A Semantic Specification for Data Protection Impact Assessments (DPIA)

Harshvardhan J. PANDIT [a,1],
[a] *ADAPT Centre, Trinity College Dublin, Dublin, Ireland*

Abstract. The GDPR requires assessing and conducting a Data Protection Impact Assessment (DPIA) for processing of personal data that may result in high risk and impact to the data subjects. Documenting this process requires information about processing activities, entities and their roles, risks, mitigations and resulting impacts, and consultations. Impact assessments are complex activities where stakeholders face difficulties to identify relevant risks and mitigations, especially for emerging technologies and specific considerations in their use-cases, and to document outcomes in a consistent and reusable manner. We address this challenge by utilising linked-data to represent DPIA related information so that it can be better managed and shared in an interoperable manner. For this, we consulted the guidance documents produced by EU Data Protection Authorities (DPA) regarding DPIA and by ENISA regarding risk management. The outcome of our efforts is an extension to the Data Privacy Vocabulary (DPV) for documenting DPIAs and an ontology for risk management based on ISO 31000 family of standards. Our contributions fill an important gap within the state of the art, and paves the way for shared impact assessments with future regulations such as for AI and Cybersecurity.

Keywords. GDPR, DPIA, Risk Management, ISO, Semantic-Web

1. Introduction

1.1. Motivation

The EU's General Data Protection Regulation (GDPR) [1] requires every Data Controller to assess and document whether their processing is "likely to result in a high risk to the rights and freedoms" of individuals (i.e. *high-risk*[2]), and if so - to carry out a 'Data Protection Impact Assessment (DPIA)'. A DPIA is essentially a three-step iterative risk governance process where the organisation first identifies its activities, then checks whether any DPIA-requiring criteria is met, and if yes - conduct a DPIA (see more in Section 2.1). GDPR does not impose a strict process for how organisations have to conduct their risk and impact assessments, but instead specifies only broad requirements. Data Protection Authorities (DPA), tasked with enforcing GDPR, have published (on respective websites) guidance and tools related to compliance, including DPIA and risk governance.

We identify five important challenges regarding DPIAs present in the current landscape that serve as motivation for this work. (1) DPIAs can involve multiple stakehold-

[1] Corresponding Author: Harshvardhan J. Pandit ; E-mail: pandith@tcd.ie

[2] Hereafter, *high-risk* is used as a shortened form of "high risk to the rights and freedoms of natural persons"

ers (e.g. Data Processors) which creates information dependencies (e.g. measures implemented by processors). (2) Since DPIAs must be specific, controllers conducting similar DPIAs will repeat information and tasks. (3) Despite existing standards for risk management, there is variance in methodologies that prevents common universal solutions. (4) Current documentation norms are heavily human-oriented (e.g. spreadsheets, PDF), which severely limit development and application of tools for DPIAs. (5) Solutions do not take into account that high-risk impact assessments are a form of shared activity i.e. they share processing activity information, risks, and impacts with other GDPR requirements (e.g. Register of Processing Activities (ROPA), data transfers), and have overlaps with similar assessments in aligned regulations, e.g. the EU's proposal for AI Act [2].

The state of the art contains multifaceted application-specific solutions for expressing risks, DPIA methodologies, and GDPR compliance. In particular, they demonstrate advantages of semantic web technologies for: (i) specialising for a use-case; (ii) interoperability between stakeholders and tools; (iii) creating shared knowledge-bases; and (iv) developing tooling for machine-based compliance. However, there are two important gaps that have not been addressed: impact assessments and documenting DPIAs.

1.2. Contributions of this Work

We take the first step towards improving the DPIA processes by enabling sharing and reuse of information required for risk/impact assessments through the use of semantic web technologies. Our approach reflects the positioning of DPIAs within a broader framework of information and compliance management associated with GDPR. Thus, rather than creating an ontology solely dedicated to representing DPIA, we extend an existing ontology - the Data Privacy Vocabulary (DPV) produced by the Data Privacy Vocabularies and Controls Community Group[3] (DPVCG) as the state of the art (see Section 2.2). DPV provides a comprehensive taxonomy of data processing related concepts, including rudimentary concepts for risks and DPIA, that are meant to be jurisdiction and domain agnostic, with a separate extension (`dpv-gdpr`) providing GDPR specific concepts. We identified and proposed concepts currently missing in (core) DPV, and from these developed a DPIA specification as an extension (called `DPV-DPIA`). For expressing risk/impact assessments - we developed an ontology based on the ISO 31000 family of risk-related standards. For expressing impacts to fundamental rights and freedoms, we created a thesauri from the EU Charter regarding rights and freedoms[4].

To ensure the specification is useful and practical for stakeholders, we based it on DPA guidelines and tools to first ensure important requirements are met (see Section 3.1). We then modelled real-world instances of (publicly available) DPIAs as a form of reflective evaluation, and to demonstrate sharing of knowledge we used the DPV-specified concepts within French DPA's (CNIL) DPIA tool (see Section 4). We conclude with a discussion (see Section 5) on identified and perceived limitations of our work, and the pragmatism of developing shared impact assessments for EU's regulatory landscape.

To summarise, our major contributions are: (i) Machine-readable DPIA specification; and (ii) Enabling reuse and sharing of risks, mitigations, and impacts through linked data. Minor contributions include: (i) Risk ontology based on ISO 31000 family of standards; (ii) Thesauri of EU fundamental rights and freedoms; (iii) Collection of risks, mit-

[3]Disclaimer: The lead author currently chairs the DPVCG.

[4]`http://data.europa.eu/eli/treaty/char_2012/oj`

igations, and impacts from literature; (iv) Extension of DPV and state of the art; and (v) Practical discussions towards developing shared impact assessments.

