Enhancing Data Use Ontology (DUO) for Health-Data Sharing by Extending it with ODRL and DPV

Harshvardhan J. Pandit [a,*] and Beatriz Esteves [b,**,***]
[a] *ADAPT Centre, Trinity College Dublin, Ireland*
E-mail: pandith@tcd.ie
[b] *Ontology Engineering Group, Universidad Politécnica de Madrid, Spain*
E-mail: beatriz.gesteves@upm.es

Abstract. The Global Alliance for Genomics and Health is an international consortium that is developing the Data Use Ontology (DUO) as a standard providing machine-readable codes for automation in data discovery and responsible sharing of genomics data. DUO concepts, which are OWL classes, only contain textual descriptions regarding the conditions for data use they represent, which limits their usefulness in automated systems. We present use of the Open Digital Rights Language (ODRL) to make these conditions explicit as rules, and combine them to create policies that can be attached to datasets, and used to identify compatibility with a data request. To associate the use of DUO and the ODRL policies with concepts relevant to privacy and data protection law, we use the Data Privacy Vocabulary (DPV). Through this, we show how policies can be declared in a jurisdiction-agnostic manner, and extended as needed for specific laws like the GDPR. Our work acknowledges the socio-technical importance of DUO, and therefore is intended to be complimentary to it rather than a replacement. To assist in the improvement of DUO, we provide ODRL rules for all of its concepts, an implementation of the matching algorithm, and a demonstration showing it in practice. All resources described in this article are available at: *https://w3id.org/duodrl/repo*.

Keywords: health data, biomedical ontologies, policy, regulatory compliance, GDPR

1. Introduction

1.1. Background & Motivation

The sharing of health-related data holds great promise for enhancing research and applying advanced computational and statistical techniques for progress in medicine. At the same time, such sharing and use of health-related data is required to be regulated at legal and institutional levels given its sensitive nature and the ability to have significant impacts. The current landscape consists of institutions such as hospitals assessing each data use request through a dedicated committee that is responsible for the evaluation and decision making regarding the release of data under their custody. To assist in this process, the Global Alliance for Genomics and Health[1] (GA4GH) was formed as

*Corresponding Author E-mails: pandith@tcd.ie, beatriz.gesteves@upm.es.
**Both authors have contributed equally to this work

[1] https://www.ga4gh.org/

an international consortium for developing standards and responsibly sharing genomics data. Of its various initiatives addressing different components and processes involved in data sharing, it has developed a machine-readable ontology called Data Use Ontology[2] (DUO) [1, 2] for expressing "Data Use Limitations" (DUL) – conditions and constraints expressed by data providers and adhered by requestors.

DUO is an OWL ontology based on (and part of) Open Biological and Biomedical Ontology[3] (OBO). Through the use of OBO upper ontologies and guidelines, DUO offers (semantic) interoperability with a variety of biomedical ontologies part of the OBO family. The intended use of DUO is to annotate datasets with DUL codes to indicate usage conditions, expressing data use requests, and identifying or discovering compatible datasets automatically by comparing the request with dataset specific DULs. More information about DUO is provided in Section.2.1.

DUO concepts specify the DULs as human-readable text within their description (using `obo:IAO_0000115` relation), which restricts their usefulness to humans or explicitly encoded systems that can only function on known concepts. In addition, DUO concepts are not linked to relevant legal concepts, which creates confusion and ambiguity as to the implications of using these in a system or jurisdiction such as the EU where the General Data Protection Regulation (GDPR) [3] introduces additional accountability and compliance requirements which must be identified and applied. The existing documentation notes that the applicability of laws is the responsibility of the adopter, and that DUO terms have not been considered for implications under the the GDPR. Compatibility with existing regulations is important for future endeavours, especially since the EU envisions a 'Health Data Space'[4] where machine-readability and automation will play an important role.

1.2. Research Objectives and Contributions

Our argument is that *true machine-readability* requires the information intended to be conveyed through DUO concepts about the specific permissions, prohibitions, constraints, requirements, and so on to be (also) represented as machine-readable *rules* that utilise semantic concepts. With this, the DULs inherent in the descriptions of each DUO concept are made explicit through formal representation as a set of *rules* that can be attached and used alongside the data as a *sticky policy*.

For assessing whether a data use request is compatible with the dataset DULs, both dataset and request conditions are expressed as policies, and are compared to evaluate whether the dataset policy's rules are satisfied by the request policy. While DUO is already being used in this manner, such as within the Data Use Oversight System[5] (DUOS), this is done by checking hierarchical compatibility between request concepts and data use conditions through subclass relations between concepts. This approach is limited in ability and expressiveness for specifying rules and their use in automated systems.

More importantly, it does not take advantage of existing research and functional solutions for expressing specifics of rules and policies, for checking their conformance and compliance with legal requirements, and the ability to have mathematical guarantees regarding correctness and consistency. Such solutions have existed for a while now - for example Answer Set Programming (ASP), or the use of logic-based semantic reasoners, and have been utilised in a variety of domains - including for legal compliance with the GDPR (see Section.2.2).

With the above motivation, we present an approach for representing the inherent rules in DUO concepts explicitly in RDF through use of the Open Digital Rights Language[6] (ODRL) [4]. We specifically chose ODRL because: (i) It is a standard developed explicitly to model rules and policies; (ii) It uses RDF and is machine-readable; (iii) Its terms are aligned with legal vocabularies; and (iv) Its use can be limited to declaration of information (e.g. what rules apply) or also for their validations and conformance checking (i.e. using an ODRL validator).

In addition to these, we also consider ODRL the most suitable candidate for representing DUO concepts as it can be used without requiring any of the existing DUO-based data use or request governance processes to make radical and incompatible changes. That is, the existing practices and processes by which DUO codes are added as

[2] http://purl.obolibrary.org/obo/duo

[3] https://obofoundry.org/ The prefix `obo` has the IRI `http://purl.obolibrary.org/obo/`

[4] https://ec.europa.eu/health/ehealth-digital-health-and-care/european-health-data-space_en

[5] https://duos.broadinstitute.org/

[6] https://www.w3.org/TR/odrl-vocab/ The prefix `odrl` has the IRI `http://www.w3.org/ns/odrl/2/`

annotations to datasets and are used to request access to them can continue without hindrance, and can choose which aspects of our ODRL solution they want to adopt within their practices.

