Further with Knowledge Graphs. M. Alam et al. (Eds.)
AKA Verlag and IOS Press, 2021

doi:10.3233/SSW210043

Building a Data Processing Activities Catalog: Representing Heterogeneous Compliance-Related Information for GDPR Using DCAT-AP and DPV

Paul RYAN[ab,1] and Harshvardhan PANDIT[c] and Rob BRENNAN[a]
[a] *ADAPT Centre, School of Computing, Dublin City University, Dublin 9, Ireland*
[b] *Uniphar PLC, Dublin 24, Ireland*
[c] *ADAPT Centre, Trinity College Dublin, Dublin 2, Ireland*

Abstract. This paper describes a new semantic metadata-based approach to describing and integrating diverse data processing activity descriptions gathered from heterogeneous organisational sources such as departments, divisions, and external processors. This information must be collated to assess and document GDPR legal compliance, such as creating a Register of Processing Activities (ROPA). Most GDPR knowledge graph research to date has focused on developing detailed compliance graphs. However, many organisations already have diverse data collection tools for documenting data processing activities, and this heterogeneity is likely to grow in the future. We provide a new approach extending the well-known DCAT-AP standard utilising the data privacy vocabulary (DPV) to express the concepts necessary to complete a ROPA. This approach enables data catalog implementations to merge and federate the metadata for a ROPA without requiring full alignment or merging all the underlying data sources. To show our approach's feasibility, we demonstrate a deployment use case and develop a prototype system based on diverse data processing records and a standard set of SPARQL queries for a Data Protection Officer preparing a ROPA to monitor compliance. Our catalog's key benefits are that it is a lightweight, metadata-level integration point with a low cost of compliance information integration, capable of representing processing activities from heterogeneous sources.

Keywords. Legal Compliance, Data Governance

1. Introduction

Organisations can be large and complex entities that perform heterogeneous processing on large volumes of diverse personal data. In practice, organisations often consist of (semi-)autonomous data processing units such as divisions, departments, or subsidiaries to achieve organisational goals. Organisations may also employ external parties like contractors, processors, or operational partners. This heterogeneity contrasts with existing LegalTech solutions for GDPR compliance that require the organisation to adhere to whatever data model is required by the solution [1].

From a legal perspective, administrative fines and actions are imposed on organisations as singular entities instead of individual units (GDPR Rec.150). Hence, the organisation is responsible for creating, maintaining, and demonstrating legal

[1] Corresponding Author, Paul Ryan, ADAPT Centre, School of Computing, Dublin City University, Dublin 9, Ireland; E-mail: paul.ryan76@mail.dcu.ie.

compliance information in its entirety. GDPR requires organisations to appoint a Data Protection Officer (DPO) to advise and assist them with compliance-related tasks. The DPO's challenge is to document all personal processing activities, which multiple parties carry out across the extended organisation. In practice, the DPO is the early warning indicator of adverse data processing activities within the organisation [2]. This challenging role requires the DPO to arduously document processing activities carried out by internal (e.g. departments) and external (e.g. contractors) units; and thereby establish, monitor, and advise the organisation on its compliance accordingly.

Processes can be intra-organisational involving internal departments or business functions, or inter-organisational where external parties are involved in the process. This information must be fed into the legal compliance 'graph' or 'product'. In practical terms, these 'sources' of data processing activities may evolve independently and have requirements and management methods that do not necessarily match the organisation's compliance processes.

As an example of the challenge, consider an organisation creating its Register of Processing Activities (ROPA), which is the first item requested by a regulator to investigate and must be produced on request (GDPR Art.30.4). The organisation must collect the information required for inclusion in the ROPA from potentially diverse sources such as business functions, departments, and affiliates. In practice, organisations rely on manual and informal methods such as spreadsheets, customised software, or internally developed systems to catalog their processing activities [1], which are then presented to the DPO in multiple heterogeneous forms by the various sources responsible for processing personal data. These practices result in organisations struggling to meet their ROPA obligations [1] and is an ongoing issue as inter and intra-organisational processes and their relevance in crafting the legal compliance documentation such as ROPA are yet to be resolved [3].