2. Background and State of the Art

2.1. GDPR and Data Protection Impact Assessments (DPIA)

GDPR's Article 35 prescribes requirements for assessing necessity of DPIAs based on potential for high-risk, and for carrying out a DPIA if a criteria is met. In this, it describes conditions that always need a DPIA and lays down the basis where DPAs can specify further rules on conditions that do/don't require DPIA. It also describes consultation of stakeholders such as Data Protection Officers (DPO) and data subjects where necessary.

In order to determine necessity, controllers require descriptions of processing activities in terms of specific criteria, for example the scale and scope of data (Art.35-3b), or whether automated decision making and profiling operations are involved (Art.35-3a). DPA guidelines provide additional nuanced descriptions of concepts that are relevant for determining risk, impact, and the basis on which DPIAs should be conducted.

While GDPR intends to provide harmonised requirements for DPIAs, individual DPAs have taken different approaches with deviations regarding use of organisational processes related to management practices and risk governance - which are not necessarily directly associated with a DPIA. For example, as part of the DPIA templates, both AEPD (Spanish DPA) and CNIL (French DPA) ask about the organisation's "internal practices and context" which includes " organisation's structure, functions and competencies, adopted policies, norms and standards, organisational maturity objectives and in general the organisation's culture". Owing to this, organisations have difficulties in determining what requirements a DPIA must meet given that the guidance is varied, complex, nuanced, and difficult to judge for sufficiency. Additionally, Georgiadis et al. [3] conducted a systemic literature review on the different privacy and data protection risks specified within the state of the art, with a conclusion on the necessity to further develop better DPIA methodologies due to organisation's limited knowledge on this topic.

2.2. Models for DPIAs and Risk Assessments

There are several domain and application specific approaches for modelling risk in ontological form. Some examples are: Agrawal's [4] ontology based on ISO/IEC 27005:2011 risk management standard, Ameida et al's [5] conceptual enterprise architecture models for organisational risk management based on ISO 31000, Rosa et al's. [6] ontology for IT risk management based on ISO 31000, Vicente et al's. [7] high-level model for organisational risk governance, and Hayes et al's. [8] ontological model of online privacy risks and harms. While these approaches model risk concepts in ontological form, they focus on organisational perspective of risks (e.g. economic), or on generalised concepts (e.g. philosophical) that are not sufficient for expressing impacts as needed for a DPIA.

In approaches that represent DPIA related information, GDPRtEXT [9] provides insufficient concepts related to DPIA. PrOnto [10] specifies DPIA as a workflow with steps and different categorisations of risk. Data Privacy Vocabulary[5] (DPV) [11] provides

[5] `https://w3id.org/dpv/`

comprehensive taxonomies for describing personal data processing activities, which includes DPIA and risk concepts. In approaches related to automating DPIA processes, Dashti et al. [12] explore automation of DPIA based on rule-based mechanisms to identify alternatives for less risky implementations. And Saniei [13] proposes use of semantic web technologies to represent DPIA related knowledge and to use rules and inferences to identify relevant obligations and actions, with ongoing work [14] in collecting competency questions and creating a vocabulary - which was useful for this work.

Of these approaches, none provided all necessary concepts or could be readily used. Of these, DPV was the most suitable choice to extend given that it is: (a) most comprehensive; (b) open access; (c) has a mechanism for updating through DPVCG. This finding is backed by a recent survey by Esteves et al [15] regarding modelling of GDPR related information flows that also included DPIA as a factor in investigation, with favourable reviews for DPV, though it found no suitably complete vocabulary for DPIAs.

3. DPIA Specification

3.1. Requirements and Objectives

For understanding DPIA information requirements, we utilised EU DPA provided guidelines, tools, and templates. For non-English documents, we utilised machine-translation to convert them, and manually inspected them for correctness (relying on the author's familiarity with information). In particular, we focused on identifying requirements regarding: (1) personal data processing activities; (2) DPIA necessity assessment and outcomes; (3) risk/impact assessments and outcome; (4) conditions regarded as high-risk, and requirement for a DPIA; and (5) documentation required for maintaining DPIAs.

As outlined earlier in Section 2.1, these documents provide a wide range of information requirements that do not necessarily relate directly to DPIAs as stated in GDPR Art.35. In particular, the DPAs from Spain, France, and UK have provided comprehensive documentation which does not provide justification for how these are connected to specific legal requirements, and often go well beyond GDPR and into describing internal risk and governance procedures. The scope and breadth of these practices necessitate a much larger study given their complexity, variance, and connection to legal requirements. We focused on representing relevant information at a 'high-level' while also being sufficient in terms of GDPR requirements. This led to identifying the following specific requirements regarding documentation of information: (1) provenance records for DPIA in terms of processes and actors; (2) representing risk/impact assessments; (3) description of processing activities; and (4) risks, mitigations, and impacts.

3.2. Specification Overview

The specification, available online[6], models three categories of information: provenance and status of DPIA, processing activities associated with a DPIA, and the risks/impacts involved in that DPIA. In this, the existing[7] DPV concept `dpv:DPIA` is reused as a focal point with further specialisation into three aspects: `DPIANecessityAssessment` repre-

[6] `https://w3id.org/dpv/dpv-gdpr/dpia`

[7] For brevity, concepts presented as contribution are specified without prefix, and existing ones with prefix.

senting determination of whether a DPIA is required; `DPIAProcedure` for risks, impacts, and mitigations being investigated and documented; and `DPIAOutcome` for documenting the outcomes of a DPIA in terms of continuation of processing. Figure 1 represents these along with other core concepts to provide an overview of the specification.

