# Extension of iStar for Big data projects

Chabane Djeddi
LIRE Laboratory,
Constantine 2- A. Mehri University, Algeria
chabane.djeddi@univ-constantine2.dz

Nacer eddine Zarour
LIRE Laboratory,
Constantine 2- A. Mehri University, Algeria
nasro.zarour@univ-constantine2.dz

Pierre-Jean Charrel
IRIT Laboratory,
Toulouse 2 Jean Jaurès University, France
charrel@univ-tlse2.fr



## Abstract

Big data is characterized by the volume, variety, velocity, and complexity of the data which make it very difficult to handle. On the other hand, requirements engineering (RE) is very important for the success of any software system. As a result, the importance of requirements engineering for Big data projects is evident, but there is no RE method to undertake them. We have analyzed the fields of Big data and RE to figure out how the RE can allow taking into account the properties of Big data. This paper presents BIStar which is an extension of iStar for Big data projects to support its properties in the elicitation step of RE process. Our extension undertakes the characteristics of Big data, which allow a better elicitation of the requirements and therefore, it facilitates data analysis.

**Keywords** - 1st Big data, 2nd Requirements engineering, 3rd iStar, 4th iStar extension

## 1. Introduction

Compagnies store a large amount of data every day as transactions that are important to them. However, over time, the management of these data by traditional systems becomes impossible, even regarding in terms of analysis time, it becomes challenging to guarantee efficient data processing in a short time. We find structured data but also semi-structured data and even unstructured data. This heterogeneity generates data incompatibility issues that threaten integrity and consistency.

Big data has its own properties (Volume, Velocity, Variety, Veracity, and Value) [ChML14], [KaWG13], [Madd12], [OtPe15], [ShOt16], which make it crucial and specific.

Authors in [OtPe15], [ShOt16] confirmed that there is the necessity for the Big data software to include all the three parameters (functional feature, time constraint, and verifiable during some period) to completely define the requirement specification for Big data projects. But all of the existing models are not including the constrained time and verifiable time, and specify the requirement only in terms of the functional features. Until now, there is no work to create or to adapt an existing RE method for Big data projects.

The modeling languages are classified into two classes: (i) Domain Specific modeling language (DSML) to model only one given domain, (ii) General Purpose Modeling Language (GPML), to model any domain [GCAH18]. iStar [Ref96] is a GPML to support the elicitation step in RE for any domain. It is widely used and adopted by the research community[GCAH18].

iStar was extended in order to be adopted in many domains like (security, data warehouse, social-technical systems, etc.) [GCAH18]. In our work, we propose BiStar (for Big data iStar) which is a DSML dedicated to the modeling of requirements for Big data. BiStar is based essentially on iStar, to not recreate all from scratch. We extend iStar to undertake the properties of Big data. Like that, we benefit from the iStar and add to it what we need and what is specific for Big data.

## 2. Literature reviews

In this section, we describe briefly the domains of the requirement engineering and the Big data that are related to our work.

### 2.1 Requirements engineering

The primary criterion for the success of any software is the degree of satisfaction of the goals fixed by the stakeholders. The requirements engineering (RE) is the process of discovery of these goals [NuEa00].


In: Proceedings of the 3rd Edition of the International Conference on Advanced Aspects of Software Engineering (ICAASE18), Constantine, Algeria, 1,2-December-2018, published at http://ceur-ws.org

‘’Requirements engineering is the branch of software engineering concerned with the real-world goals for, functions of, and constraints on software systems. It is also concerned with the relationship of these factors to precise specifications of software behavior, and to their evolution over time and across software families." [IeAI97].
The objective of RE is to know the requirements of the stakeholders and to verify them in order to arrive at an agreement on the requirements. To fulfill this, we perform the activities of elicitation, negotiation, documentation, validation, and management of the requirements. One of the difficult parts in building a software program is to decide what the software should exactly do. RE helps us to understand the problem. By studying the RE specifications precisely, we can even estimate the cost of the project. Moreover, RE also helps to know the limits of our system [MiNa11].

