GRALD: an Approach for Goal and Risk Analysis in the Development of Information Systems for the Web of Data

Marcio Louzada de Freitas, Renata S. S. Guizzardi, Vítor E. Silva Souza

Ontology & Conceptual Modeling Research Group (NEMO)
Federal University of Espírito Santo (UFES), Brazil
Av. Fernando Ferrari, 514 – Goiabeiras – Vitória, ES – 29075-910

marciolfreitas@outlook.com, {rguizzardi, vitorsouza}@inf.ufes.br

Abstract. *The publication of Linked Data on the Web regarding several application domains leads to new problems related to Requirements Engineering, which needs to take into account aspects related to new ways of developing systems and delivering information integrated with the Web of Data. Tasks such as (functional and non-functional) requirements elicitation and ontology-based conceptual modeling can be applied to the development of systems that publish Linked Data, in order to obtain a better shared conceptualization (i.e., a domain ontology) of the published data. The use of vocabularies is an intrinsic activity when publishing or consuming Linked Data and their choice can be supported by the elicited requirements and domain ontology. However, it is important to assess the risk when choosing external vocabularies, as their use can lead to problems. Thus, risk identification, modeling and analysis techniques can be employed, in order to identify risks and their impacts on stakeholder goals. In this work, we propose GRALD: Goals and Risks Analysis for Linked Data, an approach that combines existing Risk Analysis and Web Engineering approaches for modeling goals and risks for information systems for the Web of Data.*

1. Introduction

The Semantic Web was presented by Berners-Lee et al. [2001] as the Web version that seeks to make content understandable by both humans and machines, improve search engines by giving meaning to published content and take into account contextual information of time, space and states of things. According to its creators, a challenge of the Semantic Web is to ensure expressiveness and generate inference over the published content, without losing performance in the representation of the data on the Web. Ontologies are used to integrate different databases, to define the classes, subclasses and relationships between them for the creation of contents for the Semantic Web, making it possible to generate such inferences [Berners-Lee et al., 2001].

At the core of the Semantic Web idea is the concept of *Linked Data*.[1] According to Bizer et al. [2009], Linked Data is a set of data interconnected by URIs (Uniform Resource Identifiers)[2] whose contents can be processed by machines, forming a Web of Data. The published content is based on the RDF (Resource Description Framework) standard[3] and data can be extracted using SPARQL[4] queries. Published data and their

[1] http://linkeddata.org/
[2] https://www.w3.org/wiki/URI
[3] https://www.w3.org/RDF/
[4] https://www.w3.org/TR/rdf-sparql-query/

interconnections are described by means of vocabularies, i.e., schemas that describe the existing entities and the relationships between them. Moreover, such data can refer to several domains, such as Geographic, Media, Social Media, Governmental, Libraries and Education, Life Sciences and so on [Heath and Bizer, 2011].

Given that the Web follows an open and decentralized architecture [Heath and Bizer, 2011], connecting an information system with external data sources can lead to potential risks (e.g., misinterpretation of meanings due to poor documentation, connection timeouts due to infrastructure problems, etc.), thus the need to understand their impact on stakeholder goals. Hence, with the adoption and implementation of Linked Data in several areas of knowledge by companies, institutions and governments, it becomes necessary to analyze goals and requirements, as well as to identify and analyze the risks of adopting Linked Data in the development of Web-based Information Systems (WISs).

Goal-Oriented Requirements Engineering (GORE) approaches aim to analyze and model the goals of systems and stakeholders. Goals can be used to capture interactions and trade-offs between requirements and have been broadly used in Software Engineering, Information Systems Design, Conceptual Modeling and Enterprise Modeling [Horkoff et al., 2016]. GORE approaches, such as the NFR Framework [Mylopoulos et al., 1992], iStar [Yu, 2009] and KAOS [van Lamsweerde and Letier, 2000] could be applied to the modeling of WISs and, in particular, to analyze the use (publication) of Linked Data by such systems. Henceforth, we will refer to WISs that publish their data on the Semantic Web as *Linked Data Systems*.

Some approaches combine goal modeling with risk modeling, which provide tools that help analyze the impact of risks on stakeholder goals. For instance, the GR Framework [Asnar et al., 2011] allows modeling and reasoning about risks during requirements analysis. KAOS allows not only goal modeling but also obstacle analysis. The RISCOSS project [Costal et al., 2015] proposes to integrate risk modeling language RiskML with goal modeling language iStar to analyze risks in the adoption of open source software. With some effort, these approaches can be adapted to the analysis of risks in the development of Linked Data Systems.

This paper proposes *Goal and Risk Analysis for Linked Data* (GRALD), an approach that applies GORE and risk analysis techniques for the development of Linked Data Systems. GRALD is based on the RISCOSS approach [Costal et al., 2015] which seeks to align business goals and risks in the adoption of open source software, modeling risks with the RiskML language. The modeling of goals is done with iStar [Yu, 2009], aiming to understand the social domain to enable requirements engineering, defining social concepts. GRALD is integrated with our previous work, FrameWeb-LD [Celino et al., 2016], a method for building Linked Data Systems.

GRALD is motivated by the growing publication of Linked Data in various domains [Heath and Bizer, 2011], in which goal modeling and risks analysis can be applied. Tasks such as requirements elicitation, creation of a domain ontology, and modeling of system goals can also help in the choice of Linked Data vocabularies. The objective of this paper is to demonstrate the applicability of goal and risk analysis in the development of Linked Data Systems, and to assist in the process of choosing vocabularies.

This paper is an extended version of [de Freitas et al., 2018]. It extends the orig-

inal paper by: (i) presenting the approach in more detail, (ii) extending an existing iStar modeling tool in order to include RiskML concepts and, thus, support GRALD; (iii) introducing a catalog of goals and risks for the development of Linked Data Systems; and (iv) providing more detail regarding the evaluation of the approach.

The remainder of the paper is divided as follows: Section 2 summarizes the baseline of our work; Section 3 presents the GRALD process, exemplifies the artifacts produced by it and introduces the catalog of goals and risks for Linked Data Systems development; Section 4 describes how GRALD was evaluated; Section 5 discusses related works; and Section 6 concludes the paper.

2. Baseline

Goal and Risk Analysis for Linked Data (GRALD) is based on two existing approaches: RISCOSS and FrameWeb-LD. We chose to combine these two approaches because, on the one hand, RISCOSS uses two different languages for modeling goals (iStar) and risks (RiskML), which allows one to study how the same risks may affect different strategies or ecosystems [Costal et al., 2015]. Moreover, RISCOSS extends the goal analysis support in iStar, allowing us to analyze how risks are propagated in the goal graph. On the other hand, Frameweb-LD is focused on the development of Linked Data Systems. Through GRALD, we seek synergy between these two approaches.

With some effort, other approaches related to risks and goals, such as the GR Framework [Asnar et al., 2011] or KAOS [van Lamsweerde and Letier, 2000] obstacle analysis could be adapted to use in GRALD. This is, however, out of scope here.

2.1. The RISCOSS Approach

The RISCOSS project[5] [Costal et al., 2015; López, 2015] came about because of the growing adoption of OSS (*Open Source Software*) by organizations. The occurrence of risks in the adoption of OSS can impact the business goals of the organization.

In RISCOSS, risk management is based on a three-layered strategy to cover the gathering of data [López, 2015]. In layer 1, data about risks is collected from OSS communities, projects and experts that determine the risks drivers; in layer 2, risk indicators are defined and risk models are created; and in layer 3, risk models are linked with the goal models to represent the impact that the possible risk events have on strategic and business goals.

The modeling of risks is done using RiskML [Costal et al., 2015], a language that uses primitive concepts like *Goal* — something of interest for a stakeholder to obtain or maintain; *Event* — the occurrence of something that may undermine the objectives; *Situation* — circumstances where risks are likely to occur; and *Indicators* of risks — existing data measurements approved by experts, which can be simple or composite. Moreover, the impact relationship, between an event E and a goal G, indicates that the occurrence of event E impacts on the satisfaction of G [López and Siena, 2015].

On the other hand, business goals are modeled using iStar [López and Franch, 2014], which seeks to understand social concepts and applies them in systems engineering processes. The central concept is the *actor*, which can be human beings, organizations,

[5] `http://www.riscoss.eu`

hardware, software or a combination thereof. Actors are able to act independently, have autonomy, intention to perform an action and their behavior is not totally controllable. Other concepts such as *tasks*, *resources*, *goal*, *softgoal*, *agent*, *roles*, etc. are part of this approach [Yu, 2009].

iStar proposes two types of diagrams: *Strategic Dependence (SD)* and *Strategic Rationale (SR)*. In the SD diagram, the relationship of dependency is addressed: one actor (the *depender*) can depend on another (the *dependee*) for something (the *dependum*). The types of dependency are *goal dependency*, *softgoal dependency*, *task dependency* and *resource dependency*. In the SR diagram, it is possible to reason about the intentional elements that an actor wants to achieve, as well as to indicate how they can be achieved. iStar can be used in requirements engineering, enterprise engineering, security, privacy and trust modeling, etc. [Yu, 2009].