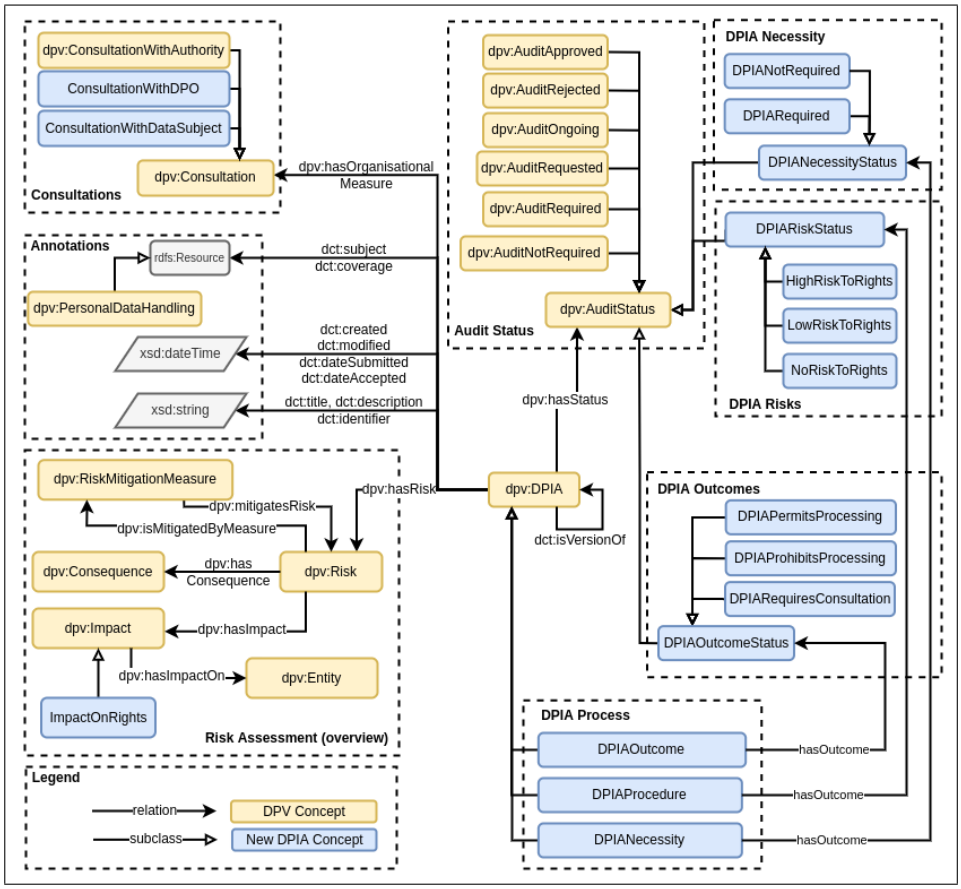

Figure 1. Overview of the DPIA Specification

DPIAs require documentation of provenance information regarding when it took place (temporal information), and who was involved (agents, e.g. approval). For these, we reuse Dublin Core Metadata Innovation[8] (DCMI) terms for temporal information (`dct:created, dct:modified, dct:dateSubmitted, dct:dateAccepted, dct:temporal, dct:valid`), conformance e.g. codes of conduct (`dct:conformsTo`), descriptions (`dct:title, dct:description`), identifier or version (`dct:identifier, dct:isVersionOf`), and subject or scope of DPIA (`dct:subject, dct:coverage`).

To record outcomes of DPIA processes, we consider a DPIA to be a form of *Audit* and use `dpv:hasStatus` with the appropriate `dpv:AuditStatus`. For example, `DPIANecessityAssessment` with `dpv:AuditRequired` indicates a necessity assessment is required, whereas `DPIAProcedure` with `dpv:AuditApproved` indicates the

[8] `https://dublincore.org/specifications/dublin-core/dcmi-terms/`

DPIA results were approved (e.g. by a DPO). The relation `hasOutcome` was created to indicate status of each DPIA process as - (i) for `dpv:DPIANecessityAssessment`: `DPIANecessityStatus` and specialisations related to whether a DPIA is required or not-required; (ii) for `dpv:DPIAProcedure`: `DPIARiskStatus` and specialisations related to level of risk as high, low, or none; and (iii) for `dpv:DPIAOutcome`: `DPIAOutcomeStatus` and specialisations for whether processing is permitted or prohibited or consultation is required[9]. These represent the broad outcomes to be recorded when carrying out a DPIA in terms of whether risks have been mitigated (or deemed acceptable) and whether processing can (or cannot) be carried out.

For indicating the different stages and processes in conducting and managing DPIA, the concepts `Audit, Approval, Investigation`, and `Review` were created with specific relations (e.g. `hasAudit`) to associate them with the relevant concepts. For indicating specific categories of consultations, the existing concept `dpv:Consultation` was extended as `ConsultationWithDataSubject` and `ConsultationWithDPO` to record their views and inputs within the DPIA process.

For indicating the scope and contents covered within a DPIA, the property `dct:coverage` is reused with `dpv:PersonalDataHandling` instances to indicate the specifics of purposes, processing operations, personal data categories, entities (e.g. controllers, recipients), technical & organisational measures, legal bases, and other details. Here, `dct:subject` can be optionally used to indicate a DPIA (and its associated processing activities) relate to a specific topic, such as a service or a product.

The existing risk concepts in DPV are used as: to indicate risks (`dpv:Risk, dpv:hasRisk`), mitigations (`dpv:RiskMitigationMeasure, dpv:mitigatesRisk`), consequences (`dpv:Consequence, dpv:hasConsequence`), and impacts (`dpv:Impact, dpv:hasImpact, dpv:hasImpactOn`). For more specific risk assessment information, such as risk levels and severity, the ISO 31000 based risk ontology is used.

3.3. Extending DPV

We found DPV currently has several concepts missing regarding not only DPIAs, but also those related to descriptions of processing activities beyond what is needed from a risk/impact perspective. For example, one of the prominent criteria in determining whether processing is likely to be high-risk is the understanding of scale and scope regarding personal data, processing activities, and data subjects. Rather than specifying their expression only within what is needed for a DPIA, we consider these concepts to be useful in other tasks and assessments, and thus propose their inclusion in DPV.

