RSLINGO4PRIVACY: AN INTEGRATED APPROACH TO IMPROVE THE SPECIFICATION AND ANALYSIS OF PRIVACY POLICIES

Alberto Rodrigues Da Silva
University of Lisbon, INESC-ID, Instituto Superior Técnico, Portugal

Abstract: *Popular web and mobile applications attract and manage a huge number of users. They collect data from their users without ensuring traceability between privacy policies and application design decisions. A particular challenge for policy authors and application developers is the need to use a common language and companion tools that supports translating important privacy policy statements into actionable requirements. For example, European Union and United States employ privacy policies as "notices" to end users and, in the U.S., these policies are often the sole means to enforce accountability. Given the pressure to post privacy policies and the pressure to keep policies honest, companies must do more to align their policies and practices. In this respect, more should be accomplished by enabling developers with new tools to better specify their data needs while policy authors, who are typically legal professionals, can work with those specifications to create more accurate policies or to enforce those policies in the context of developer data needs.*

In general, a privacy policy is a technical document that states multiple privacy-related requirements that a system should satisfy. These requirements are usually defined as ad-hoc natural language (NL) statements. Natural language is an ideal medium to express these policies because it is flexible, universal, and humans are proficient at using NL to communicate. Moreover, natural language has minimal adoption resistance as a requirements documentation technique (Ferreira & Silva, 2012) (Ferreira & Silva, 2013). However, NL has intrinsic characteristics that become the root cause of quality problems, such as incorrectness, inconsistency or incompleteness (Ferreira & Silva, 2012) (Pohl, 2010).

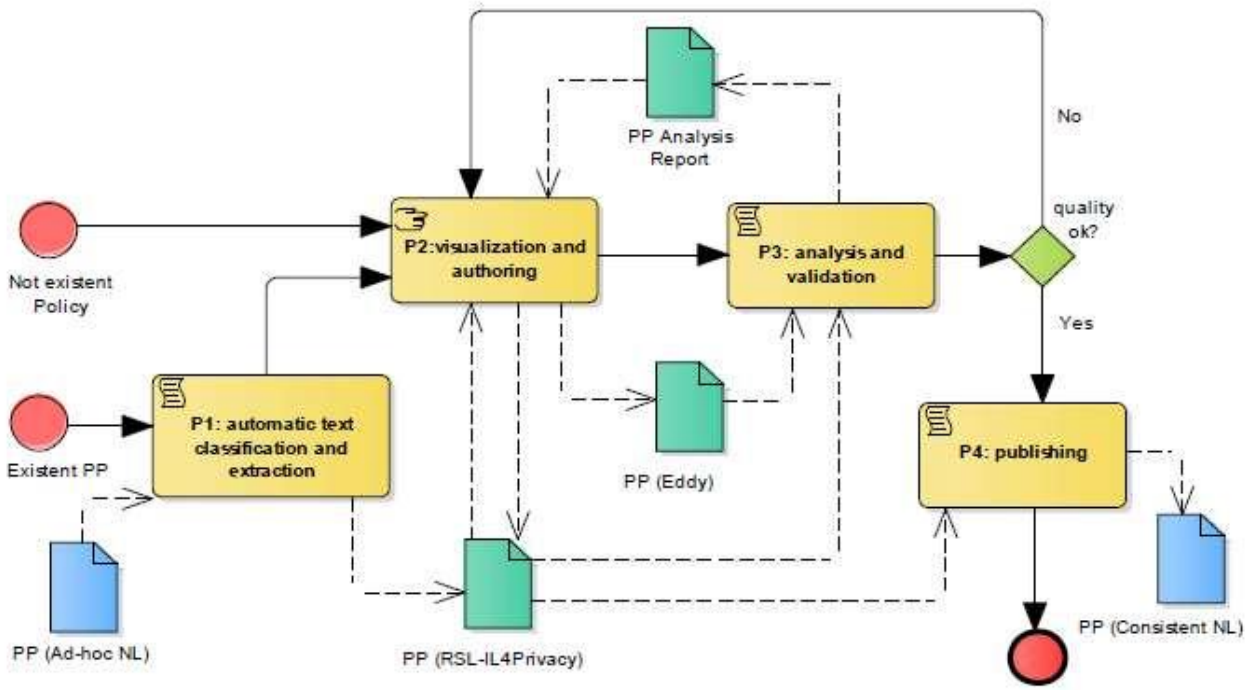

Figure 1: RSLingo4Privacy approach (defined with a BPMN business process diagram).