Our solution to this challenge is the development of DPCat. This is a profile of the well-established DCAT W3C standard for data catalog [4][5]. Our technical approach analyses the legal requirements to establish the data required to complete a ROPA. We develop DPCat, a profile of DCAT-AP [6], by supplementing it with terms from the Data Privacy Vocabulary (DPV) [7]. This solution will enable organisations to collect information under a standard form and offer a consolidated view of their processing activities. We will conduct a use case to evaluate our research goal to establish the extent that a Data Processing Activities Catalog based on DCAT-AP and Data Privacy Vocabulary (DPV)can overcome the heterogeneity of sources to facilitate a ROPA.

The structure of our paper describes the use case based on real-world examples in section 2. We describe our deployment scenario where an organisation that consists of multiple business functions and an outsourced processor holding data in many diverse heterogeneous sources is required to identify and record all personal data processing activities to meet its GDPR compliance obligations. In section 3, we evaluate the related work of the cataloguing of Data Processing activities. We identify that the development of vocabularies and ontologies in this domain, whilst prolific, would benefit from deploying a data processing catalog to collect unified metadata to be utilised for ROPA creation, particularly the ability to span graph-based and non-graph data sources. Section 4 proposes a data processing activities catalog for representing heterogeneous compliance-related Information for GDPR and identifying the key benefits of a data catalog. Section 5 presents the design of our proto-type Data Processing Activities Catalog system based upon DCAT-AP. We present the regulatory requirements of a ROPA and express these in RDF form based upon the Data Privacy Vocabulary (DPV).

We identify the key features that the Data Processing Catalog must contain to enable automatic ROPA generation. In the remainder of the paper, we implement our DCAT-AP-based catalog and evaluate our research goal to establish the extent that a Data Processing Activities Catalog based on DCAT-AP and DPV can overcome the heterogeneity of sources to facilitate the preparation of a ROPA. For the remainder of the paper, we evaluate how effective the data catalog performed to meet the research goal.

2. Use Case

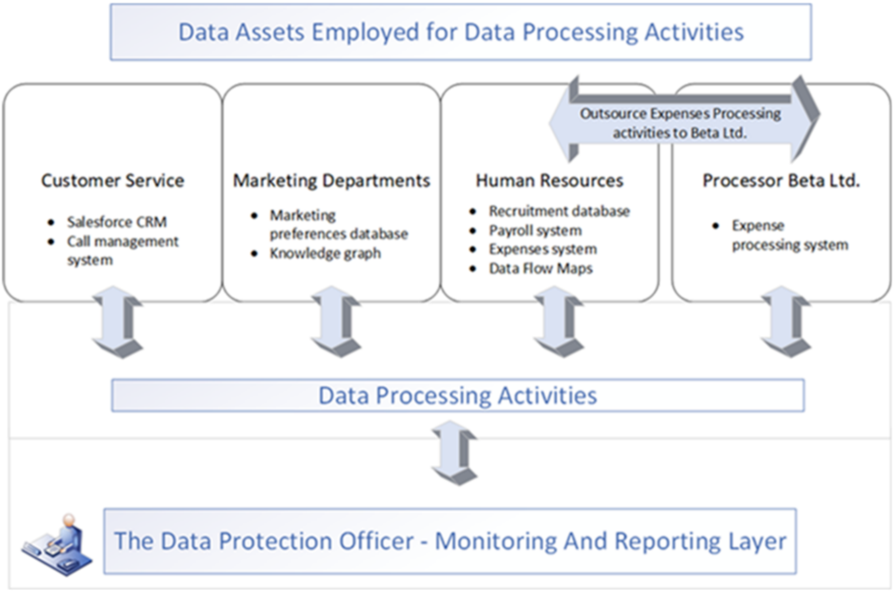
Data Assets Employed for Data Processing Activities
Outsource Expenses Processing activities to Beta Ltd.
Customer Service
Salesforce CRM
Call management system
Marketing Departments
Marketing preferences database
Knowledge graph
Human Resources
Recruitment database
Payroll system
Expenses system
Data Flow Maps
Processor Beta Ltd.
Expense processing system
Data Processing Activities
The Data Protection Officer - Monitoring And Reporting Layer

Fig. 1. Diverse Sources and Formats for Data Processing Activities in Organisations

Our use case scenario involves an organisation known as Alpha Ltd. The organisation comprises three distinct departments: Customer service, Human Resources and Marketing (see Fig.1). The departments are part of the same legal entity but carry out a variety of data processing activities. These departments collect and process different personal data according to their purposes. The tools and systems they use to manage information and processing can be distinct (see Fig.1), such as CRM systems, ERP systems, data flow models, semantic models, spreadsheets, etc. The distribution of platforms tends to reflect historical acquisitions by Alpha Ltd and local deployments by market segment leaders rather than homogeneous development of corporate IT systems, including data management or governance platforms.