ODRL, by modelling terms regarding rights and licensing, also offers a compatible segue for DUO to be linked with relevant legal concepts. We further enhance these through use of the Data Privacy Vocabulary[7] (DPV) [5], an output of the W3C Data Privacy Vocabularies and Controls Community Group[8][9] (DPVCG). DPV provides an extensive vocabulary of concepts, can be expanded or specialised for jurisdictional requirements, provides legal bases and rights - including from GDPR, is open and accessible, and can be easily integrated into DUO's use-cases.

The contributions of this work are summarised through the following research objectives:

RO1 Specifying DUO concepts and conditions for data use as machine-readable policies using ODRL
RO2 Developing an algorithm for consolidating data use conditions into a single ODRL policy
RO3 Developing an algorithm for identifying compatible datasets with data use requests based on ODRL policies
RO4 Enabling expression of legal concepts and restrictions with(in) ODRL policies for DUO concepts using DPV
RO5 Elucidating relevance of DUO concepts and associated ODRL+DPV policies for GDPR obligations

The rest of this article presents: an overview of DUO and its applications in Section.2.1, relevant work in state of the art regarding machine-readable policies for GDPR in Section.2.2, our use of ODRL to represent DUO concepts and perform matching with requests in Section.3, expression of legal concepts using DPV in Section.4, a demonstration through proof-of-concept in Section.5, a discussion on integrating this work into existing DUO-based workflows in Section.6, and concluding statements in Section.7.

2. Relevant Work and State of the Art

2.1. Data Use Ontology (DUO) and Aligned Efforts

DUO concepts are structured across three taxonomies. The *Data Use Permission* taxonomy, with base class `obo:DUO_0000001`, represents permissions for purposes regarding data use. The *Data Use Modifier* taxonomy, with base class `obo:DUO_0000017`, represents 'modifiers' or conditions to be applied in addition to permissions. The *Investigation* taxonomy, with base class `obo:OBI_0000066` from the Ontology for Biomedical Investigations[10] (OBI), represents 'investigations' or planned processes for which the data is requested for use. Along with these, the concept `obo:DUO_0000010` represents the relation *is_restricted_to* which is used to restrict or scope specific concepts to some context, for example with domain as `obo:DUO_0000022` representing limitation on use within a geographic region, and range as `obo:GAZ_00000448` from the Gazetteer[11] (places) ontology.

DUO is the result of earlier efforts to create codes regarding data use, and use them as machine-readable information towards automation. The first iteration was based on Consent Codes [6] which provided concepts representing permission to use data. The second iteration adopted some terms from the Automatable Discovery and Access Matrix[12] (ADA-M) [7] framework which has similar aims and concepts. The use of DUO as intended towards collection of consent for dataset sharing and reuse is specified in the 'Machine-readable Consent Guidance'[13]. A brief outline and summary of DUO and its use to streamline access to biomedical datasets is presented in [1], and a list of GA4GH initiatives and standards along with relevance of DUO within those is presented in [2].

The Data Use Oversight System[14] (DUOS) is a platform based on DUO that provides semi-automated data access management for use of datasets. It uses DUO annotations for adding new datasets and in data access requests, which are then matched using an algorithm based on hierarchical compatibility i.e. permitted conditions identified based on

[7]https://w3id.org/dpv The prefix `dpv` has the IRI `https://w3id.org/dpv#`
[8]https://www.w3.org/community/dpvcg/
[9]Note: Harshvardhan currently chairs the DPVCG, and Beatriz is a member. Both have contributed to DPV's development.
[10]http://obi-ontology.org/
[11]https://environmentontology.github.io/gaz/
[12]https://www.ga4gh.org/genomic-data-toolkit/regulatory-ethics-toolkit/
[13]https://www.ga4gh.org/wp-content/uploads/Machine-readable-Consent-Guidance_6JUL2020-1.pdf
[14]https://duos.broadinstitute.org/

establishing subclass relations between request and dataset DUO codes. The output of the matching process is then used as part of a review by a 'Data Access Committee' (DAC). An evaluation of DUOS's automation process found it to be comparable to human data access committees [8]. DUOS is currently being implemented in an ongoing large scale pilot [9].

Other uses of DUO include specification of informed consent for health and genomics research in Africa [10], along with ADA-M for representing consent for health data sharing in a blockchain [11], and in CTRL [12] - an online platform that uses DUO to provide dynamic consent interfaces and tools for large-scale genomics research programs. Potential uses of DUO are described in the Data Tags Suite (DATS) [13] where DUO is a candidate vocabulary in its framework for discovering data access based on metadata, and as part of a roadmap for accessing 1 million human genomes across EU infrastructures [14]. We found only one article that provided a machine-readable metadata representation of information using DUO - which used SWRL[15] to express the rules [15]. Further overview of DUO and its relevant approaches amongst other rights and licensing initiatives, approaches, and tools for health data sharing is provided by Grabus and Greenberg [16].

Of note in these identified articles and other resources is that we did not find a clear example or workflow for how the machine-readability of DUO should be associated with datasets, expressed as part of a request, or how the matching algorithm should function. The article presenting DATS [13] also refers to this difficulty in establishing the permissions and prohibitions when using DUO, and mentions ODRL as an alternative model providing clearer expression of permissions and prohibitions. The DUOS framework offers the best (available) description for how DUO can be applied, but does not offer much guidance on how the matching is performed between datasets and requests annotated with DUO concepts. From these, we establish the necessity of providing *RO1*, *RO2*, and *RO3*.

2.2. Expression of Machine-readable Information and Policies for GDPR

Given that health data may be regarded as personal data, it is subject to regulations such as the GDPR as well as other domain and sector specific laws such as Health Insurance Portability and Accountability Act[16] (HIPAA). By also annotating datasets with machine-readable metadata that relates to such laws, automation can also be used to assist stakeholders in identifying and meetings their compliance requirements [17].