### 2.1.1 The steps of RE

RE is usually divided into four steps [KoSo98] (i) Requirements elicitation (ii) Requirements analysis and negotiation (iii) Requirements documentation (vi) Requirements validation.
Requirements elicitation: serves to capture the requirements and it is usually divided into five sub-steps [ZoCo05], Understanding the application domain, Identifying the sources of requirements, Analyzing the stakeholders, Selecting the techniques, Approaches, and Tools to use, Eliciting the requirements from stakeholders and other sources.
Requirements analysis and negotiation: focuses on the review, understanding of the elicited requirements and their verification for quality in terms of accuracy, completeness, clarity, and consistency.
Requirements documentation: we document the requirements obtained from previous steps. Requirements document can be considered as a base for controlling changes and evaluating future products and processes (system design, system test cases and validation) [MiNa11].
Requirements validation: It is done for controlling the quality. it means confirming that requirements are complete and well- written and supply needs of customer. This step may continue repeating other requirements development phases because of identified deficiencies, gap between requirements, additional information and other issues. Implemented software product is validated in software life cycle test phase on the basis of its requirements.
In this work, we are interested in the first step, which is requirement elicitation; because it is indispensable for any RE step and we cannot do any others steps without it.

### 2.1.2 The approaches of RE

We find in the literature [ZoCo05] that there are three basics approaches of RE (i) Goal Based Approaches, (ii) Scenarios Based Approaches, (iii) Viewpoints Based Approaches. These approaches can be modified or mixed to create new approaches.
The fundamental premise of goal based approaches (GORE) is high-level goals. These goals represent objectives for the system, they are decomposed (e.g. usually using AND and OR relationships) and elaborated (e.g. with “Why” and “How” questioning) into sub goals and then further refined, in such a way, elementary requirements are elicited [ZoCo05].
Several methods can be considered as belonging to GORE: iStar Framework [Ref96], NFR [ChPr09], KAOS [Vanl01]. Among all GORE methods, KAOS and iStar have been the most cited [WeOP09]. In our work, we choose iStar to extend because it is very used in academic research, and it is properly extensible [GCAH18].
Scenarios Based Approaches use narrative and specific descriptions of current and future processes including actions and interactions between the users and the system. Like use cases, scenarios do not typically consider the internal structure of the system, and require an incremental and interactive approach to their development. Naturally, it is important when using scenarios to collect all the potential exceptions for each step [ZoCo05].
Viewpoints Based Approaches: aim to model the domain from different perspectives in order to develop a complete and consistent description of the target system [ZoCo05]. Initially, the requirements are opaque, informal and only expressed through personal views. These views reflect the skills, objectives and roles of each participant. The elicitation activity is, therefore, a collective activity. The expression of multiple views allows for better elicitation of requirements.

### 2.2 Big data

In this section, we will present the Big data passing briefly through its definitions, properties, and importance.
There is no exact definition of Big data even though several definitions have appeared. Big data means a large dataset that cannot be processed by traditional tools [ChML14]. Big data can be seen from several perspectives, (i) on the infrastructure perspective Big data is seen as a significant amount of data characterized by (Volume, Velocity, Variety, Veracity, and Value), (ii) on the analysis perspective Big data is seen as events, (iii) on the business perspective Big data can be considered as the output that can be used directly for the improvement of the work [OtPe15], [ShOt16]. The most crucial problem is not how to store data, but rather to analyze heterogeneous data in a short time [Madd12].
There is a solid relationship between Big data and other technologies such as Cloud and IoT. Cloud can be an infrastructure for Big data, and IoT is considered the most massive source of Big data [ChML14]. Consequently, our contribution in Big data will influence other technologies.

#### 2.2.1 The properties of Big data

The Variety: the data manipulated today are not from a single representation, we have structured data, but also we have semi-structured data and even unstructured data such as web pages, social networks, making it very difficult to manipulate these data using traditional systems [ChML14], [KaWG13].
The volume : the name itself in the word Big data means that volume takes an important role in the creation of the Big data concept since the data handled today are in quantity of zettabytes at most large companies, this is of course, one of the limitations of traditional systems [ChML14], [KaWG13].
The velocity: the speed of incoming data from various sources is so critical, which make it difficult for the traditional systems to undertake the situation [KaWG13].
The value: the stored data is important. A user can execute some queries against stored data or may misuse existing data, and this can cause false results for decision makers [KaWG13].
The Complexity: This is how to ensure the correlation and the links between the data, because the latter in a Big data is collected from several heterogeneous sources, and that is very important to guarantee the integrity of the data, and not to be found in unmanageable situations [KaWG13].