Alpha Ltd has engaged the Data Processor Beta Ltd. to assist the HR department in processing employee expense claims. Beta Ltd carries out this processing activity in Canada, outside the European Union and is designated as providing appropriate safeguards for personal data transfers (GDPR Art.46.1). As a Data Controller, Alpha Ltd must ensure that all personal data processing activities are collected and recorded in its ROPA. To do this, the DPO, as a 'compliance officer' for Alpha, needs to liaise with each of the individual departments and request required information from them. In turn, the responsible departments must identify and extract this information from the information management systems used to track activities. As a result, the information about data processing activities within the organisation is presented to the DPO in heterogeneous forms. Further, the DPO must engage with relevant people or 'contacts' within each department in case of further information, clarity, or communication needs.

Hence, the requirements that a tool for creating a ROPA must deliver are:

1. Supports the heterogeneity of data sources describing data processing activities within an organisation
2. Enables standards-based collation of the data required for completion of a ROPA
3. Recording temporal validity of processing activities, e.g. active period
4. Supports periodic or continuous changes to data processing activity descriptions to reflect the dynamic lifecycle of data processing activities in an organisation
5. Records identity of activity host and organisational unit and relevant contact, e.g. to assist the DPO to collect additional information
6. Facilitate searching records, e.g. identify activities active on a specific date
7. Enable the creation of ROPA and other compliance-related documentation using information collected in the records
8. Minimises the data to be collected and integrated
9. Easy to deploy, e.g. based on established or commonly used software platforms

Next, we examine current systems' abilities to deliver these functionalities to DPO.

3. Related Work

We have established that organisations need to capture and express data processing activities carried out by their affiliates/ business functions and associated entities irrespective of the source data's heterogeneity. These processing activity descriptions need to be recorded and maintained in a ROPA. This section will review the extent to which the existing related work can meet the requirements set out in our use case. We will discuss the ability to exist commercial solutions [8], enterprise architecture and semantic-based solutions to meet the use case requirements.

Firstly, if we examine existing commercial solutions, we find a fragmented approach to recording processing activities to prepare ROPAs [1]. Organisations most commonly create and maintain ROPAs through informal tools, such as visual data flow mapping, customised in house software, and spreadsheets [1]. Data Protection Regulators encourage this practice by providing spreadsheet-based templates to help organisations prepare and maintain ROPAs [3]. A spreadsheet, while being a simple and commonly utilised versatile medium, requires effort to enter information and keep it updated. As a human-oriented application, spreadsheets often lack the rich data structures and semantics suitable for building automated toolchains, especially when modelling complex legal concepts beyond numerical or financial models. Furthermore, these approaches present challenges in that they are stand-alone and lack interoperability [3]. The maintained ROPA fails to meet the minimum threshold in many circumstances as they fail to be "sufficiently detailed for purpose" [9].

Enterprise Architecture (EA) models have offered the potential to generate a ROPA by augmenting existing EA models with the necessary information to maintain and generate a ROPA [10]. Huth et al. propose an approach where all required ROPA information is queried and presented in a structured format. The data in this structured form can be displayed in a custom-built application or exported to a ROPA presentation spreadsheet. However, the heterogeneity of data processing activities from diverse sources, both Inter and Intra organisational, creates challenges as the EA architecture may not extend to all the business units or domains required. In addition, specialised knowledge and tools are often not in-house, are required to build and extend EA models.

Many Semantic-based projects provide vocabularies, ontologies, and policy languages that can be used to represent GDPR concepts. These solutions mainly focus on providing informational items referenced in GDPR rights and obligations. They tend to focus on modelling/advanced use cases rather than deployment and interoperability. These projects focus on legal compliance evaluation. They do not consider the critical aspect of how the information required for (a) evaluating legal obligations and (b) demonstrating legal compliance - is maintained or generated within/by organisations and the entities involved in this process. The ability to demonstrate compliance is integral to the principle of accountability (GDPR Art.5.2). In many cases, many of the open-source ontologies and vocabularies are obsolete or without new developments in recent years, except for a small number of open vocabularies such as **BPR4GDPR's IMO** [11], **GDPRov** [12], **GConsent [**13], **DPV** [7], **GDPRtEXT** [14] and **PrOnto** [15] being the only ones that continue.