In this, the state of the art consists of substantial research and development in modelling and using legal ontologies (see survey by Rodrigues et al. [18]). Of note regarding the matching of DUO dataset annotations is the policy checking algorithm for GDPR developed by SPECIAL H2020 project [19] which offers a fast matching algorithm based on subsumption between OWL2 concepts with logical consistency and correctness guarantees. In principle, this is similar to DUOS's matching algorithm where the concepts to be matched in a policy are pre-determined.

While ODRL, being a standard for expressing policies, provides concepts with legal interpretation (e.g. *Asset* or *Party*), it deviates from or does not contain terms such as *Controller* or *Legal Basis* which carry important obligations under regulations such as the GDPR. Vos et al. address this by extending ODRL as a 'Regulatory Compliance Profile' which is used for expressing policies associated with GDPR [20]. In this, the relevant concepts in ODRL are extended with those from GDPR to construct ODRL rules reflecting GDPR's compliance requirements. In approaches providing a vocabulary for use regarding GDPR, GDPRtEXT [21] provides vocabulary of concepts, and GConsent [22] provides an OWL2 modelling of consent information. While these approaches are illuminating in how to describe GDPR's requirements, their use would restrict the created policies to be operationally limited for application under GDPR.

In contrast to these, the Data Privacy Vocabulary (DPV) [5] provides a taxonomy of concepts which can be used as jurisdiction-agnostic terms with an extension for specific concepts from GDPR[17] such as its legal bases and rights. DPV is, to our knowledge, the most comprehensive vocabulary for modelling concepts associated with privacy and data protection laws. It also offers different semantic serialisations (i.e. SKOS, RDFS, OWL) which facilitate integration into use-cases.

---

[15] https://www.w3.org/Submission/SWRL/

[16] https://www.govinfo.gov/link/plaw/104/public/191?link-type=html

[17] DPV-GDPR: GDPR Extension for DPV https://w3id.org/dpv/dpv-gdpr

Two recent surveys provide an overview of existing efforts that have utilised semantic web technologies to address GDPR compliance. The first, by Kurteva et al. [23], describes the approaches associated with consent, and the second, by Esteves and Rodrigues-Doncel [24], analyses ontologies and policy languages for modelling information flows. Both highlight the variety of approaches available, and offer opinionated suggestions regarding use of ODRL and DPV - which we have incorporated in our choice of implementations[18]. Based on on these identified works and existing surveys, we chose ODRL and DPV to create jurisdiction-agnostic policies that can be specialised for GDPR, thus addressing *RO4* and *RO5*.

3. Rewriting DUO using ODRL

As presented in Section.2.1, DUO concepts are structured across three taxonomies with textual description of the DULs they represent. The goals of this work, in terms of research objective *RO1* is to analyse this implicit information and express it explicitly using ODRL, with the additional goal of keeping compatibility with existing uses and workflows that use DUO so as to not cause large disruptions to GA4GH's current and future activities.

We consider DUO's primary attractiveness to be the ease with which its concepts can be easily constructed from input mechanisms (such as a form) and simply 'tagged' onto a dataset as an annotation. In this, the textual clauses used to describe the concepts are based on well-defined clauses from consent forms[13]. The role of ODRL, therefore, is not to replace DUO, but to provide additional machine-readable information for each DUO concept that provides explicit conditions currently inherent in the textual clauses i.e. as rules that can be checked, verified, and consumed in an automated manner.

To do this, we first analysed DUO concepts and performed a manual alignment of the textual information with ODRL concepts. We then constructed rules expressing identified conditions and expressed them using ODRL. We then constructed a matching algorithm that utilised ODRL policies to compare a request policy with a dataset's policy to determine compatibility.

3.1. Identifying ODRL equivalents for DUO concepts

For each concept in DUO, we first sought to identify the constraints or conditions by interpreting the textual description and identifying whether it related to a permission, prohibition, or obligation, and the specific context of how those are to be applied. In doing this, we observed duplicity and overlap between DUO's data use permissions and modifiers as both contained *purpose-based conditions* without a clear distinction between their semantics and interpretation, and regarding permission or prohibition of that purpose as an indication of *consent*. For example, `DUO_0000011` represents permission and `DUO_0000044` represents prohibition for "population origins or ancestry research", with the former being a data user permission and the latter a data use modifier.

We suggest restructuring the taxonomies in DUO to address this by considering a single purpose-based taxonomy specifying research concepts that either have variants for permission and prohibition (i.e. two distinct concepts), or to explicitly provide a data use modifier concept representing permission or prohibition that is applied over a specified research purpose. This is based on DUOS's data collection input forms and ADA-M's concepts where each research purpose can be individually consented (or restricted) to, with possible implications arising from lack of any permission or prohibition.

After analysing DUO's concepts and identifying inherent conditions, we formulated the relevant ODRL rules for expressing those conditions. Where this was not possible because of ODRL lacking the required concept, we created proposed extensions of its concepts to enable rule expressions. For each concept, we constructed an `odrl:Set` instance representing the specific rules (see Section.3.2), and consolidated these rules into an `odrl:Offer` representing a collective singular policy for a dataset (see Section.3.3). A complete collection of the interpretations made for each DUO concept is presented in Table 1.

[18]It would be prudent to point out that while both authors of this paper are also authors on the cited surveys, the justification offered here is that these prior efforts provide clear evidence on the strengths of choices made in our implementations.