#### 2.2.2 The importance of Big data

Big data have great importance in a lot of fields (the industry, risk analysis, social networks).
In the industry: Companies store a large amount of data every day as transactions that are important to them. However, over time, the management of these data by traditional systems becomes impossible. Traditional systems cannot support a large amount of data, here we can clearly see the utility of Big data for businesses, because it allows them to store a significant amount of data for a long time [KaWG13].
In risk analysis: Many companies use their data to calculate risk. Without Big data technologies, they use small amounts. With their arrival, it becomes possible to analyze a large amount of data, which allows better risk management [KaWG13].
In social networks: the most common use of Big data is in the areas of social networks and user preferences. Social networks use a large amount of data collected from user reviews and choices. That way, they can analyze the data and make it known that they are the preferences of the users in a short time, in order to improve their products and to change their decisions to have a good position in the market [KaWG13].

## 3. Case study

We have chosen to present the case study in this section in order to be able to use it in the modeling with iStar and BiStar (iStar extended) that we propose in the following sections. This example will accompany us throughout the paper.
We will take an example of the presidential elections of 2019 in Algeria. The community of a camp wants to increase the chance of success of its candidate. For that, they want to create a Big data project to study the opinions of the people, which allow them to know the keys for which they can focus in order to lunch targeted advertisements to improve the chances of success of their candidate.

To do this, they collect data from social networks, and analyze them to know the essentials points in the opinion of the different categories of population. On these points, they make a presidential plan and present it to the people. After that, they collect the opinion of people to make changes to the plan and make targeted advertising.
This example is a Big data project, because we will manipulate a large amount of data with different natures (structured, semi-structured and even unstructured) within a limited time. Therefore, these data cannot be processed using traditional systems.

## 4. iStar

In this session, we explain the iStar method, as well as their diagrams. iStar [DaFH16], [I*wi00], [Ref96] is a goal-oriented RE method, it is very used for requirements elicitation. We first start with the identification of the actors and the relations of the strategic dependencies between them, and then we detail the reasoning of each actor. It consists of two models: The Strategic Dependency (SD) Model, and The Strategic Rationale (SR) Model.

### 4.1 The Strategic Dependency (SD) Model

The strategic dependency model represents a network of strategic dependencies between the different actors of the future system. One actor (the dependee) depends on another one (the depender) to accomplish a goal. There are nodes and links between them, the nodes represent the actors, and the links represent the dependencies. There are four types of dependencies, (i) Goal dependency serves to present a dependency to accomplish a goal, (ii) Task dependency serves to present a task dependency between two actors, (iii) Resource dependency serves to present a resource dependency, (the depender) depends on (the dependee) to offer it a resource, (iv) Softgoal dependency serves to present a dependency of performance between two actors.
Figure [1] represents the application of the strategic dependency (SD) model of the iStar method on the case study of presidential elections. The ‘’candidate‘’ depends on the ‘’elector‘’ for the goal of winning the elections. The ‘’system to be developed‘’ depends on the ‘’candidate‘’ to accomplish the task of offering him its information. The ‘’elector‘’ depends on the ‘’system to develop‘’ to launch targeted advertising. ‘’Social networks‘’ depend on the ‘’elector‘’ to collect information about their preferences. The ‘’system to be developed‘’ depends on the ‘’social networks‘’ to receive the elector information resource.
The ‘’advertising manager‘’ depends on the ‘’system to be developed‘’ to provide the summary information on electors. The ‘’system to be developed‘’ depends on the ‘’advertising manager‘’ to accomplish the goal of developing targeted advertising.

### 4.2 The Strategic Rationale (SR) Model

The Strategic Rationale (SR) Model is used to detail the reasoning of each actor apart. We represent what happens inside an actor, which allows a deep understanding of the process.
Figure [2] shows the application of the strategic rationale model on the case of study of the presidential elections.