BPR4GDPR (Business Process Re-engineering and functional toolkit for GDPR compliance) [11] is a compliance ontology used to dictate and evaluate processes by considering them as workflows where actions or operations are connected dependencies and data flows performed by actors who can include assets or artefacts. Process mining is performed on the knowledge extracted from event logs of information systems to discover, monitor, and improve processes not assumed or modelled before evaluation. BPR4GDPR is utilised to create a process monitoring architecture. These rules are intended to act as constraints in conformance checking and repair the processes by identifying components that need to be changed to satisfy rules. GDPRov, [12] is a linked data ontology for expressing consent and data lifecycles' provenance to document user compliance. GConsent [13] is an OWL2-DL ontology representing consent and associated information, such as provenance. It uses R2RML to produce mappings for generating RDF metadata and focuses on using a standard model for each consent instance. This would also facilitate using data validation of information regarding consent. GDPRtEXT [14] is a linked data resource using the European Legislation Identifier (ELI) ontology for exposing the GDPR as linked data and is published using DCAT. The dataset contains a SPARQL endpoint.

GDPRtEXT also provides a SKOS vocabulary for defining terms and concepts in GDPR. The PrOnto [15] ontology provides concepts regarding GDPR associated with data types and documents, agents and roles, processing purposes, legal bases, processing operations, and deontic operations for modelling rights and duties. It has been applied for legal compliance checking over Business Process Model and Notation (BPMN). Though several vocabularies feature concepts for GDPR compliance, none of these has been utilised in modelling ROPA (through GDPR). We identify that the development of vocabularies and ontologies in this domain is certainly prolific but would benefit from deploying a data processing catalog to collect unified metadata to be utilised for ROPA creation, particularly the ability to span graph-based and non-graph data sources.

Currently, there are no vocabularies explicitly addressing or supporting ROPAs. Of the specified existing works, the DPV is the only one deployed to represent ROPAs [3]; however, this is a conceptual initiative with no deployment to date.

4. A Data Processing Activities Catalog

Alpha Ltd. can create a ROPA using existing solutions; however, the challenge for Alpha ltd. is to do this accurately and maintain an up to date ROPA [9]. Therefore, we propose a data processing activities catalog for representing heterogeneous compliance-related data for GDPR. The key benefits of a data catalog for this task are as follows:

- The design of data catalogs span heterogeneity based on common metadata and thus only require the collection of a small amount of data to describe the processing activities
- Data catalogs are widely used by industry, with many increasing numbers of organisations having expertise in their area
- Data catalogs such as CKAN [16] offer user interfaces that facilitate use by non-technical personnel
- Data catalogs support federated and distributed systems of data processing knowledge collection
- Data catalogs have specified standards for interoperability that we show below that can align with the data required for a ROPA
- Data catalog models and tools can be extended easily to gather additional data required for the completion of a specialised dataset such as a ROPA

We will base our data processing activities catalog on DCAT-AP. This profile specification is based on W3C's Data Catalog vocabulary (DCAT) for describing public sector datasets in the EU's Open Data portals. DCAT-AP enables cross-data portal search by harmonising the metadata collected and enables common metadata collection and search about diverse datasets. This is achieved by the exchange of standard descriptions of datasets among data portals. In addition, DCAT-AP proposes mandatory, recommended, or optional classes and properties to be used for a particular application; It identifies requirements to control vocabularies for this particular application; It gathers other elements to be considered as priorities or requirements for an application such as conformance statement, agent roles or cardinalities.

Our catalog will be known as DPCat. It will be a profile of DCAT-AP and will be focused on representing data processing activities for the generation of a ROPA. DPCat will build on the specifications of DCAT-AP to represent the processing activities required for ROPA. DPCat will also utilise the DPV as the controlled vocabulary used for the catalog. The terms required for ROPA are aligned to the DPV namespace and are a controlled vocabulary for the fields in the profile. The DPV is taxonomical modelling of concepts associated with personal data processing based on the GDPR. It is an outcome of the W3C Data Privacy Vocabularies and Controls Community Group (DPVCG), representing a community agreement between different stakeholders. The creation of the DPV ontology follows guidelines and methodologies deemed 'best practice' by the semantic web community [17]. The DPV is helpful as a machine-readable representation of personal data processing and can be adopted in relevant use-cases such as legal compliance documentation and evaluation, policy specification, consent representation and requests, a taxonomy of legal terms, and annotation of text and data. The use of DPV as part of DPCat will provide an extensive personal data processing vocabulary that will sufficiently expressively represent the terms required in ROPA.