Table 1
Interpretation of conditions inherent in DUO concept descriptions as ODRL rules

Concept	Code	Rule Type	Constraint	Placeholder
DUO0000001	**Data Use Permission**			
DUO0000042	GRU	Permission	Purpose is :GRU	
DUO0000006	HMB	Permission	Purpose is :HMB	
DUO0000006	HMB	Prohibition	Purpose is not :POA	
DUO0000007	DS	Permission	Purpose is :DS and mondo:0000001	:TemplateDisease
DUO0000004	NRES	Permission	Purpose is odrl:Purpose	
DUO0000011	POA	Permission	Purpose is :POA	
DUO0000011	POA	Prohibition	Purpose is not :POA	
DUO0000017	**Data Use Modified**			
DUO0000043	CC	Permission	Purpose is :CC	
DUO0000020	COL	Duty	Action :COL	
DUO0000021	IRB	Duty	Action :IRB	
DUO0000016	GSO	Permission	Purpose is :GeneticStudies	
DUO0000016	GSO	Permission	Purpose is :GeneticStudies and :PhenotypeResearch	
DUO0000016	GSO	Prohibition	Purpose is :PhenotypeResearch and not :GeneticStudies	
DUO0000022	GS	Permission	Spatial is equal to specified :Location	:TemplateLocation
DUO0000022	GS	Permission	Spatial is not equal to specified :Location	:TemplateLocation
DUO0000028	IS	Permission	Assignee is :ApprovedInstitution	:TemplateInstitution
DUO0000028	IS	Prohibition	Assignee is not :ApprovedInstitution	:TemplateInstitution
DUO0000015	NMDS	Prohibition	Purpose is :MDS	
DUO0000018	NPUNCU	Permission	Assignee is :NotForProfitOrg and Purpose is :NotForProfit	
DUO0000018	NPUNCU	Prohibition	Assignee not :NotForProfitOrg or Purpose not :NotForProfit	
DUO0000046	NCU	Permission	Purpose is :NotForProfit	
DUO0000046	NCU	Prohibition	Purpose is not :NotForProfit	
DUO0000045	NPU	Permission	Assignee is :NotForProfitOrg	
DUO0000045	NPU	Prohibition	Assignee is not :NotForProfitOrg	
DUO0000044	NPOA	Prohibition	Purpose is not :POA	
DUO0000027	PS	Permission	Project is :ApprovedProject	:TemplateProject
DUO0000027	PS	Prohibition	Project is not :ApprovedProject	:TemplateProject
DUO0000024	MOR	Duty	Action odrl:distribute :ResultsOfStudies with odrl:dateTime	:TemplateDateTime
DUO0000019	PUB	Duty	Action odrl:distribute for :ResultsOfStudies	
DUO0000012	RS	Permission	Purpose is specified :Research	:TemplateResearch
DUO0000012	RS	Prohibition	Purpose is not specified :Research	:TemplateResearch
DUO0000029	RTN	Duty	Action :ReturnDerivedOrEnrichedData	
DUO0000025	TS	Permission	Time is less than specified :TemplateDateTime	:TemplateDateTime
DUO0000026	US	Permission	Assignee type is :ApprovedUser	:TemplateUser
DUO0000026	US	Prohibition	Assignee type is not :ApprovedUser	:TemplateUser
OBI0000066	**Data Use Permission**			
DUO0000034		Permission	Purpose is :AgeCategoryResearch for specified age categories	:TemplateAgeCategory
DUO0000033		Permission	Purpose is :POA	
DUO0000037		Permission	Purpose is :HMB	
DUO0000040		Permission	Purpose is :DS	
DUO0000039		Permission	Purpose is :DrugDevelopment	
DUO0000038		Permission	Purpose is :GeneticStudies	
DUO0000035		Permission	Purpose is :GenderCategoryResearch for specified categories	:TemplateGender
DUO0000031		Permission	Purpose is :MDS	
DUO0000032		Permission	Purpose is :PopulationResearch for specified population	:TemplatePopulation
DUO0000036		Permission	Purpose is :ResearchControl	

We faced challenges in interpreting specific phrases such as "*is limited to*" which imply that usage is permitted within and only within that specific scope. If this interpretation is correct, then DUO should clarify how potential conflicts should be resolved, for example between rules expressing exclusive limitations and other permissive expressions (e.g. "*is allowed for*"). Our suggestion is to take advantage of ODRL's ability to express these rules as code through which it can identify when a given collection of rules associated with a single dataset are contradictory or impossible to satisfy, e.g. by checking the satisfiability of a policy against itself.

Currently, DUO concepts are limited to representing conditions for data use, with suggestions referring to external ontologies for additional concepts required for expressing scope or restrictions. For example, `DUO_0000007` represents permission for disease-specific research, with recommendation to use MONDO ontology[19] for specifying diseases. Other specific concepts mentioned in the textual descriptions but not modelled explicitly include codes inherited from predecessors, such as *CC* for Clinical Care Use, or *GRU* for General Research Use. Expressing ODRL rules requires these concepts to be explicitly defined e.g. as *Disease* for the disease-specific research, upon which permissions or prohibitions are then expressed.

For our implementation, we identified and collected such 'missing terms' into an ad-hoc vocabulary to permit ODRL rules to be expressed correctly for each DUO concept. We recommend DUO to adopt these or to create a similar vocabulary for explicitly providing the concepts and their descriptions separate from the data use conditions in which they are used. This also has the added advantage of providing better documentation of information represented by those concepts. For e.g. by modelling *IRB* as a concept representing Ethics Review Board approval, it is possible to add information about what processes and requirements are needed in such reviews. It also permits further rules pertaining to ethics approvals to be semantically associated with a base concept, e.g. to indicate it must be carried out prior to data use, or periodically, or before publishing any outcomes.

For data use requests (specified as *investigations* in DUO), we again found duplicity with concepts in data use permission and data modifiers. For example, `DUO_0000040` represents a request and `DUO_0000007` represents a permission for research for specific diseases. Semantically, both refer to the same concept regarding 'research for specific diseases', with the distinction of one being a request and the other being a permission. Similar to the earlier suggestion on reorganisation of DUO's taxonomies to be based on research purposes, we also recommend applying the same approach for consistency in concepts used for requesting use of data. Doing so permits clarity, reduces disambiguity, and assists in matching as the same concept would be associated with a dataset using `odrl:Offer` and a request using `odrl:Request` (see Section. 3.4).

Apart from the expression of conditions for data use and requests to use that data, DUO concepts also have applications in *recording* the outcomes of matching processes where access has been granted. This is an important and yet unexplored area in the currently identified uses of DUO, especially since any sharing of data would be expected to be accompanied with information about the entities involved, provenance associated with grant process, and details regarding how the conditions have been met at the time or later in the future. We present how ODRL is useful in representing this information as instances of `odrl:Agreement` (see Section. 3.5) which can contain all the above information, and also be used in automated approaches that can periodically check if the pending conditions for an agreement have been met, e.g. fulfilment of publishing results.

