaliação do sucesso de projectos de desenvolvimento de software. Há alguma relação com o que é considerado noutras indústrias? Interciencia **36**(3), 200–204 (2011)
10. Wateridge, J.: How can IS/IT projects be measured for success? Int. J. Proj. Manag. **16**(1), 59–63 (1998)
11. Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P.: Knowledge discovery and data mining: towards a unifying framework. KDD **96**, 82–88 (1996)
12. Royce, W.W.: Managing the development of large software systems. In: Proceedings of IEEE WESCON, vol. 26 (1970)
13. Statistics - YouTube. https://www.youtube.com/yt/press/statistics.html
14. Radicati, S. (ed.) Email Statistics Report 2013–2017 Executive Summary, April 2013
15. Manyika, J., Chui, M., Brown, B., Bughin, J.: Big Data: the Next Frontier for Innovation, Competition, and Productivity. McKinsey & Company, May 2011. http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation
16. Traverso, M.: Presto: interacting with petabytes of data at Facebook. Research at Facebook, November 2013. https://research.facebook.com/blog/1489667567986457/presto-interacting-with-petabytes-of-data-at-facebook/
17. Pytel, P., Britos, P., García-Martínez, R.: A proposal of effort estimation method for information mining projects oriented to SMEs. In: Poels, G. (ed.) CONFENIS 2012. LNBIP, vol. 139, pp. 58–74. Springer, Heidelberg (2013). doi:10.1007/978-3-642-36611-6_5
18. Inmon, W.H.: Building the Data Warehouse. Wiley, New York (2005)
19. Bernstein, A., Provost, F., Hill, S.: Toward intelligent assistance for a data mining process: an ontology-based approach for cost-sensitive classification. IEEE Trans. Knowl. Data Eng. **17**(4), 503–518 (2005)
20. Piatetsky-Shapiro, G.: Knowledge discovery in real databases: a report on the IJCAI-89 Workshop. AI Mag. **11**(4), 68 (1990)
21. Ganesh, M., Han, E.H., Kumar, V., Shekhar, S., Srivastava, J.: Visual Data Mining: Framework and Algorithm Development. Department of Civil Engineering, University of Minnesota, MN USA (1996)
22. Adriaans, P., Zantinge, D.: Data Mining. Addison-Wesley, Reading (1996)
23. Brachman, R.J., Anand, T.: Advances in knowledge discovery and data mining. In: Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R. (eds.) American Association for Artificial Intelligence, Menlo Park, pp. 37–57 (1996)
24. Berry, M.J., Linoff, G.: Data Mining Techniques: For Marketing, Sales, and Customer Support. Wiley, New York (1997)
25. Cabena, P., Hadjinian, P., Stadler, R., Verhees, J., Zanasi, A.: Discovering Data Mining: From Concept to Implementation. Prentice Hall, Upper Saddle River (1997)

26. Lee, S.W., Kerschberg, L.: A methodology and life cycle model for data mining and knowledge discovery in precision agriculture. In: IEEE International Conference on Systems, Man, and Cybernetics, vol. 3, pp. 2882–2887 (1998)
27. Buchner, A.G., Mulvenna, M.D., Anand, S.S., Hughes, J.G.: An internet-enabled knowledge discovery process. In: Proceedings of the 9th International Database Conference, Hong Kong, vol. 1999, pp. 13–27 (1999)
28. Wirth, R., Hipp, J.: CRISP-DM: towards a standard process model for data mining. In: Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining, pp. 29–39 (2000)
29. SAS Institute: SEMMA (2005). http://www.sas.com/offices/europe/uk/technologies/analytics/datamining/miner/semma.html
30. Pyle, D.: Business Modeling and Data Mining. Morgan Kaufmann, San Mateo (2003)
31. Moyle, S., Jorge, A.: RAMSYS-A methodology for supporting rapid remote collaborative data mining projects. In: ECML/PKDD01 Workshop: Integrating Aspects of Data Mining, Decision Support and Meta-learning (IDDM-2001) (2001)
32. Solarte, J.: A proposed data mining methodology and its application to industrial engineering. Masters Theses, August 2002
33. Cios, K.J., Kurgan, L.A.: Trends in data mining and knowledge discovery. In: Pal, N.R., Jain, L. (eds.) Advanced Techniques in Knowledge Discovery and Data Mining, pp. 1–26. Springer, London (2005)
34. Gottgtroy, P.: Ontology driven knowledge discovery process: a proposal to integrate ontology engineering and KDD. (2007)
35. Rennolls, K., Al-Shawabkeh, A.: Formal structures for data mining, knowledge discovery and communication in a knowledge management environment. Intell. Data Anal. **12**(2), 147–163 (2008)
36. Alnoukari, M., Alzoabi, Z., Hanna, S.: Applying adaptive software development (ASD) agile modeling on predictive data mining applications: ASD-DM Methodology. In: International Symposium on Information Technology, ITSim 2008, vol. 2, pp. 1–6 (2008)
37. Osei-Bryson, K.-M.: A context-aware data mining process model based framework for supporting evaluation of data mining results. Expert Syst. Appl. **39**(1), 1156–1164 (2012)
38. IEEE Computer Society, "IEEE Standard Glossary of Software Engineering Terminology," IEEE Std 61012-1990, pp. 1–84, December 1990
39. Boehm, B.: A spiral model of software development and enhancement. SIGSOFT Softw. Eng. Notes **11**(4), 14–24 (1986)
40. Martin, J.: Rapid Application Development. Mac Millan (1991)
41. IBM Rational software and systems delivery, 26 August 2014. http://www-01.ibm.com/software/rational/
42. Beck, K., Beedle, M., Bennekum, A.: Agile Manifesto (2001). http://www.agilemanifesto.org/
43. Panov, P., Soldatova, L., Džeroski, S.: OntoDM-KDD: Ontology for Representing the Knowledge Discovery Process. In: Fürnkranz, J., Hüllermeier, E., Higuchi, T. (eds.) DS 2013. LNCS (LNAI), vol. 8140, pp. 126–140. Springer, Heidelberg (2013). doi:10.1007/978-3-642-40897-7_9
44. Zakova, M., Kremen, P., Zelezny, F., Lavrac, N.: Automating knowledge discovery workflow composition through ontology-based planning. IEEE Trans. Autom. Sci. Eng. **8**(2), 253–264 (2011)
45. Dietz J.L.: Enterprise ontology - understanding the essence of organizational operation. In: Chen CS., Filipe J., Seruca I., Cordeiro J. (eds) Enterprise Information Systems VII, pp. 19–30. Springer, Dordrecht (2007)
46. Piatetsky-Shapiro, G.: KDNuggets, "Poll: Data Mining Methodology," (2014). http://www.kdnuggets.com/polls/2014/analytics-data-mining-data-science-methodology.html