Once goal and risk models are created, they are integrated by identifying concepts that have the same semantics in both models and following the meta-model shown in Figure 1, which also depicts attributes of, and relations between concepts from goal and risk modeling. Moreover, Costal et al. [2015] describe alignment cases that help developers guide the process of iStar–RiskML model integration.

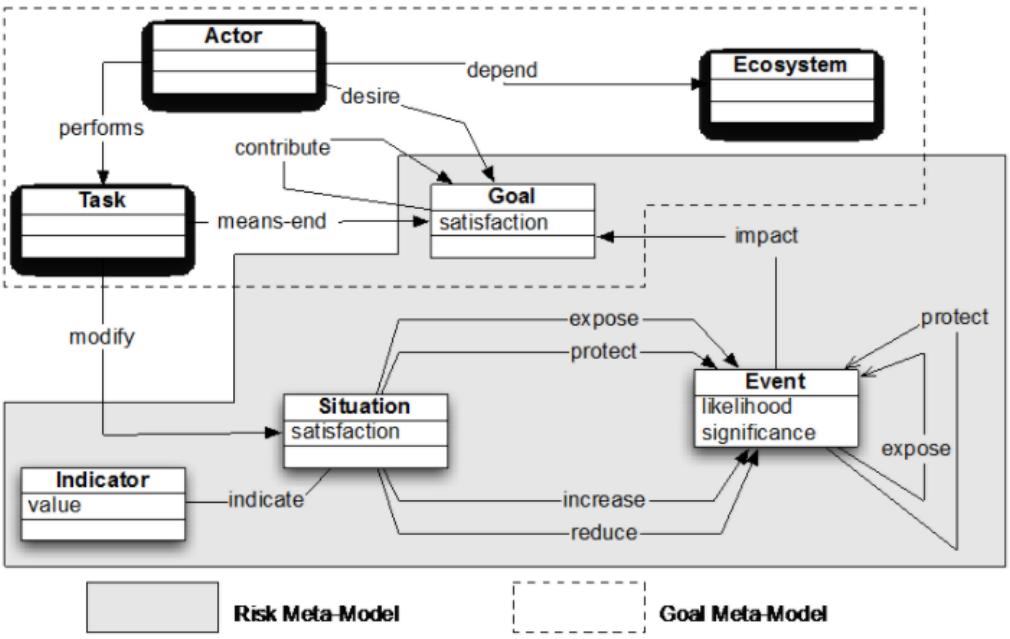

Figure 1. RiskML-iStar Integrated Metamodel [López and Siena, 2015].

2.2. The FrameWeb-LD Approach

FrameWeb-LD [Celino et al., 2016] is an approach for building Linked Data Systems, i.e., Web-based Information Systems (WISs) that publish Linked Data. It proposes a process divided in five stages: *Analysis*, *Design*, *Implementation*, *Testing* and *Deployment*. The main contributions of this approach are an extension of FrameWeb's metamodel [Martins and Souza, 2015] allowing Linked Data mappings to be represented in its design models, and a tool for code generation to assist developers in publishing Linked Data.

In the analysis stage, developers should elicit requirements and develop the domain model. Here, FrameWeb-LD suggests to use first the Ontology Engineering method SABiO [Falbo, 2014] in order to identify the ontology purpose and its intended uses, and then to perform the elicitation of requirements. These requirements can be divided into functional and non-functional requirements. Functional requirements refer to the content to be represented by the ontology and are usually written as competency questions (CQs), i.e., questions that the ontology is supposed to answer. According to Falbo [2014], non-functional requirements refer to features, qualities and general aspects not related to the ontology content. Some examples are usability, maintainability and security.

Next, the ontology is captured and formalized with the aid of tools such as OLED,[6] establishing a shared conceptualization of the domain among domain specialists, in which relevant concepts and relations are identified and organized, guided by the CQs elicited in the previous phase. FrameWeb-LD also suggests OntoUML [Guizzardi, 2005] as the ontology representation language for the domain model.

At the design stage, FrameWeb proposes the creation of an *Entity Model*, based on the conceptual models/ontologies built in the preceding Requirement Engineering phases, in order to represent domain classes and their integration to frameworks that are commonly used in the development of WISs [Martins and Souza, 2015]. FrameWeb-LD adds annotations on top of the basic FrameWeb Entity Model to specify Linked Data vocabulary mappings [Celino et al., 2016].

We illustrate this with a running example that will be used throughout the paper: an academic WIS called Marvin,[7] under development in our university department. In particular, we focus on a module of Marvin called C2D, which keeps track of members of our postgraduate program and their respective publications for evaluation purposes. Researchers and their publications are registered in the system, venues are then matched to a list of qualified conferences and journals provided by the federal government and, based on this list, each publication is assigned a score, which is then used to calculate the score of each researcher.

Figure 2 shows the FrameWeb-LD Entity Model for C2D, in which UML Classes about researchers, publications, etc. are linked to popular vocabularies, such as FOAF[8] and DBLP.[9] For instance, Researcher is equivalent to dblp:Person, given that the scope of the DBLP vocabulary is to represent researchers and their publications. Subclass relations between vocabulary classes and domain classes can be represented by inheritance, e.g., Researcher is subclass of foaf:Person (FOAF has a broader scope and represents not only researchers). The subPropertyOf constraint denotes relations between properties, in this example, the association between Publication and Venue is rdfs:subPropertyOf dblp:publicationType. In the User class, the ld-ignore stereotype represents that user data will not be published in Linked Data.

In [Celino et al., 2016], the implementation phase contains three activities: *Encode Operational Ontology in OWL* (which can be automated by tools), *Encode Web Informa-*

[6] http://nemo.inf.ufes.br/projects/oled/
[7] http://github.com/dwws-ufes/marvin
[8] http://xmlns.com/foaf/spec/
[9] http://dblp.uni-trier.de/

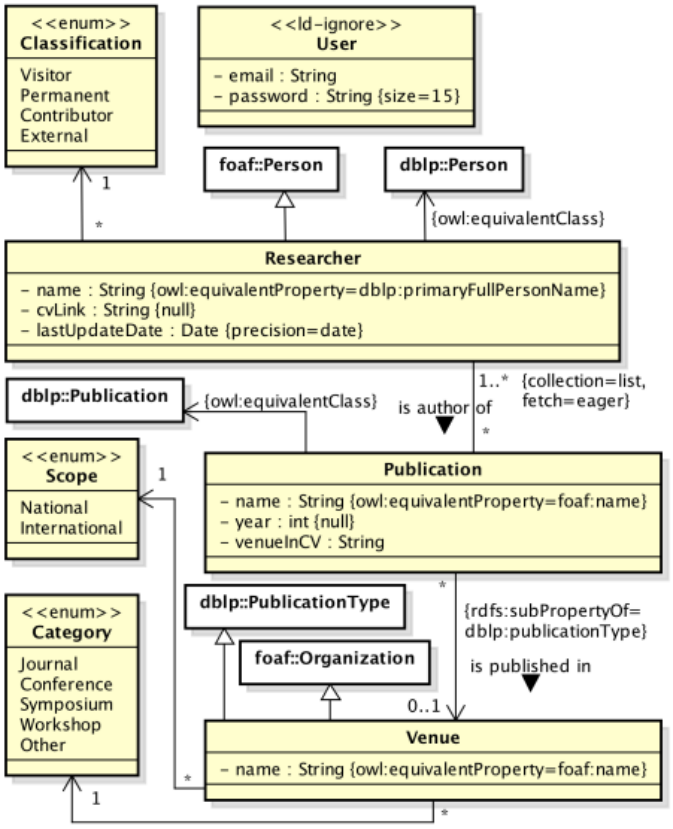

Figure 2. A FrameWeb-LD Entity Model for C2D [Celino et al., 2016].

tion System and *Build Databases*. For the latter, a relational database is created and a Linked Data layer above it is added with the use of D2RQ,[10] which provides triplestore (a database of RDF triples) features such as a SPARQL endpoint. After Implementation, Test and Deployment phases are carried on using traditional Web Engineering techniques.

It is important to note that while FrameWeb-LD allows us to link to external vocabularies, it does not aid developers in finding the most appropriate vocabularies to link. This is very important in the publication of Linked Data, as the objective is to make our data understandable by third party software which has already been programmed to understand some of these popular vocabularies [Heath and Bizer, 2011]. Linking to unknown vocabularies or to terms that do not properly represent your data can compromise this objective. Hence, it is important to properly understand the requirements and risks involved in Linked Data publication.

3. GRALD: Goal and Risk Analysis for Linked Data

In this section, we present *GRALD — Goal and Risk Analysis for Linked Data —*, a method that supports developers of Web-based Information Systems in the analysis of goals and risks in the publication of Linked Data by these systems and in the choice of appropriate vocabularies.