3.2. Data use restrictions as `odrl:Set`

Each DUO restriction is represented as an instance of `odrl:Set`, which must contain at least one permission, prohibition, or duty, and one resource (here a dataset) to be a valid ODRL policy. Its use does not grant any access or privileges, and only represents a collection or *set* or rules over the resource.

Interpreting the textual descriptions accompanying each DUO concept, we used `odrl:permission` when the condition granted access to data, `odrl:prohibition` when it denied access, and `odrl:duty` when it specified obligations to be fulfilled. We included DUO's textual descriptions using `rdfs:comment` for convenience, and indicated association with the DUO concept using `dct:source`.

It was challenging for us to construct a valid policy which required specifying the resource (dataset), because DUO concepts only represent abstract conditions that don't relate to a specific dataset. To ensure ODRL

[19] https://obofoundry.org/ontology/mondo

policies are always valid, and to clearly indicate how to later apply or *instantiate* them for a dataset, we created the class `TemplateQuery` representing a placeholder that is replaced with the value(s) retrieved by executing a SPARQL query associated with it through `sparqlExpression`. In Table.1 these are indicated as *Placeholder*. To apply queries over the specific context, the variable `?dataset` is replaced with the IRI of the dataset whose policy is being generated, and `?this` is replaced with the IRI of placeholder instance. For associating datasets, the instance `TemplateDataset` with the SPARQL query `SELECT ?dataset WHERE { ?dataset ?o ?p }` is used. We also defined placeholders for other restrictions, for example the instance `TemplateDisease` for use in disease-specific restrictions with the SPARQL query `SELECT ?disease WHERE { ?this obo:DUO_0000010 ?disease }`.

Another challenge we faced was for indication of scoped restrictions e.g. specifying the location when use is limited to a geographic location. DUO contains the property `obo:DUO_0000010` that describes the relation `is_restricted_to` which we interpret as intending to be used to specify such scopes in restrictions. However, DUO concept descriptions only state "this should be coupled with an ontology term describing the (concept) the restriction applies to", and we could not find an example showing how it should be used in this manner. Further, ODRL requires all constraints to be specified directly over the *Asset* (i.e. dataset). Therefore even if this property were available, its use would complicate the expression of rules in ODRL.

We discussed possible solutions to this, and identified four potential avenues: (i) use of OWL class expressions[20]; (ii) use of SHACL shapes to indicate a constraint; (iii) creating a new ODRL mechanism that takes property paths as *Operand*; and (iv) declaring the concept directly as an instance of the scoping concept (e.g. for disease-specific restriction, the concept would be an instance of the appropriate DUO class as well as the disease class). Each of these have a bearing on how a condition is expressed, and on the performance and capability of matching processes for comparing two policies. For example, use of (i) would require executing an OWL2 reasoner prior to the matching process, and (ii) would require a SHACL validator. In our implementation, we used (iv) by declaring the concept as an instance of both DUO and scoping classes as it was the simplest method, did not require any additional tools or changes to ODRL, and could be replaced trivially with a different method in the future. However, we explicitly indicate this issue as requiring further investigation. The `odrl:Set` defined to represent DUO's concept on "population origins or ancestry research only" is presented in Listing 1.

3.3. Dataset policies as `odrl:Offer`

When using DUO concepts to annotate datasets, each dataset can contain multiple DUO concepts that must be interpreted in combination as an offer for using that dataset. This is expressed in ODRL as an instance of `odrl:Offer` containing the union of all `odrl:Set` instances associated with DUO concepts for a given dataset. In doing this, the `Offer` represents a single policy for that dataset that can be used in matching requests, or embedded as metadata to form a *sticky policy*. When creating offers, each individual rule retrieved from the merged set policies is maintained (as an individual rule) to facilitate the matching process with rules from data use requests. This also facilitates potential annotations for rules, such as specifying their provenance or adding additional information for their interpretation within that offer. An example `odrl:Offer`, which merges `:DUO_0000042`, `:DUO_0000025` and `:DUO_0000020`, is presented in Listing 2.

The construction of the `odrl:Offer` instance uses the following algorithm:

1. For a given dataset, retrieve all DUO data use permissions and modifier concepts it was tagged with.
2. For each DUO concept retrieved, fetch its relevant `odrl:Set` policy by using the `dct:source` association.
3. If a retrieved policy uses an instance of a `:TemplateQuery`, execute its associated SPARQL query, and replace the instance with retrieved value(s).
4. Create an instance of `odrl:Offer` containing all extracted rules[21].
5. Add provenance information or other additional documentation, e.g. `dct:dateSubmitted` for when the dataset was added to a system.

[20] A short and informative summary provided by Protégé https://protegeproject.github.io/protege/class-expression-syntax/

[21] Note: each rule is still associated with DUO concepts using `dct:source` to indicate which concepts are being used in the policy

```
:DUO_0000011 a odrl:Set ;
    rdfs:label "DUO_0000011" ;
    rdfs:comment "This data use permission indicates that use of the data is limited to the
    ↪ study of population origins or ancestry (POA - population origins or ancestry research
    ↪ only)" ;
    dct:source obo:DUO_0000011 ;
    odrl:permission [
        odrl:action odrl:use ;
        odrl:target :TemplateDataset ;
        odrl:constraint [
            odrl:leftOperand odrl:purpose ;
            odrl:operator odrl:isA ;
            odrl:rightOperand :POA ] ] ;
    odrl:prohibition [
        odrl:action odrl:use ;
        odrl:target :TemplateDataset ;
        odrl:constraint [
            odrl:leftOperand odrl:purpose ;
            odrl:operator :isNotA ;
            odrl:rightOperand :POA ] ] .
```

Listing 1: An `odrl:Set` representing `DUO_0000011` regarding Population Origins or Ancestry research (POA). The permission and prohibition over the same purpose is based on interpretation of the phrase "*is limited to*" to indicate use if permitted only for that research

```
:Offer a odrl:Offer ;
    rdfs:label "Offer to use dataset for GRU within time limits" ;
    odrl:target <https://example.com/Dataset> ;
    odrl:action odrl:use ;
    dct:source :DUO_0000042, :DUO_0000025, :DUO_0000020 ;
    dct:dateSubmitted "2022-04-30"^^xsd:date ;
    odrl:permission [
        odrl:duty [ odrl:action :CollaborateWithStudyPI ] ] ;
    odrl:permission [
        odrl:constraint [
            odrl:leftOperand odrl:elapsedTime ;
            odrl:operator odrl:lteq ;
            odrl:rightOperand "2022-12-31"^^xsd:date ] ] ;
    odrl:permission [
        odrl:constraint [
            odrl:leftOperand odrl:purpose ;
            odrl:operator odrl:isA ;
            odrl:rightOperand :GRU ] ] .
```

Listing 2: An example `odrl:Offer` containing a permission for general research use, from `DUO_0000042`, a time limit on the use, from `DUO_0000025`, and a duty to collaborate with the studies' primary investigator, defined from `DUO_0000020`.