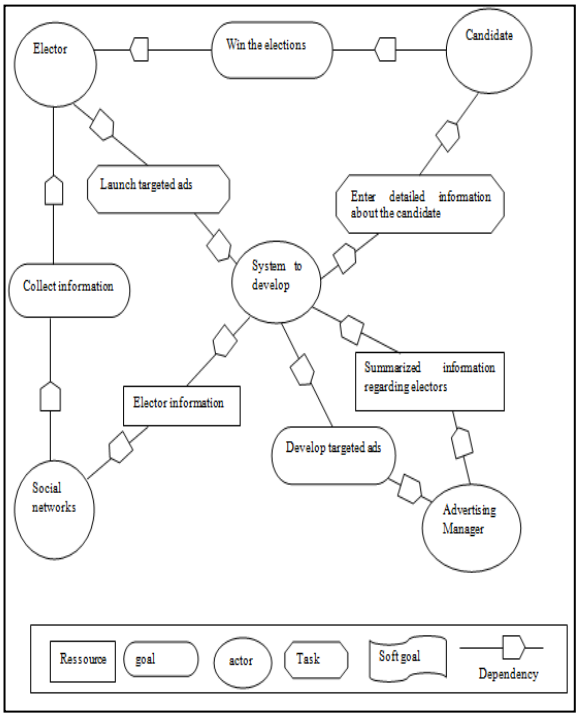


Figure 1 : Strategic dependency (SD) model for elections

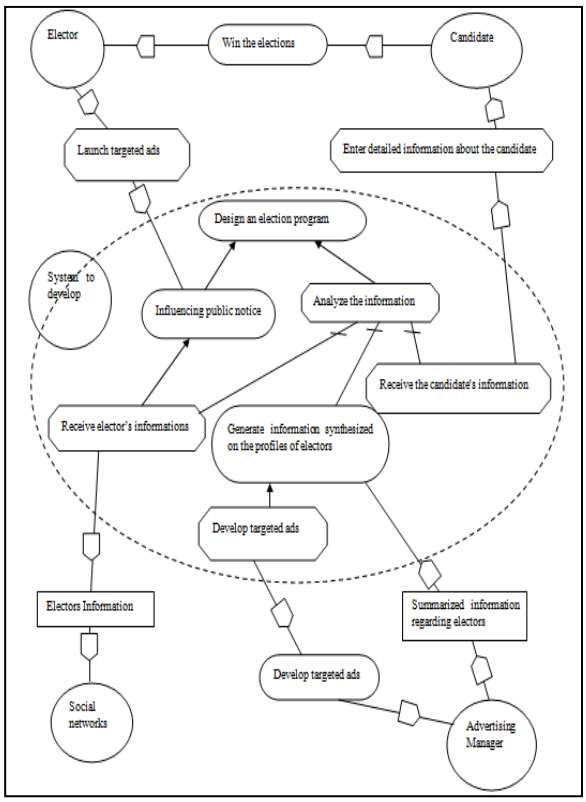


Figure 2 : Strategic rationale model for elections

## 5. BiStar: An extension of iStar for Big data projects

In this session, we present BiStar (Big data iStar) which consists of an extension of the iStar method for Big data projects. We start with clarifying the needs for an extension of iStar to support elicitation of the requirements for Big data projects; then we explain the concepts to add, after that, we perform the BiStar on the case of study of the presidential elections.

### 5.1. The needs for an extension of iStar

In this part, we explain the situation and the important points that we find them as critical ones.
The Elicitation is the most crucial step in RE, if it is not well done can lead to projects that do not respond well to the needs of stakeholders. In the case of Big data projects, it is getting more and more complicated. A Big data project must not only meet a need, but also respond in a very short time by processing a big amount and a specific nature of data (structured, semi-structured, unstructured).
During the RE phase for Big data projects, we are interested in what to collect rather than how to collect because current techniques and approaches of RE are valid for Big data projects. Big data projects are like traditional projects on how to collect requirements. It is on the Big data properties (Volume, Velocity, Variety) that we are going to focus to collect and model them in the RE phase by BiStar.
Also, The papers [OtPe15], [ShOt16] confirmed that there is the necessity for the Big data software to include all the three parameters (functional feature, time constraint, verifiable during some period) to completely define the requirement specification for Big data projects .