An important addition we propose is the indication of certain `Scale` concepts along with commonly used qualitative terms[10] that relates to a measurement of dimension of some other concept. `DataVolume` indicates the scale of personal data being processed with qualifiers (from larger to smaller in context) - {`Huge, Large, Medium, Small, Sporadic, Singular`}. `DataSubjectScale` indicates a measurement of the scale of data subjects with the same qualifiers as data volume. `GeographicScale` indicates the geo-physical scale (e.g. for processing activities or data subjects) as

[9] These reflect the possibility where a first iteration of DPIA identifies a high-risk which cannot be mitigated by the second, leading to a consultation with a DPA.

[10] Here concepts are derived from specific obligations, e.g. 'large scale of data', which gives concepts for more/less than 'large'. The specifics of whether something is 'large' is to be interpreted contextually.

{`Global, NearlyGlobal, MultiNational, National, Regional, Locality, WithinEnvironment`} with the last item referring to instances such as on device. Separate from scale, we also propose the modelling of `Scope` as a concept referring to the *extent or range* of other concepts such as processing activities. To differentiate between scale and scope, the former refers to a *measurement* such as volume or number whereas the latter relates to *variance* such as categories or dimensions.

Along with scale and scope as new concepts, we also propose remodelling existing concepts that relate to either. These include `dpv:Frequency` which indicates temporal periodicity, and should be a specialisation of `Scale` with qualifiers {`Continuous, Often, Sporadic, Singular`}. Similarly, `dpv:Duration` should also be a specialisation of `Scale` with qualifiers {`Endless, TemporalDuration, UntilEvent, UntilTime, FixedOccurences`} to represent the different categories of durations that are utilised regarding personal data processing activities.

In our analysis of the DPIA documents, a large amount of information was expected to be recorded in the form of "*justification*" for why something was or was not done regarding the requirements set out by GDPR or DPAs. This information would typically be indicated as a textual description (i.e. free-form text) accompanying some question or concept. Given the importance of this concept in legal compliance, and the necessity to record this information in a form more explicit than (mere) descriptions, we propose the property `hasJustification` for inclusion in DPV. The concept enables associating a textual statement, or document, or specific concept as the justification for its state or existence, and is also useful beyond DPIAs - such as for acknowledging legal compliance obligations or recording a DPO's statements during an investigation.

We also identified concepts missing regarding processing operations: {`Access, Assess, Filter, Monitor, Modify, Observe, Screen`} - that refer to specific kinds of actions over personal data relevant when conducting a DPIA. Other missing concepts relate to certain categories of purposes, and technical and organisational measures, in particular those that are relevant in determining whether processing activities require a DPIA. Similarly, missing concepts were also identified regarding personal data categories (for the DPV-PD extension[11]) relating to behavioural, financial, professional, and in particular their indication as sensitive and special categories. We have shared these findings with the DPVCG through the public mailing list[12].

3.4. Risk Ontology based on ISO 31000 family of Documents

As stated before, DPV offers a few abstract risk-related concepts that are not sufficient to represent risks, mitigations, consequences, impacts, and their assessments as required within a DPIA. Additionally, the state of the art does not provide a suitable risk ontology that can be used readily or adapted for this work. Due to these reasons, we initiated development of a risk ontology. For this, we looked towards existing standardised forms of risk management, but found no consistent or common modelling of risk or its associated processes. Our experience revealed a fragmented landscape consisting of often conflicting use of terms and a high degree of use-case specific solutions within both academia and industry. The few standardised approaches regarding risk limited themselves to ei-

[11] `https://w3id.org/dpv/dpv-pd/`

[12] `https://lists.w3.org/Archives/Public/public-dpvcg/2022May/0003.html`

ther providing an organisational perspective of risk or forced the use of domain-specific terms that raised questions regarding its usefulness outside those domains.

Within these, the ISO 31000 family of standards provide a set of harmonised and consensus-building documents that provide guidance, principles, and vocabularies associated with risk management and risk governance. Other approaches also exist that are more systematised - such as the US Government's NIST Risk Management framework[13] [16], or are intuitive for businesses - such as FAIR Risk Management[14].

We decided to utilise the ISO standards due to their global applicability, standardised terminology, involvement and alignment with EU standardisation bodies, and also because one of our future ambitions is to provide a way for expressing utilisation of ISO standards in processing activities, e.g. regarding cloud security. Though it must be noted that the FAIR risk management approach specifies use of an ontology in its modelling of risk concepts, we decided against adopting it in favour of ISO 31000 being standardised.

The two main standards we utilised for our risk ontology were ISO 31000:2018[15] Risk Management Guidelines and 31073:2022[16] Risk Management Vocabulary. From these, we analysed risk-related concepts, definitions, intended uses in these and other documents, and identified relations to create an ontology. Here it is important to state that the resulting ontology is our representation of how the ISO 31000 series can be used for representing risk related information, and that these documents by themselves do not prescribe any specific modelling of relations between the concepts.

We first identified and represented all risk-related concepts from ISO 31073:2022 as a SKOS vocabulary and identified taxonomic (i.e. broader/narrower) relationships between them. This provided us with an overview of what concepts are present in ISO's risk standards and how they relate to each other. We then identified additional relationships between these concepts based on statements from ISO 31073:2022 and ISO 31000:2018 and expressed them as an *OWL* ontology. An overview of the outcome is presented in Figure 2, and the risk ontology is available online[17].

