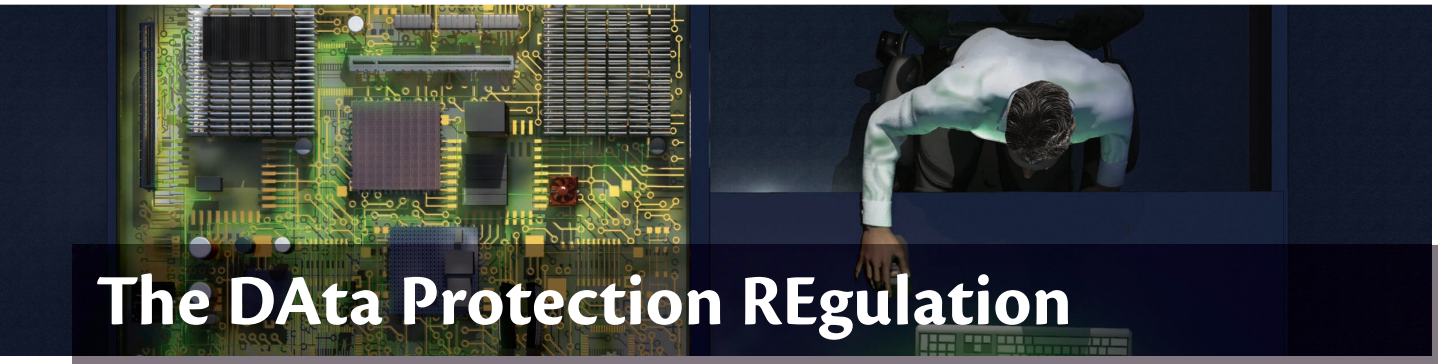

The DAta Protection REgulation COmpliance Model

Cesare Bartolini, Gabriele Lenzini, and Livio Robaldo | University of Luxembourg

Understanding whether certain technical measures comply with the General Data Protection Regulation's (GDPR's) principles is complex legal work. This article describes a model of the GDPR that allows for a semiautomatic processing of legal text and the leveraging of state-of-the-art legal informatics approaches, which are useful for legal reasoning, software design, information retrieval, or compliance checking.

Technological changes in recent decades have brought forth a shift in the market, where personal data have become a significant business asset and a major (sometimes only) source of revenue for many undertakings. Seconding this shift, but acknowledging the need to balance market needs with fundamental rights as laid out at the international level, the General Data Protection Regulation (GDPR) introduced significant changes in data-protection legislation, thus imposing changes on its processing activities and business models. A stronger protection of the rights of the data subject, more clearly defined duties of the controller and the processor, and brand-new rights, such as the right to data portability, demand a new approach in the design and development of personal data-processing activities. Although previous data-protection laws lacked an effective enforcement, the GDPR now demands that the protection of personal data not be taken lightly, extending the enforcement powers of authorities and introducing severe penalties. GDPR compliance has become a serious issue for companies in the hopes of avoiding the high fines and other sanctions imposed by Articles 82 and 83 of the regulation.[16]

Effects of the GDPR are not limited to data-processing activities, but reach upstream to the software tools used in the processing of personal data. According to data protection by design, the novel principle from Article 25, such software tools must have data protection as a driving requisite. Akin to when IT security became a critical concern several decades ago and software development responded by creating models and tools to support security, today, the European Union (EU) offers a similar approach to the protection of personal data.

In general, compliance with regulatory instruments can be problematic, considering that regulations prefer to define quality objectives as those that attain, rather than suggest, technical instruments of compliance. The GDPR fits this description exactly, and it could not be otherwise, as it aims to provide an adequate protection of personal data that stands the tests of time and technical evolution. Imposing specific techniques or tools on it would certainly produce its obsolescence in a few years. Consequently, data protection entails legal compliance issues that are very difficult to address, particularly because the GDPR sets objectives that must be attained, which are expressed as hard-to-interpret requirements.

Digital Object Identifier 10.1109/MSEC.2019.2937756
Date of current version: 20 September 2019

As a result, a new market segment has emerged: one that addresses the need to provide support and tools that improve or verify compliance. In the same way, research institutes and academia have begun considering how to ensure compliance, thus giving the provisions a meaning that is usable by nonexperts in law. Hence the need to define the legal requirements and measures of compliance for data-processing systems that 1) offer models that build data-protection-compliant tools and workflows, 2) build tools and methodologies that can detect potential violations, and 3) propose solutions to fill the void.

Incorporating the requirements expressed by the law is an activity that requires the assistance of legal experts, but modern digital technologies with the proper legal information discipline as well as the Semantic Web can partly automate this task; however, this requires the construction of a machine-readable model to represent legal texts. Such a model should be able to represent different interpretations depending on the context, because legal texts from different contexts can differ greatly in structure, objectives, and meaning of the terminology used. The context is influenced by a number of factors, such as the normative source (e.g., state law, recommendation, or judicial decision), the emitting authority (e.g., state institutions, a public agency, or a doctrinal author), the application domain (contract law, criminal law, competition law, and so on), and many others.