Recently we proposed the definition of a domain-specific language (DSL) for the specification of privacy-aware requirements, called RSL-IL4Privacy (Caramujo & Silva, 2015) . The RSL-IL4Privacy allows specify privacy policies by providing several constructs, such as statements, private data, recipients and

enforcement mechanisms, which are necessary to specify and document privacy-related requirements. The goal of the proposed approach is to use the RSL-IL4Privacy formalization as the necessary mechanism for the specification of policies while providing features for better analyzing and validating the corresponding privacy requirements.

RSLingo4Privacy is a multi-language approach that uses the following privacy-aware languages: RSL-IL4Privacy and Eddy. Figure 1 overviews RSLingo4Privacy approach as a top-level BPMN business process diagram. If a given (ad-hoc natural language) policy exists, the process P1 applies complex text classification and text extraction techniques to automatically produce the equivalent specification in RSL-IL4Privacy. In addition or otherwise, if that policy does not exist, the RSLingo4Privacy approach starts directly with process P2 to allow visualizing and authoring the policy in a rigorous and consistent way based on the RSL-IL4Privacy language. Process P3 takes as input both RSL-IL4Privacy and Eddy specifications, and provides analysis and validation features, producing, for example an analysis report with errors and warnings that can be taken into consideration during these authoring and validation processes. Finally, when the quality of the policy specified in RSL-IL4Privacy is appropriated, the process P4 is responsible for producing an improved version of the policy, specified again in natural language but in a more consistent and high-quality manner.