5. DPCat Specifications

Our system requires the representation of the legal data required to complete the ROPA and operational information to maintain the ROPA on an ongoing basis. Article 30 of the GDPR sets out the legal information required to prepare the ROPA. In addition, the regulation states that each controller and, where applicable, the controller's representative, shall maintain a record of processing activities under its responsibility. That record shall contain all the following information:

(a) the name and contact details of the controller and, where applicable, the joint controller, the controller's representative, and the data protection officer
(b) the purposes of the processing
(c) description of the categories of data subjects and the categories of personal data
(d) the categories of recipients to whom the personal data have been or will be disclosed include recipients in third countries or international organisations.
(e) where applicable, transfers of personal data to a third country or an international organisation, including the identification of that third country or international organisation and, in the case of transfers referred to in the second subparagraph of Article 49(1), the documentation of suitable safeguards
(f) where possible, the envisaged time limits for erasure of the different categories of data
(g) A general description of the technical and organisational security measures referred to in Article 32(1) is possible.

In practice, many regulators provide ROPA templates that prescribe a format for the presentation of ROPA [3]. Whilst these templates are not mandatory; they are a minimum expectation of what is required by the regulator to demonstrate the organisation's accountability. For our use case, we will create a ROPA based upon the fields specified by regulation in Article 30 of the GDPR.

In section 4, we present DPCat as a solution to represent data processing activities. We have identified the data required for representation in the ROPA from Article 30 of the GDPR. To achieve this representation in DPCat, we identify the mandatory, recommended, and optional fields already specified in DCAT-AP and build on this, as DPCat is a profile of DCAT-AP. We find that we can utilise several DCAT properties to meet our requirement list's needs for a Processing Activities catalog as set out in section 2. We utilise the DPV to specify all additional properties that we require to populate ROPA. We document this specification for representing data processing activities using DPCat in table 1 with the following notation: M for Mandatory fields, C for Conditionally applicable, R for Recommended, and O for Optional. We provide a specification overview for the DPCat catalog in Figure 2.

Table 1. Specification for Representing the Data Processing Activities in DPCat

ROPA Requirement	Obligation	DPCat Property	DPCat Property Range
Controller	M	dct:publisher	foaf:Agent, dpv:Controller, adms:PublisherType
Purpose	M	dpv:hasPurpose	dpv:Purpose
Categories of Data Subjects	M	dpv:hasDataSubject	subclass of dpv:DataSubject
Categories of Personal Data	M	dpv:hasPersonalDataCategory	subclass of dpv:PersonalDataCategory
Categories of Recipients	C	dpv:hasRecipient	subclass of foaf:Agent, adms:PublisherType, dpv:LegalEntity
Data Transfer	C	dpv:hasProcessing	dpv:Transfer
Data Transfer Location	M	dpv:hasLocation	dpv:Location
Data Transfer Recipient	M	dpv:hasRecipient	foaf:Agent, adms:PublisherType, dpv:LegalEntity
Data Transfer Safeguards (see note below)	C	dpv:hasSafeguard	dpv:Safeguard
Time limits for erasure of different categories of data	R	dpv:hasDuration	dpv:StorageDuration
Technical and Organisational Measures	R	dpv:hasTechnicalOrganisationalMeasure	dpv:TechnicalOrganisationalMeasure
Processors responsible for processing	R	dpv:hasRecipient	dpv:Processor

Note: The Property dpv:hasSafeguard and the property range dpv:Safeguard have been submitted to the Data Privacy Community Controls Group for inclusion in the DPV vocabulary.