Data protection by design as well as compliance checking are two of the main contexts that would benefit greatly from a machine-processable model of the GDPR, one that is capable of hosting both its provisions and their interpretations. This article presents such a model, called the *DAta Protection REgulation Compliance* (*DAPRECO*) model, which is the main outcome of the DAPRECO project at the University of Luxembourg (see the "Application Scenarios" section).

The DAPRECO model cannot comprise a mere representation of the legal text; rather, textual analysis operates at a naïve level and cannot perform more complex operations such as legal reasoning or decision making. Thus, whereas the model requires the legal text as an essential component, it must also be imbued with the appropriate legal knowledge, capable of giving meaning to the terms and provisions contained therein. The DAPRECO model encompasses three distinct, interoperating components:

1. a structured representation of the legal text
2. a semantic description of data-protection concepts
3. a formalization of the legal provisions.

This model is general purpose and can be used in several different ways; for example, to find correlations between the GDPR and security standards; in information retrieval; to design GDPR-compliant business processes; in decision making; or in risk assessment. This article describes the DAPRECO model in detail and suggests some potential uses, based on previous research.

Application Scenarios

The DAPRECO model was originally designed as a part of the DAPRECO project, whose objectives originate from the topic of legal compliance with the GDPR. However, the DAPRECO model is not a compliance tool per se.

One solution toward compliance is to use technical standards and certification mechanisms. These are normative documents that (generally) express more fine-grained technical requirements than do laws. Because they are managed by technical committees and do not need to undergo intricate legislative processes, technical standards are more easily and frequently updated. In addition to that, technical standards can be certified by appointed authorities; so, at least with respect to compliance with a given standard, there is no need for judicial litigation, although audits and certifications can still be expensive.

Considering that standards and certifications also exist with the purpose of complying with business laws and often refer to specific legal sources in their introductory notes, they are rarely officially endorsed by the law itself. Implementing standards endorsed by the law, i.e., harmonized standards, also gives a legal presumption of compliance; however, harmonized standards are a fortunate but uncommon case. More often, standards do not have such a direct effect on legal compliance. The adoption and certification of a standard can demonstrate a proactive attitude and best efforts to be compliant according to the state of the art in a domain, thus lightening the accountability of an undertaking in case of a lawsuit. In short, standards can provide the undertaking with an argument of compliance.

The validity of such an argument of compliance, however, depends upon a clear correspondence between the provisions of a standard and the law's requirements. Identifying such correspondences is not simple. DAPRECO tries to ease, and partly automate, the process of finding correlations between two normative texts, to understand if, and to what extent, a standard or certification mechanism can be utilized to assert compliance with a legally binding regulation. In other words, a tool using the DAPRECO model would allow, say, a controller processing personal data to know what provisions of the GDPR are covered by a given standard, and, conversely, which provisions still require verification because the standard has no correlation with them.

Scenario 1: Finding Correlations Between the GDPR and ISO 27018

The project applies this methodology to extract correlations between the GDPR and the ISO/IEC

27018-2014 (i.e., ISO 27018) standard on public clouds acting as personal data processors. To identify possible correlations, the methodology uses textual and semantic similarities. The idea, therefore, is to assess whether two legal provisions, expressed in two different texts, can possibly carry similar or identical meanings. For this to happen, two requirements are needed:

1. The expressions used in the two provisions must use similar wordings or synonyms.
2. The textual concepts used in the legal provisions, even if expressed using different terminology, must represent the same semantic concepts.

The first requirement is fairly straightforward, as text similarity natural language processing (NLP) techniques can be used. It is also possible to start with a manual analysis, tagging similar expressions in the two legal texts, and, utilizing appropriate NLP tools, use the tagged expressions to instruct the tool so that further similarities can be automatically detected. The second requirement can easily be satisfied when definitions are available (such is the case in the GDPR and in ISO 27018), so that matching definitions can correspond to equivalent concepts. For example, the definition of "personally identifiable information (PII)" in ISO 27018 (Article 3.2) is nearly identical to that of "personal data" in the GDPR [Article 4(a)], so it is safe to assume that the PII in the ISO 27018 conceptual model corresponds to the concept of personal data in privacy ontology (PrOnto) (see Palmirani et al.[1] for more concrete application examples).

With a thorough implementation, the maximum degree of simplification that this methodology can attain is a checklist of what a controller needs to do, with the exception of all the legal requirements previously addressed by the standards (and formally certified). Although the methodology does not guarantee legal compliance, as that can only be asserted by courts and appointed authorities, it can greatly assist a legal expert in addressing legal requirements.

The DAPRECO model (described in detail in the "Building the Model" section) of both the GDPR and ISO 27018 have been built. Consequently, the legal text in Akoma Ntoso, the ontology, and the logic formulas exist for the two sources. Nonetheless, this article is focused on the model of the GDPR, hence, the ISO 27018 model will not be examined in detail.