An overview of the development process proposed by GRALD is presented as a

[10] `http://d2rq.org/`

UML Activity Diagram in Figure 3. The process is divided in three stages (the names of the roles defined in each horizontal partition). Rectangles in light background represent activities proposed in FrameWeb-LD (cf. Section 2.2), whereas rectangles in dark background represent activities proposed by *GRALD*, some of them adapted from RISCOSS (cf. Section 2.1). Objects (rectangles with sharp corners, white background) represent artifacts that are produced and/or consumed by the activities of the process. Arrows represent production/consumption of artifacs and, indirectly, establish the sequence of activities, although a specific development life-cycle is not prescribed.

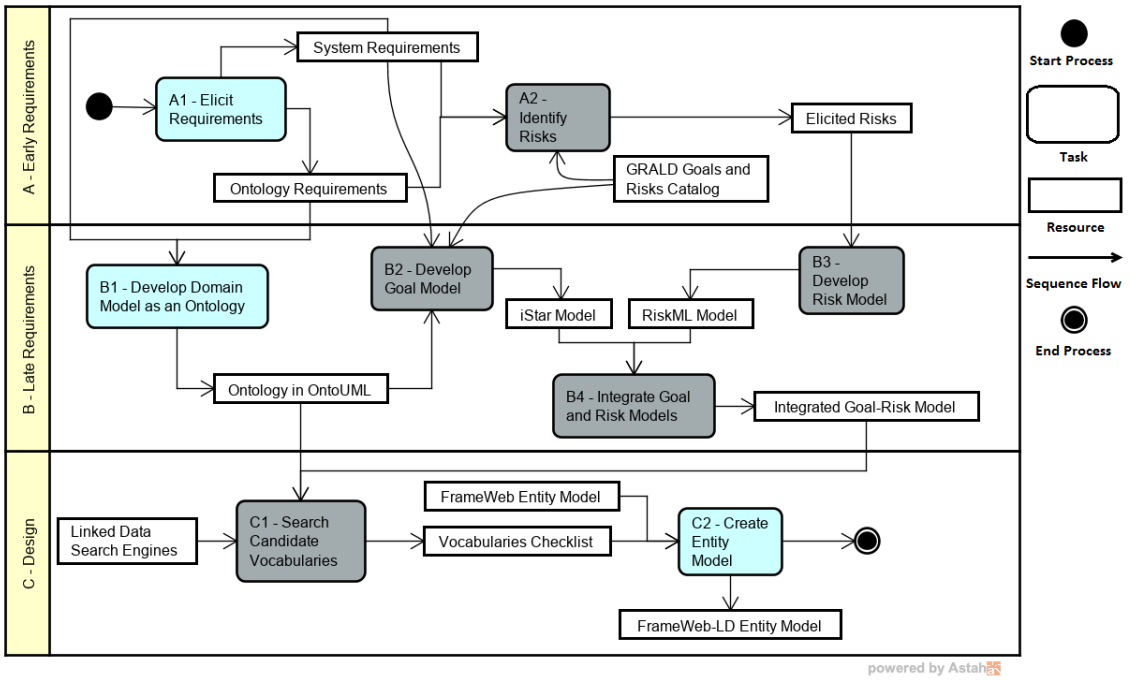

Figure 3. Overview of the GRALD process.

In our proposal we apply these approaches in a unified way, performing goal and risk modeling with iStar and RiskML (RISCOSS), respectively, for the publication of Linked Data with FrameWeb-LD. Thus, we seek synergy between these approaches to aid in the choice of vocabulary to be used by a Linked Data System, understanding the risks involved in the publication and integration of Linked Data. We also provide a catalog of goals and risks related to the development of Linked Data Systems to aid developers in risk/goal identification/modeling activities (as depicted in Figure 3).

In what follows, we detail each phase of the approach, explaining each activity and the artifacts produced at each step of the process. Section 3.1 describes *A - Early Requirements* activities, Section 3.2 presents the *B - Late Requirements* phase, Section 3.3 talks about *C - Design* and, finally, Section 3.4 introduces our catalog of goals and risks for the development of Linked Data Systems.

3.1. A - Early Requirements

In this phase the activities performed are *A1- Elicit Requirements* and *A2 - Identify Risks*. The purpose of the first activity is twofold: elicit requirements for the system to be developed (i.e., its functional and non-functional requirements); and elicit requirements for an

ontology of the domain in question (i.e., the data manipulated by the system). Also, these requirements are used to produce ontology and goal models in the next phase.

Requirements for the WIS should be elicited using any Requirements Engineering technique for early requirements and the iStar language could also be used for this purpose. As for ontology requirements, Celino et al. [2016] suggests the use of techniques prescribed by SABiO [Falbo, 2014] in order to identify the purpose and elicit the requirements for an ontology of the domain in question. Such requirements are then documented in the form of competency questions (CQs).

For example, in [Celino et al., 2016], for the C2D system introduced in Section 2.2, some of the elicited CQs are: "What is a researcher in the post-graduate program?" (CQ1), "What are the possible roles for a researcher?" (CQ2), "What is the scoring system to evaluate researchers in the program?" (CQ3). The answers obtained by these CQs serve as a basis for the creation of the conceptual model in an ontology modeling language in the next phase.

In the *A2 - Identify Risks* activity, risk identification is performed, using traditional Risk Analysis techniques [Bannerman, 2008; Boehm, 1991]. Bibliographic references related to Linked Data are used to support this phase. According to Hyland et al. [2014], best practices for publishing Linked Data should be considered, such as choice of dataset; URI creation; choice and creation of vocabulary; choice of an appropriate license for the publication of content; among others. The W3C [2017] also addresses best practices related to data on the Web. The adoption of these best practices helps prevent risks and, conversely, starting from them, we can identify possible risks related to the publication and consumption of Linked Data in our projects. Further, in [Bruwer and Rudman, 2015] traditional Web risks are extended to the Semantic Web and specific risks of Linked Data and Semantic Web, such as *SPARQL/SPARUL* injections, etc. are also analyzed. Risks related to the creation and maintenance of ontologies and trust and proof of information are also addressed.

We adapt RISCOSS' risk management strategy (cf. Section 2.1) to the case of Linked Data publication, collecting data about risks from the bibliography and Linked Data community websites. For example, in the context of C2D, Table 1 shows situations and risk events, as well as new goals related to data publication. Here, the main goal related to Linked Data is data publication, so these risk events should be taken into account when developing the system. Risks related to other categories (vocabulary adoption, creation and maintenance of ontologies, trust and proof of information, etc.) were also elicited but, for the sake of brevity, are not shown here.

Based on this risk identification activity, the risk models in the RiskML language are created in the next phase, focusing on the impact of risk events on the new identified goals, regarding the use of Linked Data by such systems (e.g., vocabulary adoption, data publication, data provenance, etc.).

3.2. B - Late Requirements

In this phase, the tasks performed are *B1 - Develop Domain Model as an Ontology*, in which FrameWeb-LD prescribes the creation of a conceptual model of the domain elements of the system in OntoUML [Guizzardi, 2005]; *B2 - Develop Goal Model*, in which iStar goal models are created with the objective of identifying and modeling actors, goals,

Table 1. Elicited risks regarding data publication in the context of C2D.

Goal	Event	Situation
Use good quality ("cool") URIs	Not provide URI in accordance to the best practices	URIs in non-compliance with best practices
Access to RDF always available	Inaccessible site	Infrastructure problem
Data updated and accurate	Data not updated or incorrect	Wrong data registration
		Low validation of data
Structured content published	Unstructured content in RDF	Encode web information system implementation error

qualities, tasks, resources and other related elements for Linked Data Systems; *B3 - Develop Risk Model* in which RiskML risk models are created based on previously identified risks and; finally, *B4 - Integrate Goal and Risk Models*, which analyzes the impact relation of risk events on goals, producing an integrated goal-risk model.

The first activity concerns the design of OntoUML models based on elicited system and ontology requirements, as briefly discussed in Section 2.2. The development of a domain model based on OntoUML aims to create a model with greater expressiveness of the domain, to establish a consensus between the experts and to obtain a shared conceptualization of the domain. Since this activity has already been proposed in FrameWeb-LD (thus, not being a novelty of GRALD), we refrain from presenting an example of a developed ontology here. For more detail on the development of the ontology, we refer the readers to [Celino et al., 2016] and [Falbo, 2014].

The purpose of the next activity, *B2 - Develop Goal Model*, is to model the goals (requirements) of the system using the iStar language, with a particular focus on publication of Linked Data. Through goal modeling we can identify actors (stakeholders) and the relationship between them, goals to be achieved, tasks to be performed, resources to be employed, links between elements, etc. The use of an iStar modeling tool is recommended. For our running example, we used the piStar[11] tool [Pimentel and Castro, 2018] in order to create iStar 2.0 models. Figure 4 shows the goal model for C2D.

The actor C2D represents the system itself, deployed and maintained in our university. The central goal for C2D, therefore, is Data Published in Linked Data, divided in subgoals, according to the data that will be published: Scores, Venues, Publications and Researchers. The goal Users not published in Linked Data represents the fact that private user data should not be published. The data is registered in the system by the tasks Calculate researcher score, Manage venues, Manage publications, Manage researchers and Manage and authenticate users.