### 5.2. The Concepts added to iStar

Based on the needs of requirements for Big data in the literature [ChML14], [KaWG13], [Madd12], [tPe15], [ShOt16], we have chosen to add the concepts of execution time, volume of data to process, variety of data, and durability of a goal. In the rest of this subsection, we will explain each concept and clarify why we added it.

#### 5.2.1. The execution time

In a Big data project, the execution time must be exact. A late result is considered a wrong one.
We take the case study of presidential elections presented in section 3. The stakeholder needs the goal ‘’Generate information synthesized on the profiles of electors‘’, and does not specify in what time it should be performed. The project will well be done and finished. But the goal must be achieved in 15 days. So the project has failed to satisfy the stakeholder’s need. We conclude that the execution time of each goal must be specified at the beginning of the project.

#### 5.2.2. The volume of data to process

The volume of data is one of the most important features of Big data projects, the volume is often large, but stakeholders are not aware of what can be done and what cannot be done. Even using Big data technologies

like (Hadoop and nosql systems), volume remains a crucial point when talking about Zettabytes [KaWG13]. In the case study of presidential elections presented in section 3, the stakeholder needs the goal ‘’Generate information synthesized on the profiles of electors‘’, and does not specify volume of data that must be proceeded. However the goal needs to analyse 100 Zettabytes of data. So the project has failed. The volume of data of each goal must also be specified at the beginning of the project.

### 5.2.3. The variety of data

In Big data projects, we find data with different presentations (structured, semi-structured data, and unstructured data). Building a Big data Project that manipulates semi-structured data is different from unstructured data.
In the above example (see section 3), the stakeholder does not specify the nature of the data that must be proceed. The goal needs to analyze semi-structured and unstructured data. Consequently, the nature of data of each goal must be also specified at the beginning of the project.

### 5.2.4. The durability of a goal

Big data projects are built to meet the needs during specified times; it turns out that their goals may become dissatisfied for stakeholders, so we need to get an agreement from the beginning on the time in which a requirement can be satisfied.
In the case study considered in section3, the stakeholder does not specify the durability of its goal. When we validate the project with the stakeholder, he says it is not what he wants; the goal must be satisfied during the hall election. So the project failed to satisfy the need of the stakeholder. Also, the durability of a goal must be specified at the beginning of the project.
iStar does not support the properties presented above, which do not allow a complete and refined elicitation of the requirements for Big data. We see that to support Big data projects by the iStar method; we must make sure that the goals are attached to their properties (execution time, the volume of data to be processed, the variety of data, and the durability of goal).
Figure [3] shows graphically the concepts added to the Strategic Dependency (SD) Model, and The Strategic Rationale (SR) Model.

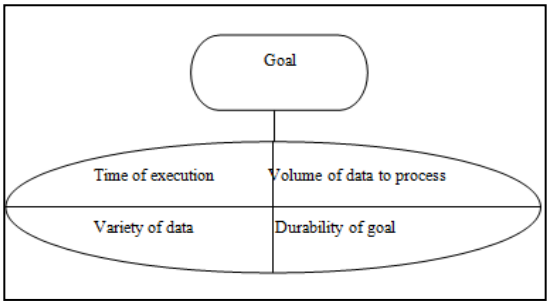


Figure 3 : Concepts added to BiStar

### 5.3. The application of BiStar on the case study

Figure [4] shows the application of BiStar strategic dependency model on the example of presidential elections.