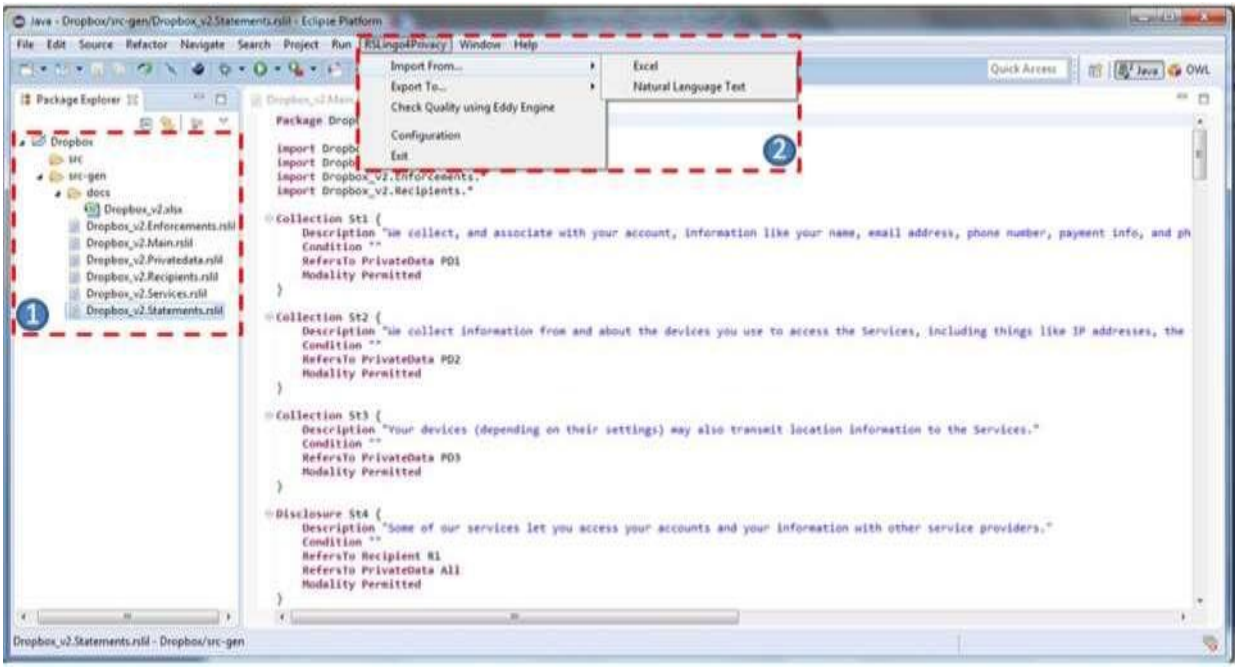

Figure 2. Structure of RSLingo project (1) and RSLingo4Privacy Main Menu bar (2).

This talk presents both the RSLingo4Privacy approach (Figure 1) and its companion tool, the RSLingo4Privacy-Studio,(Figure 2) which is particularly designed for better supporting the specification, analysis and documentation of privacy-aware requirements in the scope of privacy policies.

This work complements the current state-of-the-art by providing a versatile tool designed around the RSL-IL4Privacy domain specific language, with multiple representations while taking into account the importance of having requirements documented in a format as close to natural language as possible. This tool is built on top of the Eclipse IDE, and particularly leveraging and integrating technologies such as: Xtext, Xtend, Eclipse Modeling Framework (EMF), RapidMiner, Eddy engine and Apache POI library.

The explanation and validation with several case studies shows the potential of RSL-IL4Privacy as a rigorous language for expressing privacy requirements and, in addition, shows the relevance of the provided interoperability features. These features are classified by different classes of transformations, all of them founded on that common and intermediate format: RSL-IL4Privacy (defined with the respective Xtext grammar). First, T2M transformations intend to automatically classify NL statements and extract from them text snippets using text mining and text extraction algorithms. The implementation of such transformations is a complex task that involves the integration and tuning of tools like RapidMiner, and is still a working in progress research. Second, M2T transformations produce a consistent and easy-to-read version of a privacy policy. These versions can be produced in multiple formats, such as structured NL in Word, plain text or even HTML. Third, M2M transformations may include two variants: M2M transformations that support multiple representations of the RSL-IL4Privacy; for example, from plain text format (defined with Xtext) into tabular

format in Excel, and vice-versa; and finally, M2M transformations between RSL-IL4Privacy with other languages and formats, such as JSON or Eddy (which can be itself latter mapped in OWL or equivalent formats).

The major merit of the proposed approach is that it allows both technical and non-technical users to easily author and analyze policies using a language close to NL, but that is simultaneously readable by machines and so providing automatic validation at both syntactic and semantic levels. This fact permits RSL-IL4Privacy to act as an intermediate language and be supported by an environment that integrates multiple representations of a privacy policy addressing concerns of multiple stakeholders.

The concrete artifacts of the RSL-IL4Privacy representations for Dropbox, Facebook, LinkedIn and Twitter privacy policies, as well as the analysis of other case studies under the scope of RSLingo4Privacy are available and can be found on its GitHub repository (https://github.com/RSLingo/RSLingo4Privacy).

REFERENCES

Caramujo, J. & Silva, A.R. (2015). Analyzing Privacy Policies based on a Privacy-Aware Profile: the Facebook and LinkedIn case studies. *In Proc. of the 17th CBI conference*. IEEE, 1, 77-84.

Ferreira, D. & Silva, A.R. (2012). RSLingo: An Information Extraction Approach toward Formal Requirements Specifications. *In Proc. of the 2nd MoDRE workshop*. IEEE, 39-48.

Ferreira, D. & Silva, A.R. (2013). RSL-IL: An Interlingua for Formally Documenting Requirements. *In Proc. of the 3rd MoDRE workshop*. IEEE, 40-49.

Pohl, K. (2010). Requirements Engineering: Fundamentals, Principles, and Techniques. Springer.

Silva, A.R. et al. (2016). Improving the Specification and Analysis of Privacy Policies: The RSLingo4Privacy Approach. *In Proc. of the 8th ICEIS conference*. SCITEPRESS.