About the *qualities* of the system, the main goal Data Published in Linked Data helps C2D to Keep transparency because the data on researcher accreditation are open for the community to search; Content structured and processable by machines and Easier access to data are helped because the data is published in RDF format, allowing the possibility of a computational agent to process it. The task Calculate researcher

[11] `http://www.cin.ufpe.br/~jhcp/pistar/`.

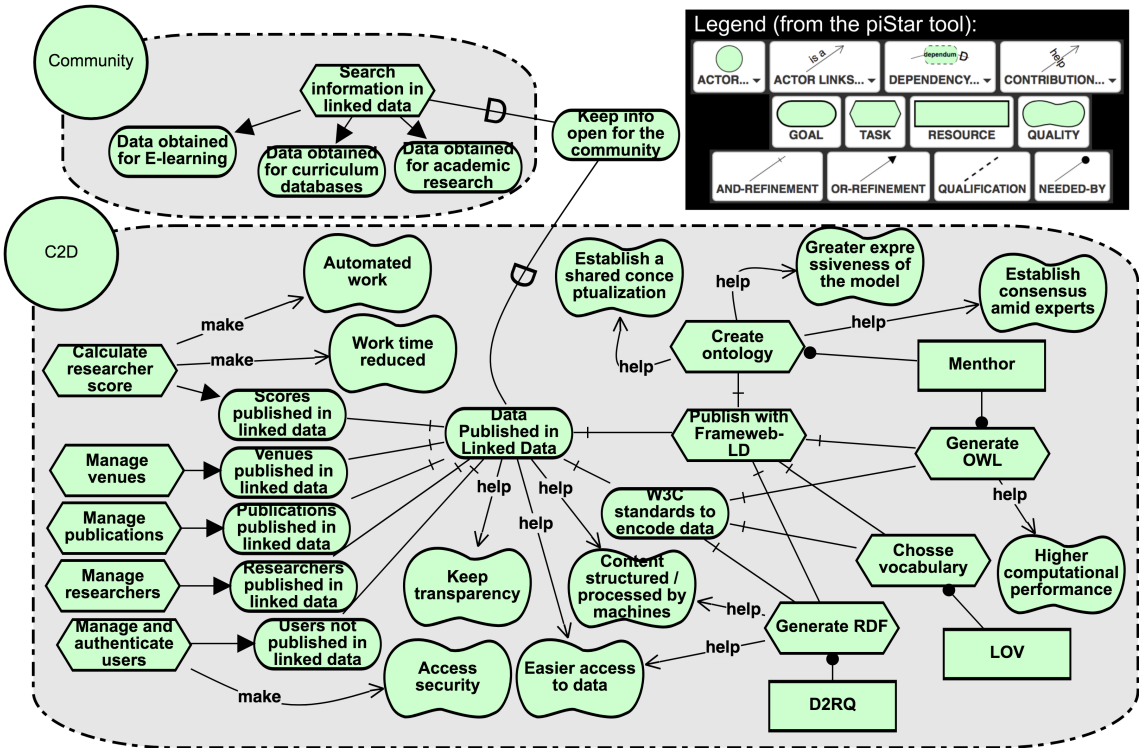

Figure 4. iStar 2.0 goal model for C2D, built with piStar.

score makes the Automated work and Work time reduced sofgoals, eliminating the manual work of calculating the scores, usually conducted by members of the program (i.e., researchers, with a doctoral degree) and considerably reducing the time taken to generate these scores for all researchers of the program. The task Manage and authenticate users makes Access security, ensuring access control to the system. The data is published by FrameWeb-LD (cf. Section 2.2).

The actor Community, in Figure 4, represents the academic community, composed by students, professors (researchers), staff, etc. As such, the community has the goals Data obtained for E-learning, Data obtained for academic research and Data obtained for curriculum databases, accomplished by the task Search information in Linked Data. To do that, the Community depends on C2D to Keep info open for the community.

Next, based on the results of the *A2 - Identify Risks* activity in the *A - Early Requirements* phase (e.g., Table 1), we *B3 - Develop Risk Model* (cf. Figure 3). The situations and events of risks, as well as the (potentially new) goals are modeled in RiskML, with the purpose of demonstrating the impact of the events on the goals.

Figure 5 shows the risk model related to data publication. For instance, the goal Use cool URI is impacted by risk event Not provide URI in accordance to the best practices exposed by risk situation URI in non-compliance with best practices. The goal Structured content published is impacted by the risk event Unstructured content in RDF exposed by the situation Encode web information system implementation error. Other risks related to vocabulary adoption, creation and maintenance of vocabulary and ontology, dataset selection, trust and proof of information and traditional Web risks

are represented in separate models, not shown here.

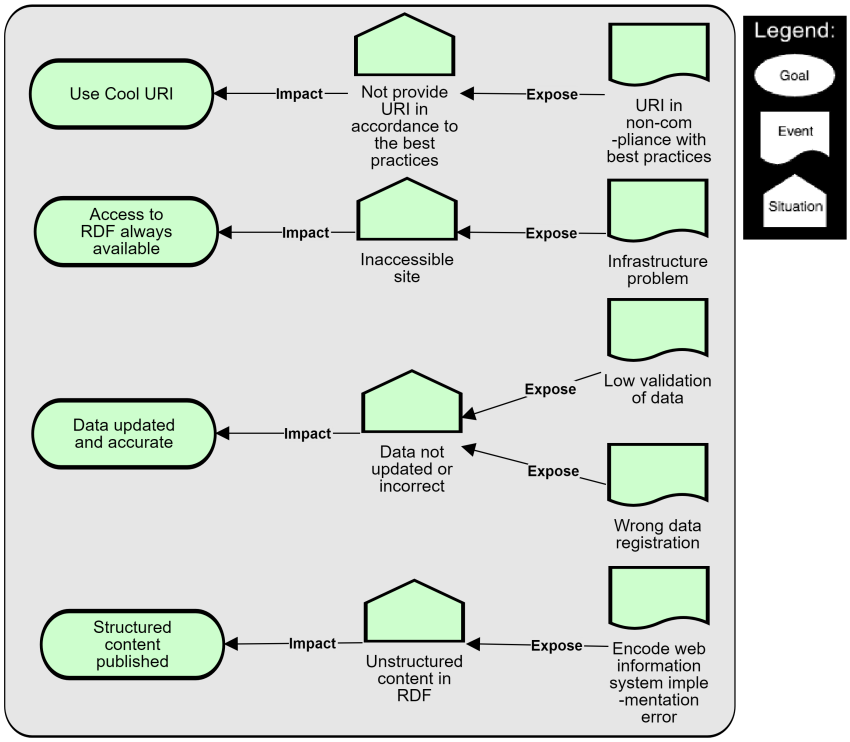

Figure 5. Data publication Risk Model in RiskML language.

Based on RISCOSS, the last activity of this phase is *B4 - Integrate Goal and Risk Models*, aligning goals and risks. To this end, new goals that were elicited during the construction of the RiskML model are added to the iStar model and are associated with existing goal model elements. At this point, elements from both models can be maintained, added or discarded in order to produce an integrated model.

In Figure 6, new goals related to data publication, elicited during risk analysis, are added to the model. The existing goal Data Published in Linked Data (already in iStar goal model of Figure 4) is connected to the new goals: Use cool URI, Access to RDF always available, Data updated and accurate and Structured content published (coming from the Risk Model in Figure 5), are impacted by the risk events Not provide URI in accordance to the best practices, Inaccessible site, Data not updated or incorrect and Unstructured content in RDF, respectively.

Once the models are integrated, risk analysis can be performed as per [Costal et al., 2015] (not detailed here). The impact relation between a risk and a goal represents a negative effect when the event is likely and significant, increasing the evidence that the goal is not achieved. Such evidence is then propagated through the goal graph calculating, for each intentional element, if it is totally/partially satisfied/denied. We are then able to see how risks affect the strategic/high-level goals of each of the involved actors and prioritize our risk mitigation efforts based on this analysis.