Scenario 2: Interoperability Between Legal Documents Beyond the DAPRECO Model

Although the previous section represents the original idea behind the DAPRECO model, during the course of the DAPRECO project, it soon became apparent that the model could have a more general-purpose scope and that its high-level structure could be a potential candidate for machine-readable representations of legal documents. The DAPRECO model is entirely based on technologies developed for the Semantic Web. Leveraging standard formats and languages and with interoperability and extensibility at the core of its design, it can be used to facilitate navigation and search in legal applications. Tools powered by the DAPRECO model, and possibly also by AI technologies, can improve the work efficiency of legal experts, especially in information retrieval.[17]

One way to use the DAPRECO model is in connection with software models. A software engineering model can be enriched in such a way to interact with the DAPRECO model. This enrichment can help in the application of the data-protection-by-design principle; however, with some minor exceptions, models currently used in software design and development do not take personal data into account.

Because Semantic Web technologies foster reuse, it is possible to connect existing notations to the DAPRECO model instead of creating a new ad hoc notation. The chosen formalism was business process model and notation (BPMN), whose extensible nature and XML serialization make it a perfect selection. Such an approach, however, can be applied to any extensive software engineering model, e.g., Unified Modeling Language.

Some of the opportunities offered by the interplay between the DAPRECO model and BPMN have already been investigated. The most straightforward idea[2] is to highlight the GDPR requirements in the data-processing workflow, appropriately assigning duties to the stakeholders involved, and showing the legal provisions that each data-processing activity (represented by special tasks in the BPMN model) entails. The DAPRECO model can also fit into a BPMN recommender,[3] i.e., a tool that suggests what activities should be placed in certain positions of the workflow, to facilitate the creation of a GDPR-compliant business process.

Aside from its correlation with security standards, the DAPRECO model rouses further interest in compliance checking. Several supervisory authorities, in particular the Commission Nationale de l'informatique et des Libertés (France) and the Commission Nationale pour la Protection des Données (Luxembourg), are currently evaluating the DAPRECO model as a possible option to integrate in their compliance tools, and, in particular, to index the flow of data-processing activities, thus extending the tools with Semantic Web features.

Compliance checking of data-processing activities is an ex post validation, in that it verifies whether an existing system correctly deals with the GDPR requirements; however, the DAPRECO model could also assist

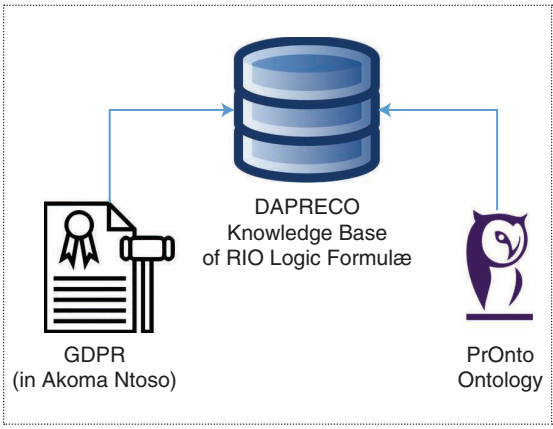

Figure 1. Components of the DAPRECO model. RIO: reified input/output.

compliance ex ante. A sample application would be to integrate the model in workflow design tools to develop systems that aim at being compliant even before they are deployed. Such an approach respects the philosophy of the data-protection-by-design principle contained in Article 25 of the GDPR.

Building the Model

Figure 1 shows a high-level view of the DAPRECO model, which consists of three distinct and connected components (see Figure 1):

1. a machine-readable version of the legal text
2. a semantic context of the law, i.e., a legal ontology
3. a machine-readable and machine-processable representation of the logic of the legal provisions.

Although the structure per se is general and could be applied to any legal text in any domain, in the case of DAPRECO it has been used to model the GDPR (and, to some extent, security standards addressing data protection); therefore, the text will refer to that specific application.

As a result, the three perspectives of the GDPR (legal text, semantic model, and provisions) are kept separate and can have individual relevance, with the possibility of using each of them separately. The full model, however, requires that all three components are in place. The components are built in such a way that they can refer to each other. For instance, from each concept in the semantic model, it is possible to retrieve all of the legal provisions that refer to that concept, regardless of the language used. From an article or paragraph in the legal text, it is possible to extract all the legal provisions that that text portion entails, and so on. Currently, the model of the legal text exists only in English. Information retrieval using different languages requires that a model of the legal text from other languages is built as well. As will be shown further in this article, the XML format used for all three components allows this connection to be very straightforward and easily navigated using simple software tools.

The following sections will briefly describe each of the three components, providing references to published works (where available) for further reading.

The Machine-Readable Legal Text: Akoma Ntoso

To translate the legal text in a machine-readable format, the Akoma Ntoso[18] standard was used. Akoma Ntoso, recently approved as an Organization for the Advancement of Structured Information Standard, is an XML tagset built to represent legal documents such as statutes, decisions, recommendations, and so on. It contains tags to structurally identify specific portions of the text. Tags can be used to identify articles within a law, paragraphs and points within an article, or to separate a preamble from the prescriptive part of the law. With respect to EU legislation, Akoma Ntoso includes appropriate tags to denote the recitals.