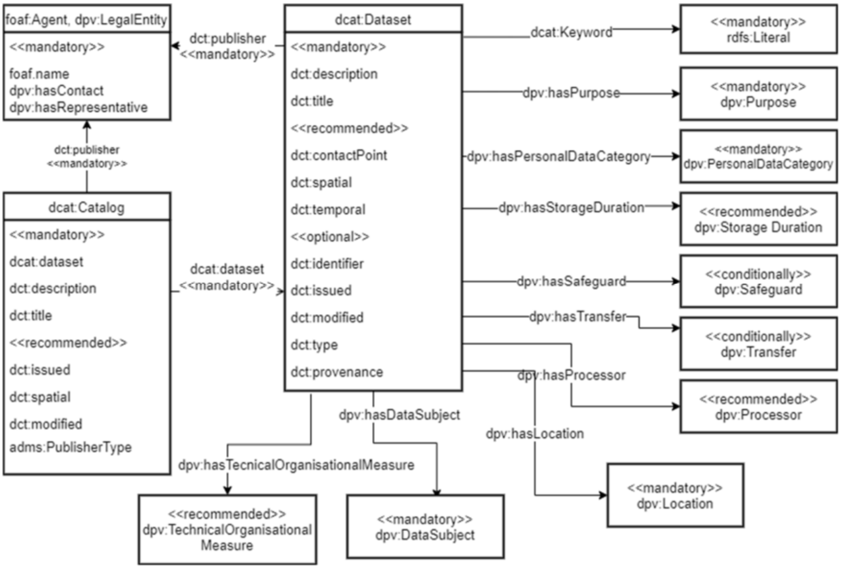

Fig. 2. DPCat specification for ROPA datasets

In section 2, we set out the requirements that a data processing catalog for ROPA must provide. We have proposed that our specialised data catalog DPCat can provide the DPO with a solution for representing a ROPA where data must be gathered from heterogeneous sources. In Table 2, we set out how DPCat can meet these requirements, and we support this with a demonstration of DPCat in section 6.

Table 2. How DPCat Meets our Requirements for a Data Processing

Req. no	Data Processing Catalog Requirement	DPCat Property
1.	Heterogeneity of data	dct:publisher ;dcat:dataset
2.	Enables standards-based collation of the data for ROPA	Refer to section 6 (Demonstration)
3.	Temporal information	dct:issued ; dct:temporal ; dct:modified
4.	Changes to the records	dct:modified ; dct:issued
5.	Identity of organisational unit	dct:publisher ; dct:contactPoint ; dpv:LegalEntity
6.	Facilitate searching records	dct:issued ; dct:temporal ; dct:identifier ; dct:modified
7.	Facilitate the creation of ROPA	Refer to section 6 (Demonstration)
8.	Minimises the data to be collected and integrated	
9.	Easy to deploy	

6. Demonstration and Discussion

To demonstrate the application of the catalog and evaluate its feasibility in addressing the requirements identified in Section 2, we created sample data reflecting the structure and operation of departments within the organisation Alpha Inc. and used queries to extract information to create ROPA. In our use-case scenario, the DPO must collect and inspect information from multiple departments for Marketing, Human Resources (HR) and Customer Services - each of which has its record-keeping practices. Also, the HR department employs the processor Beta Ltd. - which must also maintain its ROPA as a processor. The catalog, datasets, queries, and outputs for this use case are available here: https://github.com/coolharsh55/DPCat.

Each department maintains its records in our use case and has a separate catalog, while the organisation's catalog references these as datasets. The information maintained in a department's catalog and records fields are produced based on how they conduct their activities. The outcome is an RDF graph used in the catalog records. SPARQL queries were then used to create 'views' for presenting a summary and overview of activities—for example, and Table 3 specifies a snippet of processing activities in terms of information required for ROPA, their temporal periods, and the contact point for further communication with the complete ROPA available in the DPCat repository mentioned above.

Table 3. Sample Extract of Controller ROPA

Department	Customer Service Dept.	HR Dept.	Marketing Dept.
Title	Record001	Record004	Record001
Period Start	2019-01-01	2019-01-01	2019-01-01
Period End	2022-12-13	2022-12-13	2022-12-13
Contact Name	Alice	Bob	Emily
Contact e-mail	alice@example.com	bob@example.com	emily@example.com
Purpose Category	Customer care	Service Provision	Direct Marketing
Purpose	Recording of customer calls	Expenses activities	Direct marketing via e-mail
Data Subject	Customers	Employees	Customers
Personal Data Category	Voice recordings	Financial	E-mail addresses
Recipient	Null	Beta Ltd.	Null
Recipient Category	Null	Data Processor	Null
Recipient Location	Null	Canada	Null
Storage years	2.0	7.0	1.0
Measures	Standard	Standard	Standard

We used GraphDB Free [18] [2] as a triple-store to store and query the information. In the queries, we relied on utilising reasoning and inferences capabilities in GraphDB (RDFS and OWL2) to retrieve results where triples were not explicitly specified correctly. We initially opted to utilise separate named graphs for each department's information to represent independent maintenance with SPARQL CONSTRUCT queries to ingest them into a global organisation-level graph. However, we discovered that this approach creates SPARQL queries due to the requirements that each named graph be explicitly specified in the query. Therefore, we decided to use a single organisation-level graph where each department maintains its catalog for demonstration purposes. We comment on this in our discussion on practicality later in the paper.