In this work, we were particularly concerned about the creation and maintenance of the models. As already discussed, goal models are created in the iStar language, and the risk model using RiskML language. For goal modeling, we use the piStar tool, which

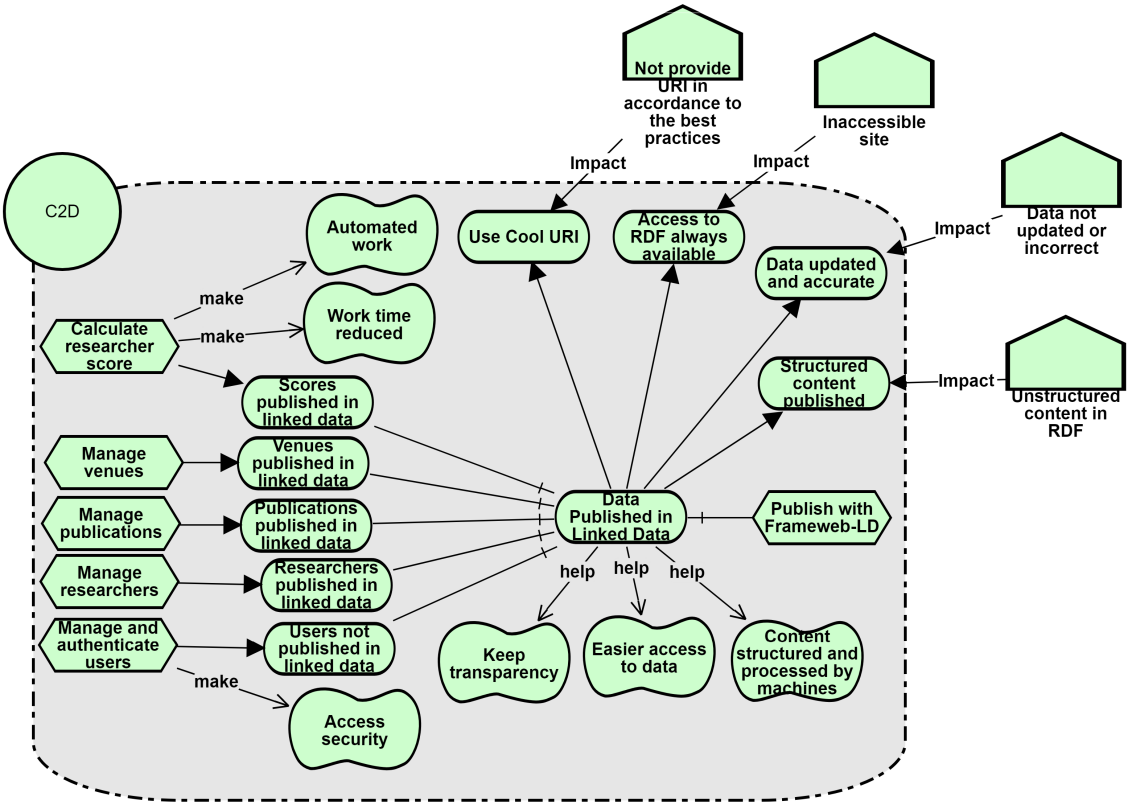

Figure 6. C2D integrated goal-risks model related to data publication.

allows you to draw iStar goal models, serialize them in the JSON format and to export them as images in SVG or PNG formats [Pimentel and Castro, 2018].

For the creation of risk and integrated (risk and goals) models, we extended the piStar tool in order for it to support elements related to the RiksML language (*Situation*, *Event*), as well as its relations (*Increase*, *Reduce*, *Protect*, *Expose* and *Impact*), some of which were already shown in figures 5 and 6. Figure 7 shows a partial screenshot of the extended piStar tool that shows the palette with RiskML elements. The source code of the tool can be found at `https://github.com/nemo-ufes/FrameWeb-GRALD`.

Figure 7. piStar adapted for RiskML modeling.

3.3. C - Design

Based on the results of the previous phase, *C - Design* begins with *C1 - Search Candidate Vocabularies* for Linked Data publication (cf. Figure 3). The models built in the previous phase identify classes and relations to be published as Linked Data and, based on these, we can search for vocabularies. After choosing the vocabularies, the task *C2 - Create Entity model* is performed.

For the first activity, the W3C [2017] suggests Linked Data search engines such as Linked Open Vocabularies (LOV),[12] Watson,[13] Prefix.cc,[14] or Bioportal[15] (for the domain of Biology), for instance. According to them, in the process of choosing a vocabulary we must take into account if the vocabularies are published by a trusted group or organization, if they have permanent URIs, if there are frequent updates published under a version control policy, if they are properly documented, if they are self-descriptive, if they are described in more than one language, if they are used by other data sets, and if they are available for access for a long or infinite time. These recommendations form a checklist developers should go through in order to determine the quality of each candidate vocabulary.

In our running example, we used the search engine Linked Open Vocabularies (LOV). To search for vocabulary classes for the Researcher, Publication and Venue domain classes, we searched LOV for categories (tags) related to the domain. Analyzing results using the aforementioned recommendation checklist resulted in the choice of new vocabularies for C2D (with respect to what had already been chosen in [Celino et al., 2016]), namely *Schema.org*, *DBPedia*, *Bio*, *Bibtex* and *Bibo*. Analyzing links between vocabularies also helped in the discovery of new vocabularies to consider.

The checklist used in this process is shown in Table 2. Vocabulary attributes are presented in different rows, whereas the columns indicate if the vocabularies being checked meet the criteria (represented by a checkmark: ✓), do not meet the criteria (represented by an ×), or partially meet the criteria (represented by a plus/minus sign: ±). To check each attribute, the data presented by LOV was analyzed, as well as the vocabularies' own documentation and their OWL schema.

Table 2. Vocabulary checklist for C2D.

#	Attributes	Dbo	Schema	Bibo	Bio	Bibtex
1	Published by a trusted group or organization	✓	✓	✓	✓	✓
2	Have permanent URIs	✓	✓	✓	✓	✓
3	Version control policy	✓	✓	✓	✓	✓
4	Documented vocabularies	✓	✓	✓	✓	✓
5	Self descriptive vocabularies	✓	✓	✓	✓	✓
6	Described in more than one language	✓	×	×	×	×
7	Used by other data sets	✓	✓	✓	✓	±
8	Available for access for a long/infinite time	✓	✓	✓	✓	✓

Linked Data search engines, such as LOV, provide vocabulary information such as label, URI, namespace, description, creator, publisher, comment and language. Also, information such as vocabulary version history is important to measure the reliability of vocabulary regarding the level of updates that may represent the addition of new classes, properties and deprecated classes. Incoming links represents the popularity of vocabulary because it means that other projects are referencing it. Below, we further describe how each item of the checklist can be verified:

[12] http://lov.okfn.org/dataset/lov/
[13] http://watson.kmi.open.ac.uk/WatsonWUI/
[14] http://prefix.cc/
[15] http://bioportal.bioontology.org/

- Item 1: check if the vocabularies have at least one creator, URI and namespace;
- Item 2: check if the URI is stable;
- Item 3: check if the vocabulary uses any sort of versioning system, e.g., are there previous versions with different numbering?
- Item 4: check if the vocabularies have websites with their respective documentation;
- Item 5: check the vocabulary OWL schema for triples that describe its classes and properties (e.g., comments or labels);
- Item 6: check the vocabulary OWL schema for strings in more than one language (in our example, Dbo was the only vocabulary that met this criterion);
- Item 7: check if the vocabulary has a substantial amount of incoming links (in our example, LOV indicated Bibtex had only a single incoming link, therefore we consider that it partially met this criterion);
- Item 8: check for how long the vocabulary has been maintained and if they are published in a stable domain.

The above checklist is, of course, not exhaustive and could be improved with further vocabularies and/or desired attributes to check, depending on the availability of resources involved in the software development project.

Once the vocabularies are chosen, we move on to *C2 - Create Entity Model*. In this activity, we build a FrameWeb-LD Entity Model as proposed by Celino et al. [2016], by adding Linked Data mapping annotations to the domain model (i.e., the FrameWeb Entity Model, cf. Section 2.2), based on FrameWeb-LD meta-model, to the vocabulary chosen in the previous activity.

Figure 8 represents the model built for C2D, based on the model previously shown in Figure 2 (cf. Section 2.2), with new vocabulary classes added by the process suggested by GRALD, which are filled in dark background. For instance, for the domain class Researcher, vocabularies schema:Person[16] and dbo:Person[17] were added; for Publication, bibo:Article[18] and bibtex:Article were chosen;[19] and for Venue, schema:Organization[20] and bio:Organization[21] were included.

3.4. A Catalog of Goals and Risks for Linked Data Systems Development

Based on the iStar metamodel, the RISCOSS approach and FrameWeb-LD, we suggest a catalog of goals and risks for Linked Data Systems development, including goals, tasks, resources, qualities and risks events that are common in the development of such systems. The objective of this catalog is to provide knowledge that can be useful for developers in the construction of models for Linked Data Systems.

Table 3 shows part of the catalog with goals related to data publication. Each element of the catalog comes with a brief description so developers can evaluate the need for that element in the models of the system being developed. The full catalog is available for the interested reader in `https://github.com/nemo-ufes/FrameWeb-GRALD/wiki`.