Akoma Ntoso also sports two important features that are very useful when dealing with legal sources. First, utilizing the Functional Requirements for Bibliographic Records (FRBR) standard,[19] it provides tags that identify important metadata concerning the law. Metadata include the author, the language, and, most importantly, the temporal and spatial context and versioning. Tracking the succession of laws over time, especially for laws that are frequently subject to amendments, can be a problematic issue. This Akoma Ntoso feature helps determine when and where a legal source is applicable and ascertain which text of the law correctly applies to a situation in a certain place and time.

Second, Akoma Ntoso labels specific portions of text with unique identifiers. This feature allows for the inclusion of references, both within the same law and to other legal sources, including specific versions of the law in case the text changes over time. Additionally, this feature guarantees the language neutrality of the model, because a portion of text is referred to using its identifier, and versions of the legal texts in different languages will all contain the same identifier. In other words, a first step allows for the selection of a specific provision, while a subsequent step allows for the specification of a language for the text of the provision.

Structuring the GDPR in the Akoma Ntoso format is the most straightforward of the three components. Because it is only structured text, there is no semantic meaning involved and therefore, no legal knowledge, apart from a very basic understanding of the structure of a legal document, is required to build this component. Still, building this document requires careful work, as the GDPR is a rather large legal text. In the DAPRECO model, only

the English version of the GDPR has been converted to Akoma Ntoso. The GIT project[17] contains the Akoma Ntoso model of the GDPR in the Resources folder.

Although the legal text in Akoma Ntoso is a component of a larger model, it can be used in a stand-alone fashion, integrated in tools that are capable of exploiting its features. For example, textual searches in Akoma Ntoso can provide more fine-grained results than on pure text documents, returning not only text snippets but specific provisions of the GDPR. Also, the aforementioned versioning tags enable quick retrieval of the regulation's evolution over time. However, the Akoma Ntoso component functions much better when combined with other parts of the DAPRECO model.

The Semantic Model: PrOnto Ontology

Notwithstanding its textual importance, Akoma Ntoso remains a structured content with human-readable text, but conveys little meaning in a computer environment. That is the role of the semantic model: to describe the concepts used (implicitly or explicitly) in the legal text and the relationships between them.

The formal model chosen for the semantic description of the GDPR in DAPRECO is that of an ontology, and, more specifically, a legal ontology. Legal ontologies are generally created to describe a legal system or norms;[4] however, they can express a number of different perspectives, from general knowledge to specific domain terminology. As is widely known, a domain ontology does not define anything; rather, it describes the concepts that already exist in its domain and the relations among them.

The ontology used in the DAPRECO model is called *PrOnto*. As the extended name implies, it aims at describing not only the concepts used in data protection but also in privacy, since there is a significant overlap between the two domains. Furthermore, the ontology does not aim at modeling solely the GDPR, but data-protection legislation in a more general perspective, including Member State laws and also laws from countries outside the EU (and thus not subject to the GDPR). This does not mean, however, that the ontology covers all such legislation. Rather, it describes the concepts in a general way so that concepts from legislation other than the GDPR can easily be connected to PrOnto, following a modular structure. This is facilitated by the fact that data-protection laws outside the EU generally borrow legal concepts and data-protection rules from the terminology that has developed in Europe over the past decades and that is now at the core of the GDPR. Some examples can be seen in the data-protection legislation of Canada, Japan, Singapore, Western Africa, and so on.

PrOnto[5–8] has been designed according to the methodology for building legal ontology (MeLOn).[9] MeLOn is an ontology engineering methodology, roughly based on the following steps:

- the creation of competency questions (CQs), i.e., questions that the ontology is expected to answer
- the definition of metrics that evaluate the ontology
- the creation of a glossary of the concepts used in the domain
- the construction of the ontology using appropriate tools
- the evaluation and measurement of CQs using quality metrics
- publication and documentation.

The MeLOn methodology follows standard principles of minimization, which may be found within the main surveys on computational ontology design and evaluation.[10] As a general rule in ontology engineering, design principles such as minimization and avoiding redundancy are needed to achieve computational efficiency.[10]

As shown in Figure 2, PrOnto is structurally made up of five interconnected modules. The five modules, each with its own classes, represent the following portions of the ontology:

1. document and data
2. actors and roles
3. processing and workflows
4. legal rules and deontic formulas
5. purposes and legal bases.

Documents and data are mainly personal data, which refer to the data subject. The role of data subject can be filled by an agent who is a physical person. Data are processed following a given workflow plan of actions and must be performed according to a legal basis that provides the lawfulness of the processing. Each processing activity involves a controller, a processor, and other roles played by many actors (physical or legal persons). Depending on their roles, actors have specific obligations and rights. These rights and obligations are often linked to documents such as terms of use, information, privacy policies, and consent forms.