By itself, this risk ontology is sufficient to represent risk-related information required for DPIAs i.e. risk, risk sources, threat actors, consequences and impacts of risks, and their attributes such as likelihoods, severity, and levels. However, in practice, we found variance in how these attributes are used by adopters, for example as differences in risk scale where one set of levels goes from 1 to 5 and another goes from 1 to 10, and yet another that uses only qualitative labels (e.g. high/low). This represented a challenge in modelling use-cases as it prevents a consistent representation of risk-related information.

To address this, we created top-level concepts (e.g. `RiskLevel`) with guidance that any attributes (e.g. risk levels) must follow existing norms where statistical distributions are used to harmonise differences in scales across use-cases. For example, by representing 0 as the lowest possible scale and 1 as the highest, qualitative terms like 'high risk' or 'frequently occurring' are forced to be expressed as values or ranges between `0..1`. While the exact values may differ between use-cases (for example, `0.5` may be high-risk in one situation and `0.9` in another), they are useful to compare the actual importance of concepts and harmonise them when information is shared, reused, or imported. To aid

[13] `https://www.nist.gov/risk-management`
[14] `https://www.fairinstitute.org/fair-risk-management`
[15] `https://www.iso.org/standard/65694.html`
[16] `https://www.iso.org/standard/79637.html`
[17] `https://w3id.org/riskonto`

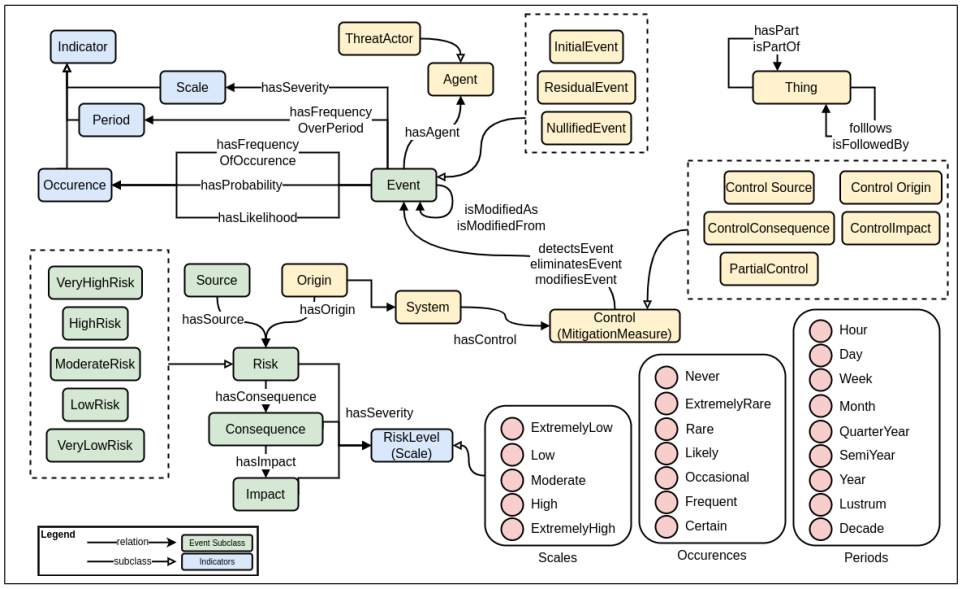

Figure 2. Overview of the Risk Ontology

with the DPIA processes, we provide a list of commonly used scales regarding risk levels and severity based on DPA guidance documents. The risk ontology can thus be used alongside DPV to represent risk-related information with more detail and granularity.

3.5. Thesauri of Fundamental Rights and Freedoms

The consideration of whether processing activities are likely to have an impact on fundamental rights and freedoms of individuals is what determines high-risk within DPIAs. While the actual assessment of whether a use-case has impacts on fundamental rights requires expertise and nuanced interpretations, we focused on the information required for documenting these impacts. This required creating a vocabulary of fundamental rights and freedoms, as defined within the EU Charter of Fundamental Rights, and associating them with impacts of a risk as well as the effect of those impacts.

We started by representing each Article in the document as a `skos:Concept` (e.g. Article 7 as *RightToPrivacy*), with the Title used to structure them (e.g. Title II as Freedoms). Further distinction between these, in particular regarding what constitutes a *right* and a *freedom* would require philosophical interpretations and their application within the legal domain of EU laws [17]. In future, we aim to expand this vocabulary by modelling the concepts from specific clauses within each article, and investigating whether they can be represented as an ontology. The current iteration is available online[18].

To indicate impacts affecting rights and freedoms, we created the concept `ImpactOnRights` as a specialisation of `dpv:Impact` within the DPIA specification, with further specialisations for each identified rights (e.g. `ImpactOnRightToPrivacy`). The decision on whether concepts related to impact on rights should be part of the main DPV vocabu-

[18] `https://w3id.org/rights-vocabulary`

lary requires careful deliberation as the notion of *rights* is not uniformly represented or interpreted in laws across the globe.

3.6. Populating Risks and Mitigation Concepts

Along with concepts related to DPIAs, providing commonly used terms related to risks and mitigations would also benefit adopters in representing their use-cases and documentations. As DPA guidance documents provide a small but good number of examples, we looked for additional concepts to better model industry challenges, and to incorporate and represent as much of the commonly utilised terms and 'good practices'.

We first referred to documents published by the European Union Agency for Cybersecurity[19] (ENISA) which provide an expert collection and overview of cybersecurity related incidents, issues, and methods for addressing them. We identified four candidate documents: (i) Risk Management Standards; (ii) Compendium of Risk Management Frameworks with Potential Interoperability; (iii) Interoperable EU Risk Management Framework; and (iv) Guidelines for SMEs on the security of personal data processing.