[16] http://schema.org/Person
[17] http://dbpedia.org/ontology/Person
[18] http://purl.org/ontology/bibo/
[19] http://zeitkunst.org/bibtex/0.2/bibtex.owl
[20] http://schema.org/Organization
[21] http://vocab.org/bio/

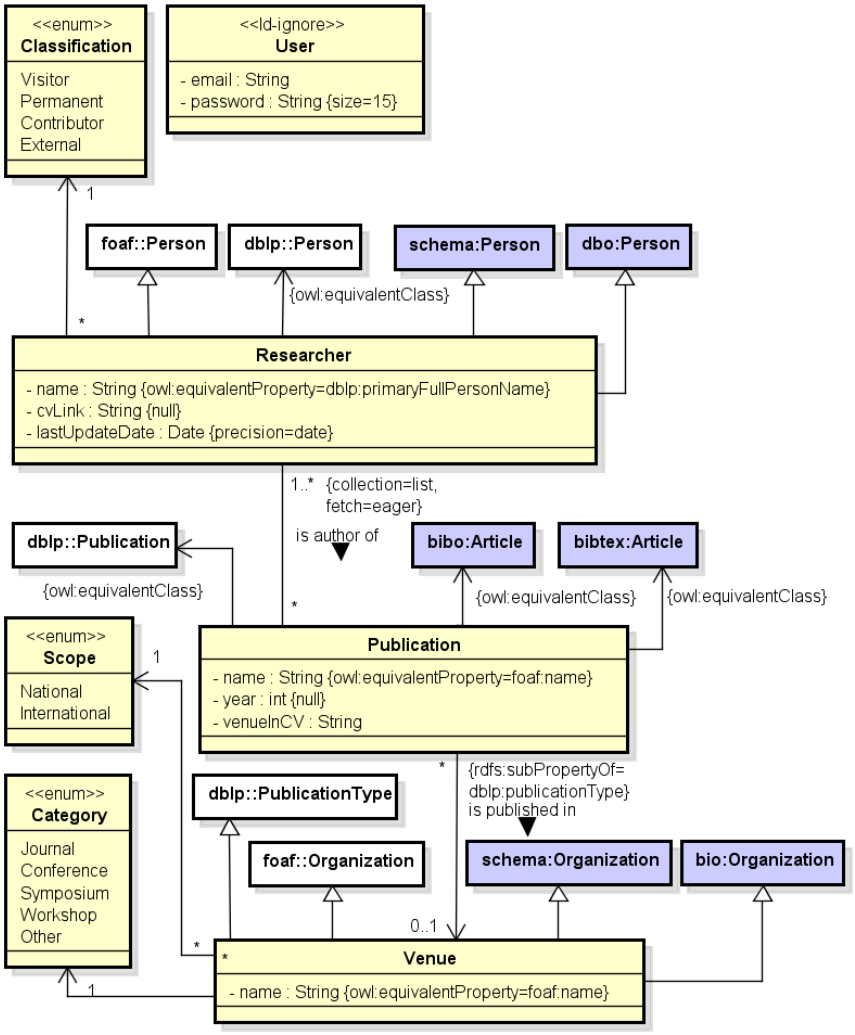

Figure 8. The FrameWeb-LD Entity Model for C2D with newly added vocabularies during the GRALD process.

4. Evaluation

The evaluation of this proposal was conducted by the first author of this paper and an undergraduate Computer Science student [Silva, 2017], using Web-based Information Systems developed by students of the *Web Development & Semantic Web* course of our Postgraduate Program in Computer Science, all of which aim to publish Linked Data. We evaluated our proposal by creating goal and risk models for these systems and searching for vocabularies based on these models. Artifacts are available in a public source code repository: `https://github.com/nemo-ufes/FrameWeb-GRALD`.

During evaluation, we particularly focused on four research questions: **RQ1**: are RISCOSS and FrameWeb-LD integratable in a useful manner? **RQ2**: can GRALD be applied to different systems and domains? **RQ3**: can GRALD be applied to identify risks and new related GORE elements? **RQ4**: can GRALD aid in the identification of vocabularies?

We applied GRALD to five different systems: RightPlace (a system that helps people find a place to live according to their preferences), Rural (management of rural

Table 3. Catalog of goals related to data publication in Linked Data Systems.

Goal	Description
Access to RDF always available	The need to have RDF data always available to be processed by machines.
Data Published in Linked Data	The main goal of the system regarding data publication.
Keep info open for the community	The need to keep system data open to a particular community.
Obtain RDF	The need for a system to obtain data in RDF for the publication of linked data.
Use cool URIs	The need to use adequate URIs for data publication according the *URI Design Principles* and *URI Construction* suggested by the W3C [2017].
Provide details about the data origin	The need to have the origin of the data properly specified.
Ensure the provenance of the data	The need to ensure that provenance data is made available as Linked Data.
Provide credibility and data integrity	The need to provide assurances of credibility and data integrity regarding the published data.

properties), Semed (information system for a medical practice), TransparencyPortal (display government data for citizens) and TravelNM (storefront for a travel agency). By applying GRALD on these existing systems, we were able to identify their goals, tasks, resources and actors, and build their goal models. Moreover, we were able to identify risks related to Linked Data, producing their risk models. These new evaluation efforts complement that of our running example, C2D, already discussed in previous sections.

For reasons of brevity, we present here the results of applying GRALD to the TransparencyPortal system only. This system addresses the issue of open data in the public sector. In Figure 9, the actor Expenses Manager is a system which has the main goal Government expenses published in LD, in order to Keep transparency and achieve Growth of social control. For data publication, we use FrameWeb-LD.

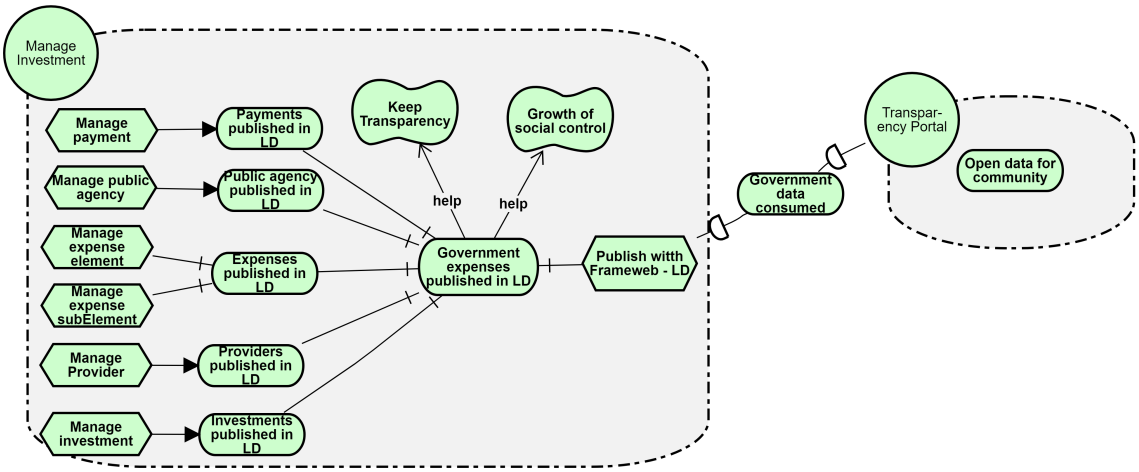

Figure 9. iStar goal model for the Transparency Portal.

Regarding risk modeling, for illustrative purposes, risks related to data provenance will be taken into account here. According to the W3C [2017], the challenge is to publish data and provide details on its origin. Through the provenance of the data, consumers can rely on the integrity and credibility of the data being shared. Based on [W3C, 2017], the Risk Model of Figure 10 was produced.

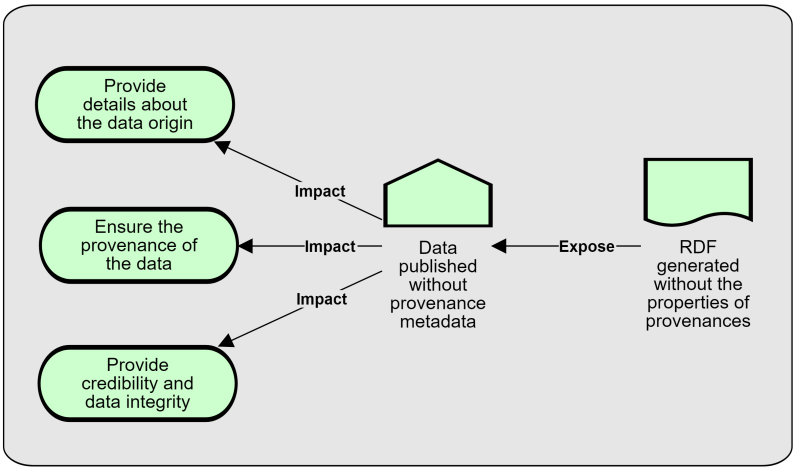

Figure 10. RiskML risk model for the Transparency Portal, related to data provenance.

In this model, goals related to data provenance are: Provide details about the data origin, Ensure the provenance of the data and Provide credibility and data integrity. The situation RDF generated without the properties of provenances exposes the risk event Data published without provenance metadata. According to the W3C [2017], properties such as dct:creator, dct:publisher and dct:issued, in the *Data Catalog Vocabulary* (DCAT)[22] can be used to provide information about the data origin.

Figure 11 shows the integrated goal-risk model for the Transparency Portal. In the figure, the risk event Data published without provenance metadata impacts the new goals Provide details about the data origin, Ensure the provenance of the data and Provide credibility and data integrity, because an RDF generated without the properties of provenance is not in accordance with the best practices, and, in this case, machines will not be able to automatically process information of provenance [W3C, 2017].

Based on previous models, LOV (Linked Open Vocabularies, cf. Section 3.3) was used to discover and analyze Linked Data vocabularies that could be used in the Transparency Portal. Table 4 presents the checklist for these vocabularies, namely, *Schema.org*, *Bio*, *Org*, *Dbo* and *Frapo*.