PrOnto also relies on several existing ontologies, such as those that express concepts used by the FRBR model or time and context. For a full list of the external ontologies used by PrOnto.[5–8]

Technically speaking, PrOnto is built using Web Ontology Language (OWL). OWL, in itself, is an abstract language that can be expressed using several different syntaxes. OWL syntaxes can differ significantly, from those more suited to be understood by humans to those that are more fit for machine processing. Some of the syntaxes are XML languages, and therefore ideal for interoperability purposes, and, in the case of DAPRECO, used to communicate with other components of the

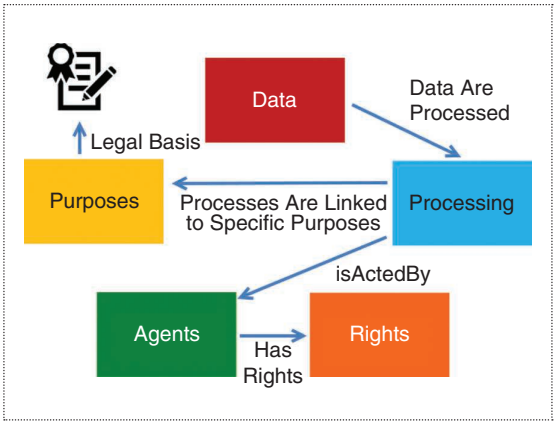

Figure 2. The modular structure of PrOnto.

DAPRECO model. In particular, the Resource Description Framework/XML syntax is supported by a huge variety of tools, some of which are libraries that can be easily integrated into software programs.

The ontology is also language neutral, as the classes have identifiers that follow the English terminology; however, labels can easily be added to the ontological elements to include the various language translations. Language neutrality is, pointedly, a feature of OWL.

Similar to the other DAPRECO components, PrOnto can be used as a stand-alone asset in tools that can use ontologies. Among the various possibilities, it is possible to query the ontology (using languages such as SPARQL) for various purposes connected to information retrieval, such as legal reasoning[6] or assessing legal compliance.[7]

Although theoretically possible, (by means of universal resource names) to connect specific terms in the text to elements in the ontology, such a connection would deviate from the purposes of the Akoma Ntoso representation of the GDPR, as it is meant to contain only the legal text. On the other hand, the ontology does not contain any connection to GDPR provisions. This is perfectly reasonable because the same concept is used repeatedly throughout the legal text, and there would be no point in connecting one concept to specific provisions, except, perhaps, to include an official definition (where available) as a comment, e.g., from Article 4 of the GDPR. The connection between the various elements and the pivot of the DAPRECO model, is therefore in the third component.

Legal Rules: The DAPRECO Knowledge Base

PrOnto was built in light of minimization principles, which are intended to avoid redundancy and enhance computational efficiency as well as human readability. The (negative) side effect of this architectural choice is that PrOnto, although being a useful "conceptual backbone" to achieve cross-document navigation and search, cannot be used to also achieve fine-grained GDPR compliance checking.

To this end, a separate repository, called the *DAPRECO knowledge base*, the current version of which is freely available online,[20] constitutes the third component of the DAPRECO model. The DAPRECO knowledge base contains the deontic rules of the GDPR provisions, expressed in a logic formalism called *reified Input/Output* (*RIO*) logic.[11] RIO logic extends I/O logic using the benefits of reification[11] to simplify the logic structure and avoid nested formulas. The meaning that the logic formulas contained herein convey tries to match that of the legal provisions. In other words, where the legal text carries the textual perspective and the ontology carries the semantic one, the formulas carry the deontic perspective.

The provisions of the GDPR can express different kinds of rules, matched by the corresponding type of formulas:

- constitutive rules or standard logic implications (if *A* is true, then *B* is true)
- deontic rules, further divided into:
  - obligations (if *A* is true, then *B* must be true)
  - prohibitions (if *A* is true, then *B* must not be true)
  - permissions (if *A* is true, then *B* may be true).

Each formula in the knowledge base represents one specific legal provision, which can be a paragraph of the legal text or part of it.

These assertions entail some degree of interpretation. Because they might be overridden by more authoritative or posterior interpretations, the defeasible rules in the DAPRECO knowledge base can be replaced by different rules. Defeasibility guarantees that no conflicting rules are ever present in the knowledge base, as one of them will definitely prevail. Using defeasibility, which is one of the features of RIO logic, the DAPRECO knowledge base can be extended by incorporating additional legal interpretations (possibly incompatible with each other) of the GDPR provisions.[1] Defeasibility is not supported in PrOnto, as it is not a feature of the OWL language.

The formulas are encoded in LegalRuleML,[21] a new XML legal standard aimed at representing the semantic/logical content of legal documents. LegalRuleML is an XML format that can be used to express legal provisions, and its artifacts can be suited to model RIO formulas. Through LegalRuleML, the logic formulas were connected to both the Akoma Ntoso representation of the GDPR and the concepts in PrOnto, so that it was possible to navigate between the formulas, PrOnto, and the structural items (e.g., articles, paragraphs, and so on) of the GDPR.