3.4. Data use requests as `odrl:Request`

To represent data use requests, termed as investigations within DUO, instances of `odrl:Request` are used along with permissions for specific research purposes. In this, the DUO concepts representing requests for use are defined as instances of `odrl:Set`, similar to Section.3.2, and are combined together to create a single request, similar to Section.3.3. A request for genetic studies (`:DUO_0000038`), and the respective `odrl:Set` which was used to generate it, is presented in Listing 3.

```
:DUO_0000038 a odrl:Set ;
    rdfs:label "DUO_0000038" ;
    rdfs:comment "Request for biomedical research concerning genetics (i.e., the study of
    ↪ genes, genetic variations and heredity)" ;
    dct:source obo:DUO_0000038 ;
    odrl:permission [
        odrl:action odrl:use ;
        odrl:target :TemplateDataset ;
        odrl:assignee :TemplateAssignee ;
        odrl:constraint [
            odrl:leftOperand odrl:purpose ;
            odrl:operator odrl:isA ;
            odrl:rightOperand :GS ] ] .
:Request_for_GS a odrl:Request ;
    rdfs:label "A request for GS (DUO_0000038)" ;
    rdfs:comment "Request for biomedical research concerning genetics (i.e., the study of
    ↪ genes, genetic variations and heredity)" ;
    dct:source :DUO_0000038 ;
    dct:dateSubmitted "2022-05-01"^^xsd:date ;
    odrl:permission [
        odrl:action odrl:use ;
        odrl:target :TemplateDataset ;
        odrl:assignee <https://example.com/SomeRequestor> ;
        odrl:constraint [
            odrl:leftOperand odrl:purpose ;
            odrl:operator odrl:isA ;
            odrl:rightOperand :GS ] ] .
```

Listing 3: An `odrl:Set` and `odrl:Request` containing a request for the purpose of genetic research created from `DUO_0000038`.

3.5. Data use decisions as `odrl:Agreement`

Instances of `odrl:Agreement` are recorded outcomes of matching processes where access to the data has been granted. In this, the ODRL terms assist in specifying who has granted the access (`odrl:Assigner`, to whom (`odrl:Assignee`, for what (`odrl:Asset`), and the conditions over it (`odrl:Rule`). The rules mentioned in an agreement are the same specific rules and obligations as that specified for a dataset (i.e. from `odrl:Offer`) and in a request (i.e. `odrl:Request`). Through these rules, an agreement references the specific DUO concepts part of the agreement. An example representation of a data use decision as an `odrl:Agreement` between a data depositor and a data requester for the purpose of genetic studies is presented in Listing 4.

```
:Agreement a odrl:Agreement ;
    dct:dateAccepted "2022-05-31"^^xsd:date ;
    odrl:permission [
        odrl:action odrl:use ;
        odrl:target <https://example.com/Dataset> ;
        odrl:assignee <https://example.com/SomeDepositor> ;
        odrl:assigner <https://example.com/SomeRequestor> ;
        odrl:constraint [
            odrl:leftOperand odrl:purpose ;
            odrl:operator odrl:isA ;
            odrl:rightOperand :GS ] ] .
```

Listing 4: An `odrl:Agreement` representing a decision for use of a dataset

Note that ODRL defines `odrl:Agreement` as the granting or acknowledgement of a rule between the parties. This definition is agnostic to the contents of that agreement, which means that the agreement could be a permission granting access to a dataset, or one that prohibits or denies it. While the above example considered use of agreements only in the cases where access was granted, this definition makes it clear that they can also be used to record instances where the request was denied.

3.6. Matching algorithm using ODRL for identifying compatible datasets for a request

The matching algorithm in DUO is based on comparing and identifying compatibility between a dataset's data use conditions with data use requests. In our ODRL implementation, this is done by comparing the dataset's *odrl:Offer* with an *odrl:Request*.

Given two sets of concepts representing an offer and a request, the matching algorithm can utilise two different and incompatible notions for how access is determined. The first, which is the more common semantic interpretation, is based on considering classes as sets and determining access based on set membership. For a class P and its subclass C, a request for accessing P would also permit use of C since a member of C is always a member of P. But a request for C would not permit use of P as not all members of P are members of C. This approach has been used in matching policies for GDPR compliance [19] and for granting access to resources in Solid [24].

The second approach, which is what DUO describes in its documentation, is based on identifying applicability of a concept based on its specificity. For a class P and its subclass C, a request for accessing P would not grant access to C since it is more specific, but a request for accessing C would grant use of P as it is less specific. Using subsumption as a criteria, the first approach grants access when the data policy subsumes the request policy, whereas the second approach grants access when the request policy subsumes the data policy. Thus, both of the former mentioned approaches (i.e. [19] and [24]) can be reused here by *reversing* the direction of subsumption.

Another consideration for the matching algorithm is the resolution of permissions and prohibitions in terms of their order of evaluation and conflicts. It is possible to interpret a policy in several incompatible ways, such as first checking for permissions and granting access at the first satisfied permission i.e. a permissive model, and its opposite where prohibitions are first checked and access is denied for first unsatisfied prohibition i.e. a prohibitive model. When a conflict occurs for a permission and a prohibition over the same resource, the resolution would be based on the precedence of one over the other. In DUO, the matching algorithm is prohibitive since prohibitions take precedence over permissions. This means that if a request either does not satisfy a permission or satisfies a prohibition, use is denied. The policies are considered compatible only when all permissions are satisfied and all prohibitions remain unsatisfied.