Finally, Figure 12 shows the FrameWeb-LD Entity Model for Transparency Portal, linking its domain entities with the selected vocabularies. The class Payment was linked to frapo:Payment[23] (*equivalent class*); the class *PublicAgency* to

[22] `https://www.w3.org/TR/vocab-dcat/`
[23] `http://purl.org/cerif/frapo/Payment`

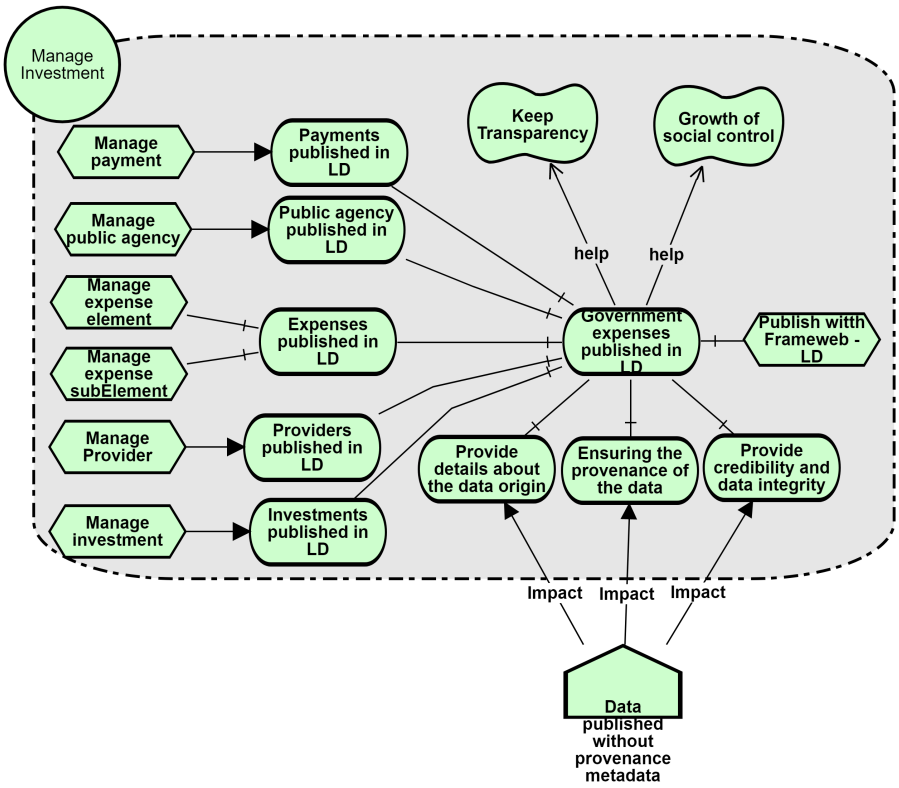

Figure 11. Integrated goal-risk model for the Transparency Portal, related to data provenance.

org:OrganizationalUnit,[24] bio:Organization[25] and schema:Organization[26] (*subclass of*) and, finally, the class *Provider* to org:FormalOrganization[27] and dbo:Person[28] (*equivalent class*).

After applying GRALD to the development of these Linked Data Systems, we analyze the proposed research questions:

1. **RQ1: are RISCOSS and FrameWeb-LD integratable in a useful manner?** We applied GRALD, which integrates RISCOSS and FrameWeb-LD approaches, to six different systems (counting our running example) and in all cases new vocabularies were identified and risks related to their adoption were analyzed, which indicates a positive answer to this RQ.
2. **RQ2: Can GRALD be applied to different systems and domains?** The systems in which GRALD was successfully applied during this evaluation involved many different domains, such as education, geographical, government, medical, etc., which indicates a positive answer to this RQ.
3. **RQ3: Can GRALD be applied to identify risks and new related GORE elements?** Applying GRALD to the aforementioned systems, although very simple and small, allowed us to elicit and model risk elements, then augment the goal

[24] https://www.w3.org/TR/vocab-org/#org:OrganizationalUnit
[25] http://vocab.org/bio/
[26] https://schema.org/Organization
[27] https://www.w3.org/ns/org#FormalOrganization
[28] http://dbpedia.org/ontology/Person

Table 4. Vocabulary checklist for the Transparency Portal.

#	Attributes	Schema	Bio	Org	Dbo	Frapo
1	Published by a trusted group or organization	✓	✓	✓	✓	✓
2	Have permanent URIs	✓	✓	✓	✓	✓
3	Version control policy	✓	✓	✓	✓	✓
4	Documented vocabularies	✓	✓	✓	✓	✓
5	Self descriptive vocabularies	✓	✓	✓	✓	✓
6	Described in more than one language	×	×	✓	✓	×
7	Used by other data sets	✓	✓	✓	✓	±
8	Available for access for a long or infinite time	✓	✓	✓	✓	✓

model with new elements (goals) related to these risks. Further risks could be found with the use of risk identification techniques that are out of scope here.

4. **RQ4: Can GRALD aid in the identification of vocabularies?** GRALD activities *Elicit Requirements*, *Develop Domain Model* and *Develop Goal Model* allowed us to model the classes of the system and clearly specify those that will have the published objects in Linked Data. The checklist used during *Design* aided us in the definition of at least two new (i.e., not previously found by the students) links to external vocabularies per class.

We are aware that the conducted evaluation has some limitations, for instance, having been performed by one of the authors, instead of having had different Web engineers experiment with the methodology and express their opinion regarding its usefulness. Another concern is that the produced models were rather small, and thus we were not able to verify the scalability of the proposed models (an issue which is usually tricky with goal modeling). Moreover, important parts of GRALD, namely the catalog of risks and goals and the vocabulary checklist, were not properly evaluated. To overcome such limitations, new validations are part of our research agenda for the near future.

5. Related Work

There are many works published on Linked Data, but in our case we are particularly interested in publications that involve requirements elicitation, risk identification, risk modeling and goal modeling for the development of systems that publish or consume Linked Data. In our search, we had difficulty to find specific references the to related to the above subjects, which seems to imply that this is an open area of research. In this context, this section refers to proposals on risk/goal modeling for software in general.

In [Giorgini et al., 2005], requirements analysis is performed with the iStar-based Tropos methodology in two phases: *Early Requirements*, which seeks to understand the organizational context where the system can work, and *Late Requirements*, which seeks to define functional (goals) and non-functional (softgoals) requirements for the system-to-be. The authors also propose reasoning with goal models using forward and backward reasoning. In our proposal we have requirements elicitation and risk identification performed in *Early Requirements* and the creation of the models for WIS that use Linked Data in *Late Requirements*.

Kenett et al. [2014] propose capturing, filtering, analyzing and reasoning about risks, based on RISCOSS, using a three layered approach to risk management in FLOSS

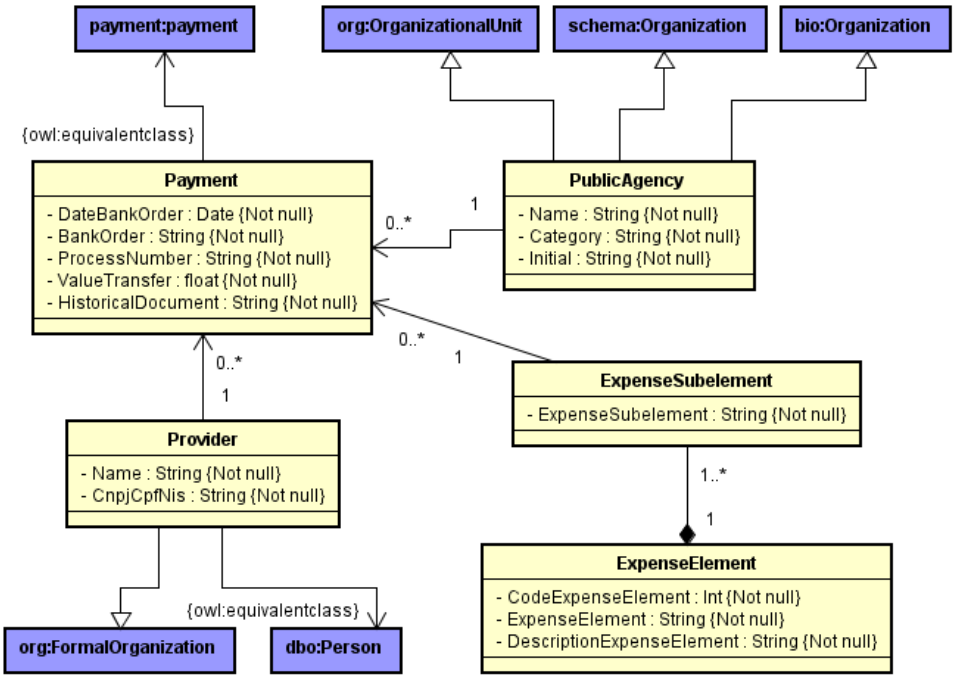

Figure 12. The FrameWeb-LD Entity Model for the Transparency Portal with vocabularies added during the GRALD process.