We also identified three existing privacy risk methodologies and taxonomies that we plan to integrate into our work: Jakobi et al's list of user-perceived privacy risks [18], Solove's Privacy Harms [19], and LINDDUN [20]. Of these, LINDDUN is notable in that it provides a privacy engineering framework that provides knowledge bases and taxonomies for threats and mitigations associated with software systems. It models 7 threat categories and their mitigations, structured according to the LINDDUN acronym as: Linkability, Identifiability, Non-repudiation, Detectability, Disclosure of Information, Unawareness, and Non-compliance. These will be used to categorise and structure risk concepts from other sources for DPIAs, with the 'threats' in LINDDUN modelled as 'risks' in our work, and 'mitigations' modelled as technical and organisational measures in DPV or risk mitigation measures in DPIA (as appropriate).

4. Applying to Real-World Use-Cases and Tools

4.1. Documenting Real-world Use-cases

To better understand how our specification fits its purpose, we looked for publicly available documents and selected three prominent ones based on quality of information, conclusion of investigation, and their topicality. These relate to DPIAs carried out in Netherlands (and involving government bodies and authorities) for use of Zoom [21], Microsoft Office 365 [22], and Google Apps (GSuite) [23]. All three cases represent complex services and infrastructures, and the large length of reports produced reflect the scope and breadth of information that is considered relevant for their DPIAs.

As we stated in the motivation, these DPIAs are also produced as human-readable documents with no ability to extract, query, or reuse their information. First we analysed the kind of information represented in these reports and whether our work (along with DPV) was sufficient in expressing it. We found that we could represent most of the concepts associated with how the processing takes place, e.g. personal data involved or purposes or data transfers. What we could not represent related to complexities of

[19] `https://www.enisa.europa.eu/publications/`

data collections and transfers, such as where Microsoft and Google combine their data across different services and transfer them outside EU/EEA. We also could not represent information about *absence* - such as a specific measure not being present, or *negation* - such as when a company asserted that they do not perform some activity. This resulted in gaps associated with information the DPIA was generated based upon.

The information regarding risks, mitigations, consequences, and impacts in most cases was directly associated with specific implementation details and technologies, and therefore could be represented using DPV and our DPIA and risk ontologies. However some of the consequences and impacts were difficult to quantify since they related to specific behaviours of individuals or groups, and were hypothetical scenarios that could not be specified with likelihood or severity. We observed this pattern in all three documents. We perform a self-reflection on this experience in Section 5.

4.2. Use with CNIL's PIA Tool

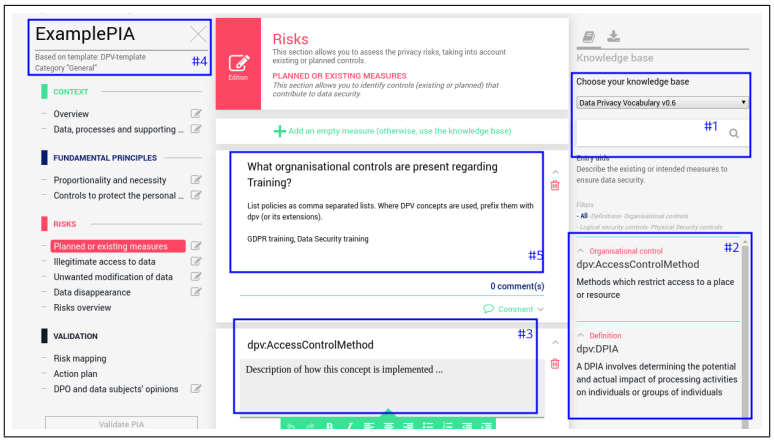

Figure 3. Example of CNIL's PIA tool modified for using DPV as: (1) a knowledge base; (2) providing concepts in relevant sections as controls and definitions; (3) selectively adding concepts to DPIA with description; (4) custom templates explaining how to use DPV concepts; (5) guided data entry for using DPV concepts.

CNIL, the French DPA, has developed the PIA (Privacy Impact Assessment) tool that assists organisations in documenting, reviewing, and sharing information regarding DPIAs. The tool is open source[20], free to use, and can be used as standalone software or on a server (e.g. for sharing). A DPIA is conducted by filling in free-form text or selecting one of specified options within the different form-like sections that relate to description of processing activities, and identifying risks and mitigations. The user can create and select 'templates' that contain pre-populated questions and guidance, and 'knowledge bases' that enable creating concepts for definitions, principles, risks, and mitigations. At the end of input, the tool provides an overview of risk scores based on entered information, and provides the ability for reviewing and approving (e.g. by a DPO).

The PIA tool provides import/export functionality using JSON for DPIA, templates, and knowledge bases. However, it is not documented in terms of structure and content,

[20] `https://github.com/LINCnil/pia/`

as well as how the tool interprets (or *parses*) the content and uses it within the layout. We investigated how our DPV-based DPIA information could be integrated or reused within this tool. This required reverse-engineering the import formats by experimenting with different data exports and analysing them. See Figure 3 for work in progress.

We are investigating the full extent of PIA's undocumented format and attempting to liaise with the developers on how to integrate RDF-based concepts within it. For this work, we used a script to convert and import DPV's concept using JSON. However, this removes the usefulness of DPV's semantics, e.g. identifying relevant risks associated with a parent concept. We hope to utilise (and advance) our DPIA specification so that it can be used within the PIA tool as a knowledge base, to describe various DPIA templates, and to provide consistent and interoperable access to exported information. From this, we also hope to investigate the capability of assisting stakeholders with automated forms of: risk discovery - in particular high-risk, suitable mitigations, and expressing impacts.

5. Discussion

Sufficiency of Concepts in DPV: Sufficiency as a criteria refers to the extent to which our concepts can represent information. The DPIA specification (including DPV) is sufficient to represent the information as specified in GDPR Art.35, but lacks representing concepts associated with other parts of the GDPR - in particular the principles in Art.5. This is because the focus of DPV has been on providing only a *conceptual vocabulary*, whereas tasks such as DPIAs require also *principles* and *controls* - both of which have specific meaning within law and industry practices. In addition, the DPA guidance clearly points to a need to represent organisational processes regarding governance and risk management in the same document as processing activities and GDPR compliance.