Our approach also strived to create each dataset record as a self-contained graph since the information maintained represents a 'snapshot' of activities for that organisation or its unit in a specific temporal period. This process involved using blank nodes and owl:sameAs to related entities within the organisation's global graph. This also helped validate the dataset on its own by using SHACL to check that mandatory fields are present and the correctness of the information. This approach has further benefits by making documentation and validation possible at any arbitrary stage - from individual records and organisational units to the entire organisation without conflicts or dependencies. Thus, the ROPA queries could target a specific catalog, department, or the entire organisation.

In addressing the requirements specified in section 2, the use-case sufficiently demonstrates that catalogs are a good design paradigm for record-keeping connected with GDPR compliance and ROPA documentation. The approach enables documenting data processing activities in terms of their temporal period, limiting the scope to organisational units, and assigning contact points within the organisation for further information. The inherent design of catalogs as a 'collection of records' permits the responsible unit to continue updating and maintaining records while reducing the burden on DPOs by utilising the catalog itself as a single point of reference for all related information. The use of SPARQL facilitates information searching, filtering, and exporting for ROPA creation. The paper's contribution is that the organisation can span heterogeneity based on common metadata requiring the collection of only a small amount of data to describe the processing activities. The organisation can thus generate, maintain and query a ROPA efficiently by relying on the common metadata-based records provided by DPCat to aggregate and homogenise access within the diverse sources of information required for compliance.

6.1. Discussion on Practicality and Avenues for future research

Automation. In terms of functioning and integration with existing organisational tools, the creation of datasets and records in RDF can be automated using approaches such as R2RML - which is a standardised specification for mappings from relational/SQL databases to RDF, or using data cataloging tools such as CKAN provides tools for catalog creation and maintenance. More importantly, the catalog is a DCAT-AP profile based on the standardised DCAT vocabulary and is itself a standard maintained by the EU to provide interoperability for sharing data between its data portals.

[2] https://www.ontotext.com/products/graphdb/graphdb-free/

Data sources. As we mentioned earlier in this section, we discovered the complexity of querying information when departments utilised individual named graphs for housing catalog records. In practical terms, whether each department should independently maintain compliance-related information or only submit it to a single monolithic repository is based on the organisation's practices. However, for interoperability, this information needs to be present somewhere in the catalog. We, therefore, intend on further exploring the suitability of existing fields within DCAT-AP and the more recent developments in DCAT v2 to represent information regarding sources, data formats, access controls, and SPARQL endpoints. This can also allow the specification to facilitate appropriate tooling and programmatic interfaces that can actively search and accommodate other heterogeneous tools and data sources.

Controlled Vocabularies. Currently, the specification uses DPV as a pseudo-controlled vocabulary to ensure information is expressed using the same concepts as those required for a ROPA (or broadly for GDPR compliance). Utilising a different vocabulary to specify the fields (such as purpose or recipient) is possible but requires changing the catalog specification in its entirety. Furthermore, any vocabulary chosen cannot foresee all possible concepts owing to the reality of how purposes and personal data categories can be defined. However, DPV, by being a 'community-driven standard', provides stability and interoperability in addition to expressing taxonomies from a top-down approach which makes it possible to extend and customise to situations. Therefore, it is recommended that other controlled vocabularies, where they are needed and used, be aligned to DPV concepts to ensure continued interoperability of the catalog information.

Representing complexity, e.g. Catalog of Catalogs. The use-case demonstrates functionality for a dataset catalog, which is more straightforward to understand due to its smaller scope and size. However, practical requirements may dictate many records and organisational units represented within the catalog's catalog. For the specification and tooling to function correctly in such situations, it is essential to formalise how such catalogs should be defined and the resulting interpretation.

Shared Information. The use-case considers complete dissociation between organisational departments, which may not be the case in practice. For example, the IT department may be responsible for ensuring appropriate technical and organisational measures are implemented, or a Controller may wish to record what measures a Processor has in place. In this case, organisations may want to delegate or import some catalog information from specific units. It is not currently possible to denote this with the outlined specification. We, therefore, specify this as an open research question regarding how to represent and maintain heterogeneous information within a catalog.