Based on these considerations, our matching algorithm consists of checking for subsumption or satisfiability between `odrl:Offer` and `odrl:Request` instances. We adapted it from a prior implementation that also utilised ODRL in a matching algorithm for granting access [25]. The algorithm simply checks whether the dataset policy conditions are satisfied by the request policy in case of permission, or violated in case of prohibition. If any prohibitions are found, access is rejected. If no prohibitions are found, and all permissions are satisfied, then access is granted i.e. the term *access* indicates an outcome where policies match - no *actual access* is provided.

4. Expressing Legal Compliance Concepts using DPV

The DUO concepts and terms used are different from those as used in legal compliance tasks. By using ODRL concepts, the terms involved are expressed in a language that has legal interpretation (e.g. *Asset* or *Party*). The ODRL vocabulary also contains additional terms which may be used with DUO for specific legal interpretations, such as *ConsentingParty*, *InformedParty*, and *obtainConsent*. While these terms are sufficient for a policy to have legal interpretations, they are insufficient to incorporate the specifics of laws such as GDPR which assign specific roles to parties and require use of specific legal basis in processing of data. At the same time, if the terms are made specific only for a single law such as the GDPR, the usefulness and applicability of the resulting policies would be restricted to only that law without a clear recourse for adopting other laws and jurisdictions. To address this gap, we

utilised the Data Privacy Vocabulary (DPV) which provides terms that are intended to be jurisdiction-agnostic and can be used without being restricted to a specific law.

To utilise DPV, we first performed an alignment between its concepts and ODRL where DPV concepts that have an overlap with ODRL concepts are defined as their subclasses (e.g. `dpv:Entity` and `odrl:Party`). This utilised the approach from existing work regarding extending ODRL concepts for GDPR [20]. Where DPV concepts had no direct equivalent in ODRL, such as for legal basis, we used them directly within ODRL rules as instances of the relevant concepts (e.g. `dpv:hasLegalBasis` as `odrl:LeftOperand`). Table 2 describes the performed alignment[22] between ODRL and DPV concepts to define DUO concepts.

Table 2
Alignment between DPV and ODRL for use in policies expressing DUO concepts

DPV Concept	ODRL Concept	Relationship
dpv:Entity	odrl:Party	subclass
dpv:Purpose	odrl:Purpose	subclass
dpv:Processing	odrl:Action	subclass
dpv:PersonalData	odrl:Asset	subclass
dpv:LegalAgreement	odrl:Policy	subclass
dpv:hasTechnicalOrganisationalMeasure	odrl:LeftOperand	instance
dpv:hasLocation	odrl:LeftOperand	instance
dpv:hasJurisdiction	odrl:LeftOperand	instance
dpv:hasApplicableLaw	odrl:LeftOperand	instance
dpv:hasLegalBasis	odrl:LeftOperand	instance
dpv:hasRecipient	odrl:LeftOperand	instance
dpv:hasRight	odrl:LeftOperand	instance
dpv:hasRisk	odrl:LeftOperand	instance

Using DPV enabled modelling rules regarding restrictions on legal basis (e.g. consent), explicit acknowledgement of roles (e.g. data controllers), limitations on third-party recipients, and indicating applicability of a specific law using `dpv:hasApplicableLaw`. The DPV's "technical and organisational measures", which consist of concepts such as data security and impact assessments, can be used to further enrich DUO's data use modifiers and create a clear delineation between research purposes, measures required, and limitations or conditions of use.

To explicitly specify GDPR as the applicable law and utilise its legal bases and rights, we utilised the DPV-GDPR[23] extension which provides these concepts. Through this separation (between DPV and DPV-GDPR), the policies can be declared in an jurisdiction-agnostic manner using DPV, and made specific to a law such as the GDPR by checking additional contextual information such as the locations of patients whose data is involved, or that of the requesting party. The separation also provides a clear path for applying other jurisdictional laws and concepts on top of DPV by creating extensions of its concepts similar to DPV-GDPR. Listing 5 includes two ODRL offer policies that use DPV and DPV-GDPR to invoke jurisdiction-agnostic data protection and GDPR-specific terms, respectively.

DUO states the interpretation and applicability of GDPR's requirements is the responsibility of the adopter. This follows from the complexities of determining their applicability before any request is known, or because of the differences between stakeholder jurisdictions. To assist with this process, we recommend adding or providing relevant methods that are necessary to identify the applicability of the GDPR (or other laws). For example, GDPR is applicable (to simplify the condition) when an organisation operates within the EU or processes the personal data of people in EU. This translates to knowing the locations of people whose data is being offered for use as well as the requesting entity location.

[22] We intentionally restricted the alignment to only concepts required for using DUO so as to not introduce additional external interpretations.
[23] https://w3id.org/dpv/dpv-gdpr

```
@prefix dpv: <https://w3id.org/dpv#>
@prefix dpv-gdpr: <https://w3id.org/dpv/dpv-gdpr#>
:Offer1 a odrl:Offer ;
    rdfs:label "Offer to use dataset using Consent, and requiring an Impact Assessment" ;
    odrl:target <https://example.com/Dataset> ;
    odrl:action dpv:Use ;
    odrl:permission [
        odrl:constraint [
            odrl:leftOperand dpv:hasLegalBasis ;
            odrl:operator odrl:isA ;
            odrl:rightOperand dpv:Consent ] ] ;
    odrl:permission [
        odrl:constraint [
            odrl:leftOperand dpv:hasOrganisationalMeasure ;
            odrl:operator odrl:isA ;
            odrl:rightOperand dpv:ImpactAssessment ] ] ;
:Offer2 a odrl:Offer ;
    rdfs:label "Offer to use dataset using GDPR's Explicit Consent, and requiring a DPIA" ;
    odrl:target <https://example.com/Dataset> ;
    odrl:action dpv:Use ;
    dpv:hasApplicableLaw dpv-geo:GDPR ;
    odrl:permission [
        odrl:constraint [
            odrl:leftOperand dpv:hasLegalBasis ;
            odrl:operator odrl:isA ;
            odrl:rightOperand dpv-gdpr:A6-1-a-explicit-consent ] ] ;
    odrl:permission [
        odrl:constraint [
            odrl:leftOperand dpv:hasOrganisationalMeasure ;
            odrl:operator odrl:isA ;
            odrl:rightOperand dpv:DPIA ] ] ;
```

Listing 5: Two `odrl:Offer` policy instances that use DPV concepts to indicate conditions of use. `Offer1` is jurisdiction-agnostic and requires use of consent and an Impact Assessment. `Offer2` is GDPR-specific, and requires use of Explicit Consent and DPIA.