We therefore recommend undertaking an evaluation of what aspects of GDPR are currently represented within the DPV, and to prioritise inclusion of concepts such as principles which are important in legal investigations - such as DPIAs. A relevant resource in this is the Standard Data Protection Model (SDM) [24] produced by the German body of DPAs, which provides interpretations of the GDPR in the form of technical and organisational measures. That said, our approach as compared to the SotA definitively is novel, and extends the available methods for conducting and documenting DPIAs as machine-readable information that can be shared and reused. It provides the advantage of machine-readability for using the same information for multiple tasks e.g. to carry out DPIAs (this work) and ROPA - another obligation under GDPR Art.30 [25].

Knowledge Representation vs Practical Considerations: GDPR and DPIAs are a relatively new legal requirement. As a result, both DPAs and organisations are still understanding the intricacies, complexities, and requirements associated with it. We have only laid the groundwork for creating DPIA-related knowledge bases and tools, and there is abundant scope for enriching this work - such as adding more concepts from existing sources. At the same time, the work needs grounding and analysis of specific DPIA approaches to ensure that whatever knowledge is generated is of practical use and beneficial to stakeholders. Our experience with the three DPIAs and the use of the PIA tool shows that automation of processes such as DPIAs have a long road ahead.

We believe impact assessments such as DPIAs are an important aspect of accountability and responsibility, and that completely automating them disregards the intended

purpose, and creates false or incorrect notions of safety. Instead, we advocate technology (and technologists) should aim to assist rather than replace a human with related DPIA tasks. Therefore, in addition to adding concepts or using rules or similar mechanisms, DPIA-related approaches should also investigate their role and usefulness in conducting *actual DPIAs* to better understand the disparity between investigation and documentation, and to provide better solutions for capturing the human-generated inputs that can be used for enriching the underlying semantics in future updates. This requires time, financing, and domain expertise - which are difficult to obtain and efficiently utilise in smaller capacities. We therefore recommend undertaking this at larger avenues, such as national and EU frameworks and projects so that a culture of shared knowledge (based on use of semantics) can be established and exploited by public and private bodies alike.

Shared Impact Assessments: The lack of domain-specific knowledge regarding what is being investigated, who it affects, technologies involved, requirements of laws such as GDPR, and governance processes associated with risk management is a challenge in DPIAs. Our motivation was to address this through sharing and interoperability of information by using semantic web technologies. Through this, common shared resources for risks and impact management can be developed and shared for reuse. However, a DPIA is not the only impact assessment that concerns risks, mitigations, and fundamental rights and freedoms. The GDPR itself specifies similar assessments regarding data transfers and legitimate interests. In addition, future regulation proposed by the EU, in particular the AI Act [2] and Health Data Space[21], include impact assessment for high-risk as obligations. Such impact assessments have a large degree of commonality and overlap.

While researchers have investigated the overlap between DPIAs and the proposed AI impact assessment [26], there is no work to date that effectively shows how one can benefit from the other. Instead of developing separate and fragmented approaches for how these risk and impact assessments are carried out, documented, and investigated, a good solution would be to 'share' them as much as possible to reduce the burden on both organisations and auditors. In this, the shared information could relate to risks, mitigations, or categories of impacts, or even the structuring of information for reusing the same tools. This requires undertaking exercises similar to this one for other kinds of impact assessments, which has not been done within the state of the art, and to then identify avenues for shared impact assessments. We plan to undertake such an exercise for combining DPIAs with AI Act's impact assessments in the future.

6. Conclusion

Data Protection Impact Assessments (DPIAs), obligated by the EU General Data Protection Regulation (GDPR), are an important part of ensuring accountability and responsibility of personal data processing, and to identify and minimise harmful impacts to individuals regarding their fundamental rights. We presented the first step towards expressing DPIA and its relevant information as a machine-readable specification that can be used to document risks, mitigations, and their impacts in a formal manner and reused in information systems based on semantic web technologies. To better understand and explore how this work would be of practical use, we utilised three real-world com-

[21] `https://ec.europa.eu/health/ehealth-digital-health-and-care/`

plex DPIAs and identified limitations and important gaps within use of automation and human-involvement in DPIA investigations. Based on this, we have provided discussions on practicality and benefits of our approach in sharing information regarding risks and mitigations, and that this needs to incorporate human-generated information as an important aspect of DPIA documentations. In terms of future work, we have clearly identified concrete steps - such as enrichment of vocabularies based on available sources, and several promising directions - such as the creation of *shared impact assessments* based on commonalities between DPIA and EU's proposed AI Act.

Post-review Changes: We thank the reviewers for comprehensive and useful comments, and have incorporated them in this version. The changes made within DPV during the review period have also been incorporated, and the provided links have been edited to point to the resulting adoption of this work within DPV and DPV-GDPR. The original unedited article is available at `https://doi.org/10.5281/zenodo.6783204`.

Funding Acknowledgements: This work has been funded by Irish Research Council Government of Ireland Postdoctoral Fellowship Grant#GOIPD/2020/790. The ADAPT SFI Centre for Digital Media Technology is funded by Science Foundation Ireland through the SFI Research Centres Programme and is co-funded under the European Regional Development Fund (ERDF) through Grant#13/RC/2106_P2.

Thanks: We thank Rana Saniei for early discussions relevant to DPIAs, Delaram Golpayegani for discussions related to ISO risk management, and Georg P. Krog and other members of DPVCG for discussions on concepts.

References

[1] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation). Official Journal of the European Union. 2016 May;L119. Available from: `http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:L:2016:119:TOC`.

[2] Regulation Of The European Parliament And Of The Council Laying Down Harmonised Rules On Artificial Intelligence (Artificial Intelligence Act) And Amending Certain Union Legislative Acts. European Commission; 2021. Available from: `https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52021PC0206&from=EN`.