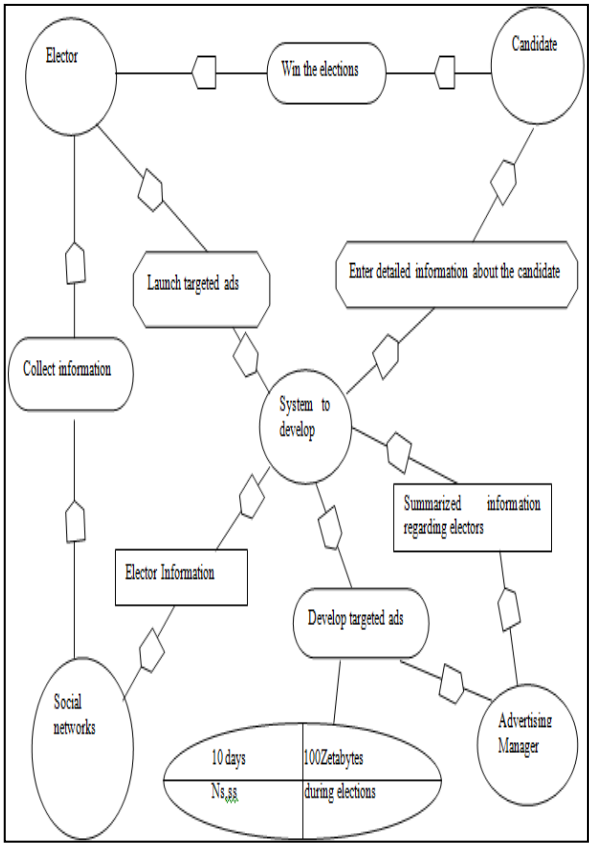


Figure 4 : Strategic Dependency model of BiStar for the elections

We keep the same meaning explained in section 4.1. However, we find in BiStar that new concepts are linked to the goal "Develop targeted advertising" which means that this goal must be done within 10 days, by analyzing 100 Zettabytes of unstructured and semi-structured data nature, and it must be in operation during the elections. Like that we give more completeness and refinement to the requirements.
Figure [5] shows the application of BiStar's strategic rationale model on the example of presidential elections. We keep the same meaning explained in section 4.2, but also, we find in BiStar that new concepts are linked to the goal "Design an election program" and to the goal "Generate synthesized information about the profile of electors". We understand that the goal "Designe an Election Program" must be done within 2 days, by analyzing 30 unstructured and semi-structured nature Petabytes and it must be functional during the elections. And for the goal "Generate information synthesized on the profiles of electors", it must be done within 15 days, by analyzing 100 Zettabytes of unstructured and semi-structured nature, and it must be functional during the elections.

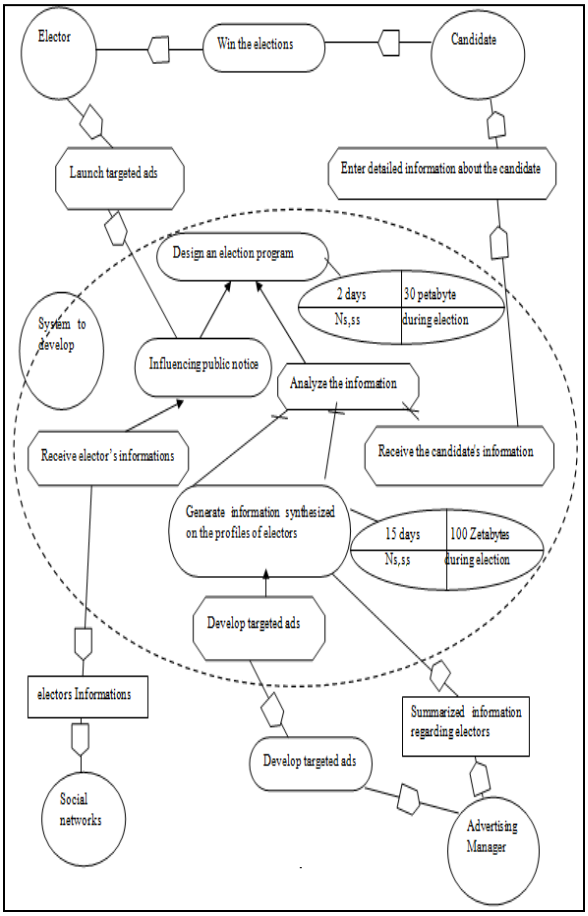


Figure 5: Strategic Rationale model of BiStar for elections

## 6. Conclusion

In this work, we have proposed BiStar (Big data iStar) a new extension of iStar to elicit the requirements for Big data projects. This extension takes into account the properties of Big data projects to ensure a proper elicitation of the requirements.
We applied iStar and BiStar on the same case study of the presidential elections to show the utility of BiStar. We can note that without BiStar we can mess some important requirements. This modest research was the first attempt to feel the gap in the field of the adaptation of RE methods for Big data projects.
We are completing this work by applying the rest of the life cycle activities of RE (specification, validation...). We hope that the research community gives more attention to this field.