The XML structure of LegalRuleML allows for it to be easily managed by leveraging existing (de)serializers. Conversely, as there are currently no tools capable of processing RIO logic, a prototype ad hoc parser was created to work with the RIO formulas.[17]

Model Validation

To be effectively used and accepted by a wide community, the DAPRECO model must be validated. This is not an easy task, more so in light of the heterogeneous nature of its three components, which stem from different research domains that cause their testing techniques to differ as well.

Akoma Ntoso

The legal text in Akoma Ntoso does not require a thorough testing, but merely the application of a tagset to the text. These tags have a fixed structure and (almost) rigid meanings, meanings that are not about linguistics or legal interpretation but about the text structure (e.g., articles, paragraphs, recitals, cross-reference links, links to external documents, and so on). Of course, some design choices are made during the construction of the model (such as the naming of identifiers), resulting in error-prone tagging. However, its well-formed structure and conformity to the Akoma Ntoso schema are guaranteed by using simple XML Schema Definition validators. That removed, the only testing that is needed for this component is a double check of the structure, verifying in particular, the following issues:

- unusual paragraph numbering (for example, Article 8.1 is made up of two separate portions of text that are in the same paragraph)
- complex article structure (for example, Articles 6.1, 6.2, and 12.5 have lists of points and then a final portion of text that is still part of the paragraph)
- cross-references, especially in articles comprising many (such as Article 82)
- references to external sources, such as in Article 43.1(a).

PrOnto

Ontology testing is a well-consolidated research branch. Because PrOnto was designed according to a solid ontology engineering methodology, its testing leverages existing methodologies. A preliminary ontology evaluation under the MeLOn methodology used to develop PrOnto is OntoCheck,[22] whose purpose is to provide a first-level assessment of the ontology by checking the correctness of the metadata (and in particular of cardinalities) and of the naming conventions used. A correctness check of the ontology can also be achieved using visualization tools such as WebVOWL.[23] Further testing techniques for the ontology, borrowed from software engineering literature, include coverage testing and mutation testing.[12]

The evaluation of the ontology's semantic soundness, i.e., how well the ontology describes the domain, is based on CQs. This kind of evaluation is currently ongoing, and some partial results (actual CQs with their SPARQL translation) are shown in previous works.[6,7] Furthermore, MeLOn intrinsically addresses an evaluation phase using specific indicators. These include

- completeness of its legal concept definitions
- correctness of its explicit relationships among legal concepts
- coherence of its legal concept modelizations
- applicability to concrete use cases
- effectiveness for the goals
- intuitiveness for nonlegal experts
- computational soundness of its logic and reasoning
- reusability of its ontology and its mapping with other similar ontologies.

The ontology evaluation using MeLOn indicators is currently an ongoing work.

DAPRECO Knowledge Base

Unlike ontology testing, the testing of legal knowledge bases is an uncommon discipline, with few available approaches and tools. This makes the task daunting. Generally,[13] the evaluation of a knowledge base is performed through domain expert validation. Legal knowledge bases, however, present a nearly insurmountable obstacle: logic formulas are typically (as is the case in DAPRECO) built by IT experts who normally have a very basic understanding of law and, quite likely, are unfamiliar with the many subtleties of legal interpretation. The semantics of the formulas that IT experts produce, on the other hand, should undergo the scrutiny of legal experts, who are normally not skilled in logic or XML.

The solution used to overcome this conundrum was to translate the formulas back into a human-readable format. This approach is uncommon in legal informatics, as generally, the objective is the opposite (i.e., from a human-readable legal text to a machine-readable model). It is also quite uncommon in logic, too. Formulas have precise semantics of their own, making the task of rephrasing a superfluous one for a person familiar with the logic. Some works exist on drawing formulas as graphs to visualize the structure in two dimensions, but still, this is meant to improve its readability for logicians, not laymen. This territory still appears to be quite barren. Here, it was performed in two separate steps: 1) an automated translation and 2) a manual postprocessing.

The first step strips the logic symbols, removes redundancy, and introduces words that are not used in logic (such as prepositions and articles) but that make the sentence intelligible. This intermediate output is somewhat comprehensible, but still sounds awkward and faltering and difficult for a legal expert to read. For

this reason, the second step breaks down the intermediate language, extracting the relevant concepts and highlighting them, and presents a kind of report on the content of the provision modeled by the formula. At the time, manual postprocessing was conducted by an IT expert (who is different from the developer) with no legal expertise, and we believe it can be automated in the future. This topic is currently under investigation.