(*Free Libre Open Source Software*) projects. In the first layer, raw data is collected from FLOSS communities and projects; in the second layer risk indicators are defined and models are produced, in which the risks can be linked to the objectives; finally, in the third layer the risks indicators are converted in Business Risks and, linked with iStar, model business goals to see how risks impact them. In our case we apply a similar process, however with special focus on Linked Data.

In [Westfall and Road, 2001], the software risk management process is performed with steps Risk Identification, Risk Analysis, Plan and Tracking. In our work, we propose risk identification and modeling related to the development of information systems for the Web of Data. Moreover, we combine risk modeling with goal modeling, taking advantage of the benefits of Goal-Oriented Requirements Engineering.

Moreno et al. [2018] propose a requirements engineering framework for Big Data, with special focus on security requirements. Their work aims at providing methodological support to link new data nodes to the existing Big Data cloud. These data sources are usually linked via LOD (Linked Open Data) vocabularies. The authors consider five dimensions when analyzing the data sources: volume, velocity, variety, veracity and value. As in our work, stakeholders goals are considered from the beginning. However, besides specifically focusing on security, their work differs from ours by not considering risks, and by being inspired in agile software development approaches.

Proposals exist that integrate risk analysis activities into Requirements Engineering. For instance, the GR (Goal-Risk) Framework [Asnar et al., 2011] is based in three layers, namely: assets, events, and treatments. The GR Model is defined as a triple

$\{N, R, I\}$ where N is a set of nodes, R is a set of relations among the nodes and I represents a impact relation of a Event wich affecting the asset layer. Impact relations are depicted as dashed line-arrows, the severity of the impact ratio is distinguished in four levels $+$, $++$, $-$, and $--$, where $++$ and $--$ are stronger than $+$ and $-$, respectively. Proposals such as these are generic, whereas our approach is focused on risks of Linked Data publication and integrates with a Web Engineering method (FrameWeb-LD).

6. Conclusions

In this paper, we presented GRALD — Goal and Risk Analysis for Linked Data —, an approach based on RISCOSS, which applies Goal-Oriented Requirements Engineering (GORE) for the development of Linked Data Systems, integrating goal models with risk models in order to perform risk analysis.

GORE is applied in order to help developers to analyze their system objectives, as well as the goals and actors related to the implementation of Linked Data, mapping the necessary resources and tasks to accomplish it. Moreover, performing risk analysis helps to analyze the impact of the occurrence of risk events on system/business goals, as well as to carry out the prevention/mitigation of these risks. Also, GRALD assists developers in the choice of vocabularies based on the tasks performed in the phases of early and late requirements. The search of such vocabularies accomplished using Linked Data search engines following guidelines from a checklist. Finally, the catalog of goals and risks for Linked Data Systems development serves as a knowledge base to aid developers in the elicitation of goals and risks when using GRALD.

Our research proposal is a work in progress and with some limitations, which we intend to address in future work, such as (i) evaluate the proposal with more systems and practitioners, going through goal-oriented modeling and risk analysis; (ii) evaluate the scalability of our models; (iii) improve the catalog of risks and goals for the development of Linked Data Systems; and (iv) develop a tool integrated with Linked Data search engines (e.g., LOV) to assist developers in the task of choosing vocabularies.

References

Asnar, Y., Giorgini, P., and Mylopoulos, J. (2011). Goal-driven risk assessment in requirements engineering. *Requir. Eng.*, 16(2):101–116.

Bannerman, P. L. (2008). Risk and risk management in software projects: A reassessment. *Journal of Systems and Software*, 81(12):2118 – 2133.

Berners-Lee, T., Hendler, J., and Lassila, O. (2001). The Semantic Web. *Scientific American*, 284(5):34–43.

Bizer, C., Heath, T., and Berners-Lee, T. (2009). Linked data-the story so far. *Int. J. Semant. Web Inf. Syst.*, 5(3):1–22.

Boehm, B. W. (1991). Software risk management: principles and practices. *IEEE Software*, 8(1):32–41.

Bruwer, R. and Rudman, R. (2015). Web 3.0: governance, risks and safeguards. *Journal of Applied Business Research*, 31(3):1037.

Celino, D. R., Reis, L. V., Martins, B. F., and Souza, V. E. S. (2016). A Framework-based Approach for the Integration of Web-based Information Systems on the Semantic Web. In *Proc. of the 22nd Brazilian Symposium on Multimedia and the Web*, pages 231–238. ACM.

Costal, D., López, L., Morandini, M., Siena, A., Annosi, M. C., Gross, D., Méndez, L., Franch, X., and Susi, A. (2015). Aligning business goals and risks in oss adoption. In *International Conference on Conceptual Modeling*, pages 35–49. Springer.

de Freitas, M. L., Silva, A. A., Guizzardi, R. S. S., and Souza, V. E. S. (2018). Goal and Risk Analysis in the Development of Information Systems for the Web of Data. In *Proc. of the 21st Ibero-American Conference on Software Engineering (CIbSE 2018), Requirements Engineering track*, pages 473–486, Bogota, Colombia. Curran Associates.

Falbo, R. A. (2014). SABiO: Systematic Approach for Building Ontologies. In Guizzardi, G., Pastor, O., Wand, Y., de Cesare, S., Gailly, F., Lycett, M., and Partridge, C., editors, *Proc. of the 1st Joint Workshop ONTO.COM / ODISE on Ontologies in Conceptual Modeling and Information Systems Engineering*. CEUR.

Giorgini, P., Mylopoulos, J., and Sebastiani, R. (2005). Goal-oriented requirements analysis and reasoning in the tropos methodology. *Eng. Appl. Artif. Intell.*, 18(2):159–171.

Guizzardi, G. (2005). *Ontological Foundations for Structural Conceptual Models*. Phd thesis, University of Twente, The Netherlands.

Heath, T. and Bizer, C. (2011). *Linked Data: Evolving the Web into a Global Data Space*. Synthesis Lectures on the Semantic Web: Theory and Technology. Morgan & Claypool Publishers.

Horkoff, J., Aydemir, F. B., Cardoso, E., Li, T., Maté, A., Paja, E., Salnitri, M., Mylopoulos, J., and Giorgini, P. (2016). Goal-oriented requirements engineering: A systematic literature map. In *2016 IEEE 24th International Requirements Engineering Conference (RE)*, pages 106–115.

Hyland, B., Atemezing, G., and Villazón-Terrazas, B. (2014). Best Practices for Publishing Linked Data, `https://www.w3.org/TR/ld-bp/` (last access: May 26th, 2017).

Kenett, R. S., Franch, X., Susi, A., and Galanis, N. (2014). Adoption of free libre open source software (floss): A risk management perspective. In *2014 IEEE 38th Annual Computer Software and Applications Conference*, pages 171–180.

López, L. (2015). Welcome to the RISCOSS Risk Modeling repository, `https://github.com/RISCOSS/riscoss-risk-modeling/wiki` (last access: May 26th, 2017).

López, L. and Franch, X. (2014). Applying business strategy models in organizations. In *Proc. of the 7th International i* Workshop*. CEUR.

López, L. and Siena, A. (2015). How to create a Risk Model, `https://github.com/RISCOSS/riscoss-risk-modeling/wiki/How-to-create-a-Risk-Model` (last access: May 25th, 2017).

Martins, B. F. and Souza, V. E. S. (2015). A Model-Driven Approach for the Design of Web Information Systems based on Frameworks. In *Proc. of the 21st Brazilian Symposium on Multimedia and the Web*, pages 41–48. ACM.

Moreno, J., Serrano, M., and Fernandez-Medina, E. (2018). Modelado de Requisitos de Seguridad para Big Data. In *Proc. of the 21st Ibero-American Conference on Software Engineering (CIbSE 2018), Requirements Engineering track*, pages 515–522, Bogota, Colombia. Curran Associates.

Mylopoulos, J., Chung, L., and Nixon, B. (1992). Representing and using nonfunctional requirements: a process-oriented approach. *IEEE Transactions on Software Engineering*, 18(6):483–497.

Pimentel, J. and Castro, J. (2018). piStar Tool – A Pluggable Online Tool for Goal Modeling. In *Proc. of the IEEE 26th International Requirements Engineering Conference (RE 2018)*, pages 498–499, Banff, AB, Canada. IEEE.

Silva, A. A. (2017). C2D - Módulo de Credenciamento e Classificação de Docentes do Sistema Marvin. Undergraduate Project, Federal University of Espírito Santo.

van Lamsweerde, A. and Letier, E. (2000). Handling obstacles in goal-oriented requirements engineering. *IEEE Trans. Softw. Eng.*, 26(10):978–1005.

W3C (2017). Data on the Web Best Practices, `https://www.w3.org/TR/dwbp/` (last access: Jul 25th, 2018).

Westfall, L. and Road, C. (2001). Software Risk Management. *Risk Manag.*, pages 1–8.

Yu, E. S. K. (2009). Social Modeling and i*. In Borgida, A., Chaudhri, V., Giorgini, P., and Yu, E., editors, *Conceptual Modeling: Foundations and Applications*, chapter 7, pages 99–121. Springer.