Common registries. The specification for a catalog of data processing activities provides an exciting possibility where a data portal can be set up for representing associated information. This can have several use-cases ranging from an open-source catalog of an organisation's practices and policies to enabling communication between controllers and/or processors. Another practical application of the specification is that it enables authorities to request and manage information about data processing activities through a dedicated data portal. This is promising given the drive for digital services and inter-jurisdictional information sharing for compliance within the EU.

Conclusions

The heterogeneity of data sources representing the organisation's data processing activities presents significant challenges when completing a ROPA. Our research sought to establish the extent to which implementing a Data Processing Activities Catalog based on DCAT-AP and DPV can overcome the heterogeneity of sources to facilitate the preparation of a ROPA. For this, we presented a use case and developed a prototype system to catalog the organisation's diverse data processing activities using SPARQL queries to output a ROPA document. Its key benefits are providing a lightweight, low cost, and metadata-level integration for compliance information regarding processing activities from heterogeneous sources. In addition, our DPCat solution advances alignments between disciplinary and domain-specific metadata standards. Finally, it enables data catalog implementations by providing a common interoperable base for ROPA without requiring full alignment or merging all the underlying data sources.

Acknowledgements

The ADAPT Centre supports this work for Digital Content Technology, funded under the SFI Research Centres Programme (Grant 13/RC/2106_P2) and co-funded under the European Regional Development Fund. Uniphar PLC supports Paul Ryan. Harshvardhan J. Pandit is funded under the Irish Research Council Government of Ireland Postdoctoral Fellowship Grant#GOIPD/2020/790 and the European Union's Horizon 2020 research and innovation programme under NGI TRUST Grant#825618 for Project#3.40 Privacy-as-Expected: Consent Gateway. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission

References

[1] International Association of Privacy Professionals (IAPP), Trust Arc.: Measuring Privacy Operations. (2019).
[2] Drewer, D., Miladinova, V.: The Canary in the Data Mine. Computer Law and Security Review 34, 806-815 (2018).
[3] Ryan, P., Pandit, H., Brennan, R.: A Common Semantic Model of the GDPR Register of Processing Activities (2020), doi:10.3233/FAIA200876.
[4] Profiles Ontology Homepage, http://www.w3.org/TR/dx-prof/ , last accessed 2021/04/11.
[5] DCAT Homepage, http://www.w3.org/TR/vocab-dcat-2/ , last accessed 2021/04/11.
[6] DCAT-AP Homepage, https://data.gov.ie/dataset/dcat-ap , last accessed 2021/04/11.
[7] DPV Homepage,https://w3.org/ns/dpv, last accessed 2021/04/11
[8] International Association of Privacy Professionals (IAPP),: 2020 Privacy Tech Vendor Report. (2020).
[9] Castlebridge Report (2020), https://castlebridge.ie/research/2020/ropa-report/ , last accessed 2021/04/11.
[10] Huth, D., Tanakol, A., Matthes, F.: Using Enterprise Architecture Models for Creating the Record of Processing Activities. IEEE 23rd International Enterprise Distributed Object Computing Conference (EDOC), 98-104 (2019) DOI: 10.1109/EDOC.2019.00021.
[11] BPR4GDPR Homepage, http://www.bpr4gdpr.eu/ , last accessed 2021/04/11.
[12] Pandit, H.J., Lewis, D.: Modelling provenance for gdpr compliance using linked open data vocabularies. (2017).

[13] Pandit, H.J., et al.: GConsent - A Consent Ontology based on the GDPR, Lecture Notes in Computer Science, Vol. 11530, 270-282, (2019)

[14] Pandit, H.J., et al.: GDPRtEXT - GDPR as a Linked Data Resource, The semantic web— 15th international conference, ESWC (2018), Notes in Computer Science, vol 10843, 481–495, (2018) https://doi.org/10.1007/978-3-319-93417-4_31

[15] 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)

[16] CKAN Homepage, https://ckan.org/ , last accessed 2021/04/11.

[17] Pandit, H.J., et al.: Creating a Vocabulary for Data Privacy. In: Panetto H., Debruyne C., Hepp M., Lewis D., Ardagna C., Meersman R. (eds) On the Move to Meaningful Internet Systems: OTM, (2019).

[18] GraphDB Homepage, https://www.ontotext.com/products/graphdb/graphdb-free/ , last accessed 2021/04/11.