This process required a prephase to test its experimental validity. In particular, the objective was to verify the claim that the human-readable output is both consistent in meaning with the formula that it aims to represent, and understandable by law experts. Thus far, consistency has been ensured by 1) the automated translation, which was checked with the IT expert that proposed the logic in the first place, and 2) the fact that the manual process was performed by another researcher knowledgeable of modal logic and usability. Understandability was gauged using a mixed methodology: a questionnaire given to a small population of legal experts (Ph.D.s or researchers) and the measure of consistency among their answers. The legal experts were asked to answer the same questionnaire about the three expressions of the formula (i.e., logic, automatic translation, and manual breakdown). The results showed an increase in the readability of the formula through its three stages.[14,15]

The next step was to test the usability of the whole approach, that is, whether the human-readable format can be used to provide feedbacks on the quality of the translation of the GDPR into a logic formalism. In other words, to provide quality feedback that an IT professional with expertise in logic may lack due to his or her unfamiliarity with legal issues. To do this, another questionnaire was used, including questions meant to inspect the legal quality of the meaning expressed by a formula. This quality can be measured with several metrics, such as completeness (Is all the required domain knowledge explicitly stated?), conciseness (Is there redundancy in the representation?), and accuracy (Does the formula match the corresponding legal provisions?). Although the results are still preliminary, this approach allows for an extensive evaluation of the legal knowledge base.

The GDPR represents a milestone in the digital agenda of the EU. Not only because it was the first step in a significant revision of existing European laws concerning information technology, nor simply because of the long and strenuous debate that preceded its entry into force, but also for the huge impact that it has had on digital stakeholders. The GDPR also marked a turning point, after which the protection of personal data cannot be taken lightly. The new context calls for a revision of the tools and methodologies that deal with personal data, which must therefore become "GDPR aware."

This article aimed at facilitating this GDPR awareness through a legal model of the regulation. Computer-readable models of legal documents are an expanding segment of research and industry, especially for the purposes of legal compliance.

A model representing the GDPR can be used in many different ways by all of the stakeholders involved in data-processing activities. Design, development, and validation tools can be enhanced to support GDPR requirements. By using GDPR-aware tools and techniques, controllers and processors can more easily build GDPR-compliant systems, thus lowering the risk of sanctions. Data subjects can enjoy a more thorough implementation of the systems used, thanks to stronger enforcement of data-protection rights. Consultants and auditors as well as supervisory and judicial authorities, can more easily check for compliance to detect GDPR violations.

Although still at an early stage, the model is an extremely complex structure currently under testing and revision. It does, however, contain a complete representation of the GDPR, suited for integration with various software tools.

The model presented in this article can also model other legal sources. The GDPR was an ideal field for which to design the model, as it is a rather self-contained normative document, with a limited number of references to other legal sources. However, nothing prevents the creation of similar models for other laws or legal domains. ■

Acknowledgments

This work was supported by the Luxembourg National Research Fund CORE project C16/IS/11333956 "DAPRECO: DAta Protection REgulation Compliance" and the European Union Horizon 2020 Marie Skłodowska-Curie project 690974 "MIREL: MIning and REasoning With Legal Texts."

References

1. C. Bartolini, A. Giurgiu, G. Lenzini, and L. Robaldo, "Towards legal compliance by correlating standards and laws with a semi-automated methodology," in *BNAIC 2016: Modern Trends in Artificial Intelligence*, vol. 765. T. Bosse and B. Bredeweg, Eds. Cham, Switzerland: Springer, 2017, pp. 47–62.
2. C. Bartolini, A. Calabrò, and E. Marchetti, "Enhancing business process modelling with data protection compliance: An ontology-based proposal," in *Proc. 5th Int. Conf. Information Systems Security and Privacy (ICISSP)*, 2019, pp. 421–428. doi: 10.5220/0007392304210428.
3. C. Bartolini, A. Calabrò, and E. Marchetti, "GDPR and business processes: An effective solution," in *Proc. Applications of Intelligent Systems Conf.*, 2019. doi: 10.1145/3309772.3309779.
4. J. Breuker, A. Valente, and R. Winkels, "Use and reuse of legal ontologies in knowledge engineering and information