## 7. References

[ChML14] CHEN, MIN ; MAO, SHIWEN ; LIU, YUNHAO: Big Data: A Survey. In: *Mobile Networks and Applications* Bd. 19 (2014), Nr. 2, S. 171–209

[ChPr09] CHUNG, LAWRENCE ; DO PRADO LEITE, JULIO CESAR SAMPAIO: On non-functional requirements in software engineering. In: *Conceptual modeling: Foundations and applications* : Springer, 2009, S. 363–379

[DaFH16] DALPIAZ, FABIANO ; FRANCH, XAVIER ; HORKOFF, JENNIFER: istar 2.0 language guide. In: *arXiv preprint arXiv:1605.07767* (2016)

[GCAH18] GONÇALVES, ENYO ; CASTRO, JAELSON ; ARAÚJO, JOÃO ; HEINECK, TIAGO: A Systematic Literature Review of iStar extensions. In: *Journal of Systems and Software* Bd. 137 (2018), S. 1–33

[IeAI97] IEEE COMPUTER SOCIETY ; ACM SIGSOFT ; IFIP WORKING GROUP 2.9 (Hrsg.): *Classification of Research Efforts in Requirements Engineering*. Los Alamitos, Calif : IEEE Computer Society Press, 1997 — ISBN 978-0-8186-7740-3

[I*wi00] *i* Wiki | i* Guide*. URL http://istar.rwth-aachen.de/tiki-index.php?page=i*+Guide. - abgerufen am 2017-12-05

[KaWG13] KATAL, AVITA ; WAZID, MOHAMMAD ; GOUDAR, R. H.: Big data: issues, challenges, tools and good practices. In: *Contemporary Computing (IC3), 2013 Sixth International Conference on* : IEEE, 2013, S. 404–409

[KoSo98] KOTONYA, GERALD ; SOMMERVILLE, IAN: *Requirements engineering: processes and techniques* : Wiley Publishing, 1998

[Madd12] MADDEN, SAM: From databases to big data. In: *IEEE Internet Computing* Bd. 16 (2012), Nr. 3, S. 4–6

[MiNa11] MINA ATTARHA ; NASSER MODIRI: Focusing on the Importance and the Role of Requirement Engineering. In: *Interaction Sciences (ICIS), 2011 4th International Conference on* : IEEE, 2011, S. 181–184

[NuEa00] NUSEIBEH, BASHAR ; EASTERBROOK, STEVE: Requirements engineering: a roadmap. In: *Proceedings of the Conference on the Future of Software Engineering* : ACM, 2000, S. 35–46

[OtPe15] OTERO, CARLOS E. ; PETER, ADRIAN: Research Directions for Engineering Big Data Analytics Software. In: *IEEE Intelligent Systems* Bd. 30 (2015), Nr. 1, S. 13–19

[Ref96] ERIC SIU-KWONG YU.: *Modelling strategic relationships for process reengineering*, University of Toronto, PhD Thesis, 1996

[ShOt16] SHARMA, KAPIL ; OTHERS: Quality issues with big data analytics. In: *Computing for Sustainable Global Development (INDIACom), 2016 3rd International Conference on* : IEEE, 2016, S. 3589–3591

[Vanl01] VAN LAMSWEERDE, AXEL: Goal-oriented requirements engineering: A guided tour. In: *Requirements Engineering, 2001. Proceedings. Fifth IEEE International Symposium on* : IEEE, 2001, S. 249–262

[WeOP09] WERNECK, VERA MARIA BENJAMIM ; OLIVEIRA, ANTONIO DE PADUA ALBUQUERQUE ; DO PRADO LEITE, JULIO CESAR SAMPAIO: Comparing GORE Frameworks: i-star and KAOS. In: *WER*, 2009

[ZoCo05] ZOWGHI, DIDAR ; COULIN, CHAD: Requirements elicitation: A survey of techniques, approaches, and tools. In: *Engineering and managing software requirements* : Springer, 2005, S. 19–46