[3] Georgiadis G, Poels G. Towards a Privacy Impact Assessment Methodology to Support the Requirements of the General Data Protection Regulation in a Big Data Analytics Context: A Systematic Literature Review. Computer Law & Security Review. 2022 Apr;44:105640.

[4] Agrawal V. Towards the Ontology of ISO/IEC 27005:2011 Risk Management Standard. In: Proceedings of the Tenth International Symposium on Human Aspects of Information Security & Assurance (HAISA 2016); 2016. p. 11.

[5] Almeida R, Teixeira JM, Mira da Silva M, Faroleiro P. A Conceptual Model for Enterprise Risk Management. Journal of Enterprise Information Management. 2019 Sep;32(5):843-68.

[6] Rosa M, Guerreiro S, Pereira R. Designing an IT Risk Management Ontology Grounded on Systematic Literature Review. In: Hawaii International Conference on System Sciences; 2021. .

[7] Vicente P, Mira da Silva M. A Conceptual Model for Integrated Governance, Risk and Compliance. In: King R, editor. Advanced Information Systems Engineering (CAiSE). vol. 141. Cham: Springer International Publishing; 2011. p. 199-213.

[8] Haynes D. Understanding Personal Online Risk to Individuals Via Ontology Development. In: Lykke M, Svarre T, Skov M, Martínez-Ávila D, International Societey for Knowledge Organziation (ISKO), editors. Knowledge Organization at the Interface. Ergon; 2020. p. 171-80.

[9] Pandit HJ, Fatema K, O'Sullivan D, Lewis D. GDPRtEXT - GDPR as a Linked Data Resource. In: European Semantic Web Conference. LNCS. Springer, Cham; 2018. p. 481-95.

[10] Palmirani M, Martoni M, Rossi A, Bartolini C, Robaldo L. PrOnto: Privacy Ontology for Legal Compliance. In: Proceedings of the 18th European Conference on Digital Government (ECDG); 2018. p. 10. Available from: `http://hdl.handle.net/11576/2691050`.

[11] Pandit HJ, Polleres A, Bos B, Brennan R, Bruegger B, Ekaputra FJ, et al. Creating A Vocabulary for Data Privacy. In: The 18th International Conference on Ontologies, DataBases, and Applications of Semantics (ODBASE2019). Rhodes, Greece; 2019. p. 17.

[12] Dashti S, Sharif A, Carbone R, Ranise S. Automated Risk Assessment and What-if Analysis of OpenID Connect and OAuth 2.0 Deployments. In: Data and Applications Security and Privacy XXXV. Lecture Notes in Computer Science. Cham: Springer International Publishing; 2021. p. 325-37.

[13] Saniei R. Challenges in the Implementation of Privacy Enhancing Semantic Technologies (PESTs) Supporting GDPR. In: AI Approaches to the Complexity of Legal Systems XI-XII. vol. 13048. Cham: Springer International Publishing; 2021. p. 283-97.

[14] Saniei R. Data Protection Impact Assessment (DPIA) Vocabulary v0.1; 2021. Available from: `https://protect.oeg.fi.upm.es/def/gdpia/`.

[15] Esteves B, Rodriguez-Doncel V. Analysis of Ontologies and Policy Languages to Represent Information Flows in GDPR. Semantic Web J. 2022;Forthcoming.

[16] National Institute of Standards and Technology. Nist Privacy Framework:: A Tool For Improving Privacy Through Enterprise Risk Management, Version 1.0. Gaithersburg, MD: National Institute of Standards and Technology; 2020. NIST CSWP 01162020.

[17] van Dijk N, Gellert R, Rommetveit K. A Risk to a Right? Beyond Data Protection Risk Assessments. Computer Law & Security Review. 2016 Apr;32(2):286-306.

[18] Jakobi T, von Grafenstein M, Smieskol P, Stevens G. A Taxonomy of User-Perceived Privacy Risks to Foster Accountability of Data-Based Services. Journal of Responsible Technology. 2022 Jul;10:100029.

[19] Citron DK, Solove DJ. Privacy Harms. Rochester, NY: SSRN; 2021. 3782222.

[20] Wuyts K, Sion L, Joosen W. LINDDUN GO: A Lightweight Approach to Privacy Threat Modeling. In: 2020 IEEE European Symposium on Security and Privacy Workshops (EuroS PW); 2020. p. 302-9.

[21] DPIA for SURF and Dutch Government on Zoom. Privacy Company; 2022. Available from: `https://www.privacycompany.eu/blogpost-en/new-dpia-for-surf-and-dutch-government-on-zoom-all-high-risks-solved`.

[22] DPIA Office 365 for the Web and Mobile Office Apps. Privacy Company; 2020. Available from: `https://www.privacycompany.eu/blogpost-en/new-dpia-on-microsoft-office-and-windows-software-still-privacy-risks-remaining-short-blog`.

[23] DPIA Google G Suite Enterprise. Data Protection Authority Netherlands; 2021. Available from: `https://www.privacycompany.eu/blogpost-en/google-mitigates-8-high-privacy-risks-for-workspace-for-education`.

[24] The Standard Data Protection Model. Conference of the Independent Data Protection Supervisory Authorities of the Federation and the Länder; 2020. Available from: `https://www.datenschutz-mv.de/datenschutz/datenschutzmodell/`.

[25] Ryan P, Brennan R, Pandit HJ. DPCat: Specification for an Interoperable and Machine-Readable Data Processing Catalogue Based on GDPR. Information. 2022 May;13(5):244.

[26] Selbst AD. An Institutional View Of Algorithmic Impact Assessments. Harvard Journal of Law & Technology. 2021. Available from: `https://papers.ssrn.com/abstract=3867634`.