management," in *Law and the Semantic Web*, vol. 3369. V. R. Benjamins, P. Casanovas, J. Breuker, and A. Gangemi, Eds. Berlin, Germany: Springer-Verlag, 2005, pp. 36–64.
5. M. Palmirani, M. Martoni, A. Rossi, C. Bartolini, and L. Robaldo, "PrOnto: Privacy ontology for legal reasoning," in *Proc. Internationales Rechtsinformatik Symp. (IRIS)*, 2018.
6. M. Palmirani, M. Martoni, A. Rossi, C. Bartolini, and L. Robaldo, "PrOnto: Privacy ontology for legal reasoning," in *Proc. Electronic Government and the Information Systems Perspective, 7th Int. Conf., EGOVIS 2018*, vol. 11,032. A. Kö and E. Francesconi, Eds. Cham, Switzerland: Springer, 2018, pp. 139–152.
7. M. Palmirani, M. Martoni, A. Rossi, C. Bartolini, and L. Robaldo, "PrOnto: Privacy ontology for legal compliance," in *Proc. 18th Eur. Conf. Digital Government (ECDG)*, 2018, pp. 142–151.
8. M. Palmirani, M. Martoni, A. Rossi, C. Bartolini, and L. Robaldo, "Legal ontology for modelling GDPR concepts and norms," in *Proc. 31st Int. Conf. Legal Knowledge and Information Systems (JURIX)*, 2018, pp. 91–100. doi: 10.3233/978-1-61499-935-5-91.
9. M. Mockus and M. Palmirani, "Legal ontology for open government data mashups," in *Proc. 7th Int. Conf. E-Democracy and Open Government (CeDEM)*, 2017, pp. 113–124.
10. J. Bandeira, I. I. Bittencourt, P. Espinheira, and S. Isotani, FOCA: A methodology for ontology evaluation. 2016. [Online]. Available: https://arxiv.org/abs/1612.03353
11. L. Robaldo and X. Sun, "Reified input/output logic, combining input/output logic and reification to represent norms coming from existing legislation," *J. Logic Comput.*, vol. 27, no. 8, pp. 2471–2503, 2017. doi: 10.1093/logcom/exx009, [Online]. Available: https://academic.oup.com/logcom/article/27/8/2471/3098296
12. C. Bartolini, "Software testing techniques revisited for OWL ontologies," in *Proc. Model-Driven Engineering and Software Development, 4th Int. Conf., MODELSWARD 2016*, vol. 692. S. Hammoudi, L. Ferreira Pires, B. Selic, and P. Desfray, Eds. Cham, Switzerland: Springer, 2017, pp. 132–153.
13. A. Stranieri and J. Zeleznikow, "The evaluation of legal knowledge based systems," in *Proc. 7th Int. Conf. Artificial Intelligence and Law (ICAIL)*, 1999, pp. 18–24.
14. C. Bartolini, G. Lenzini, and C. Santos, "An agile approach to validate a formal representation of the GDPR," in *New Frontiers in Artificial Intelligence* (Lecture Notes in Artificial Intelligence Series 11,717). Cham, Switzerland: Springer, 2019, to be published.
15. C. Bartolini, G. Lenzini, and C. Santos, "A legal validation of a formal representation of articles of the GDPR," in *Proc. 2nd Workshop Technologies Regulatory Compliance (TeReCom), 31st Int. Conf. Legal Knowledge and Information Systems (JURIX)*, 2018, pp. 111–124. [Online]. Available: http://ceur-ws.org/Vol-2309/10.pdf
16. CNIL, "The CNIL's restricted committee imposes a financial penalty of 50 million euros against Google LLC," Jan. 21, 2019. [Online]. Available: https://www.cnil.fr/en/cnils-restricted-committee-imposes-financial-penalty-50-million-euros-against-google-llc
17. GitHub, "Guerret/lu.uni.dapreco.parser." Accessed on: Sept. 15, 2019. [Online]. Available: https://github.com/guerret/lu.uni.dapreco.parser
18. Akoma Ntoso. Accessed on: Sept. 15, 2019. [Online]. Available: http://www.akomantoso.org/
19. IFLA Study Group on the Functional Requirements for Bibliographic Records. (1997). Functional requirements for bibliographic records. IFLA. The Hague, The Netherlands. [Online]. Available: https://www.ifla.org/publications/functional-requirements-for-bibliographic-records
20. GitHub, "Dapreco/daprecokb." Accessed on: Sept. 15, 2019. [Online]. Available: https://github.com/dapreco/daprecokb
21. OASIS, "OASIS LegalRuleML TC." Accessed on: Sept. 15, 2019. [Online]. Available: https://www.oasis-open.org/committees/legalruleml/
22. Protégé Wiki, "OntoCheck." Accessed on: Sept. 15, 2019. [Online]. Available: https://protegewiki.stanford.edu/wiki/OntoCheck
23. WebVOWL, "WebVOWL: Web-based visualization of ontologies." Accessed on: Sept. 15, 2019. [Online]. Available: http://vowl.visualdataweb.org/webvowl.html

Cesare Bartolini is a notary public and contributes to research activities at the University of Luxembourg. His current research is focused on law and IT, including computer-related law such as data protection and copyright and computer models for legal applications. Bartolini received a Ph.D. in computer engineering from the Scuola Superiore Sant'Anna, Pisa, Italy, in 2007. Contact him at cbartolini@notariato.it.

Gabriele Lenzini is a senior research scientist and an assistant professor at the University of Luxembourg. His current research is focused on the design and analysis of security and privacy in physical and digital systems, particularly in domains where security and privacy lose their sheer technical perspective. Lenzini received a Ph.D. in computer science from the University of Twente, The Netherlands, in 2005. Contact him at gabriele.lenzini@uni.lu.

Livio Robaldo is a postdoc researcher at the University of Luxembourg. His current research is focused on legal informatics, specifically the application of natural language processing and natural language semantics to the legal domain. Robaldo received a Ph.D. in computer science from the University of Turin, Italy, in 2007. Contact him at livio.robaldo@uni.lu.