Using DPV, both of these can be expressed using the appropriate Entity concepts and *dpv:hasLocation*. This enables creating further data use limitations such as data being available only when the request acknowledges the applicability of the GDPR, or permitting use only within GDPR-governed jurisdictions, and checking this condition when matching a request with a dataset. The DPV-LEGAL[24] extension providing Jurisdictions, Laws, and Authorities for DPV is helpful in representing these conditions.

5. Demonstration and Evaluation Using a Proof-of-Concept

In this section, we describe the implementation of a User Interface to generate dataset policies and a prototype implementation of the matching algorithm is available at https://w3id.org/duodrl/demo/.

Figure 1 shows two examples of the developed UI to edit `odrl:Offer` policies, which relies on both ODRL and the ad-hoc vocabulary created to cover missing terms[25]. The first example (a) uses only the DUO concepts, and the second example (b) includes both DUO and DPV to construct `odrl:Offer`.

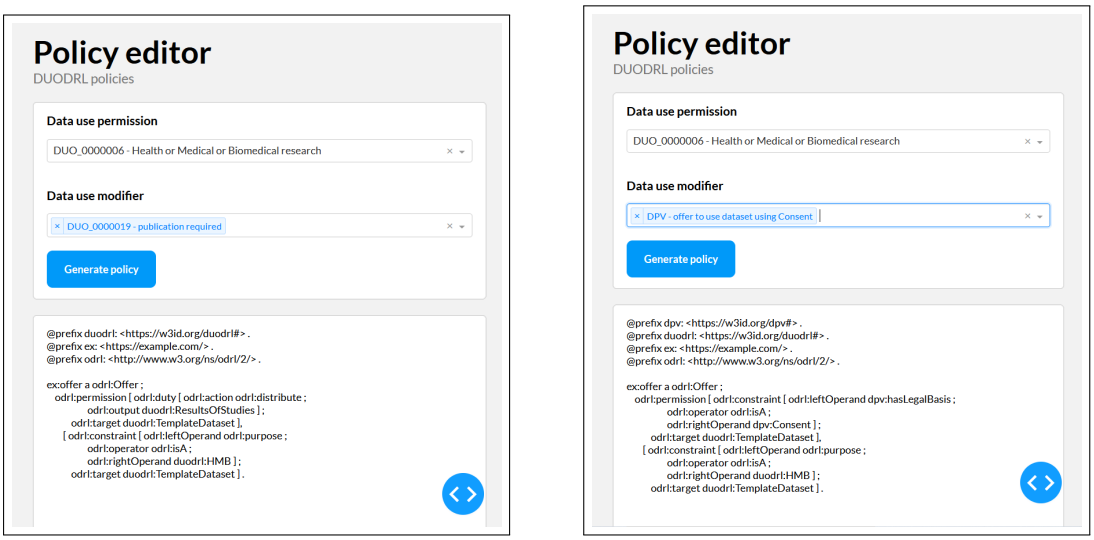
Upon selecting the relevant DUO concept in the UI, the application retrieves the associated `odrl:Set` instance representing data use permissions and data use modifiers as ODRL policies, combines them, and displays them

[24] https://w3id.org/dpv/dpv-legal

[25] Ad-hoc vocabulary available at https://w3id.org/duodrl.

on screen. The code and data used in this is available online (link in abstract). The matching algorithm used here was adapted from prior work in utilising ODRL for GDPR-based policy matching [25]. We modified it as per the requirements elicited in Section.3.

Fig. 1. Proof-of-concept implementation showing generation of `odrl:Offer` policies



(a) from DUO concepts

(b) from DUO and DPV concepts

For the matching process, the conditions represented in an `odrl:Request` instance should be a subset of those specified in the `odrl:Offer` instance associated with a dataset. This means the permissions and prohibitions from the offer instance should be satisfiable by the one from request. Once this is determined to be valid, the policies are considered compatible and access can be authorised. For data discovery, a request policy must be compared with policies of every dataset. This process can be made faster and convenient through pre-computations and optimisations - though we did not do these in our implementation as it is intended to only be a proof-of-concept.

The data discovery algorithm starts by checking if there is a specific rule within a dataset's policy for the purposes stated in the `odrl:Request` – if a permission is found for a purpose *P* then access to the dataset can be granted and if a prohibition is found then access is rejected. A similar exercise is then performed to check for additional restrictions related to other constraints (described on Table 1), e.g. restrictions on the type of assignee of the offer, or on the location or time of data use, and in case a prohibition is found then access to the dataset is denied and in case a permission is found access can be granted.

In the event additional duties are imposed for dataset use, such as agreeing to collaborate with the primary study investigator or providing documentation of ethical approval, these are included in the `odrl:Agreement` that establishes the final conditions for dataset use. If there are conflicting policies, resulting from the merging of different DUO permissions and modifiers, by default, the prohibition takes precedence, similar to the default behaviour of the algorithm for the case where no permission or prohibition is specified for a particular purpose. In such cases, access is denied.

6. Discussion on Integration into Existing DUO-based Workflows

DUO represents one facet of GA4GH's ambition to facilitate responsible genomics data sharing for health and medicine related research. It plays an important part given that its role is to increase automation in data discovery and assist in ensuring data use is permitted with accountability and oversight. Its use is thus part of a workflow consisting of different components, processes, and stakeholders who have differing requirements for how they use

DUO. Any changes proposed to the way in which DUO is modelled, is applied for dataset discovery, or is used in automation for identifying compatibility with requests may have consequences on these existing workflows. While better design and performance are valid technological goals, they should be evaluated within the lens of socio-technical applications they are a part of. This section therefore discusses the influence and impact of our work on existing DUO-based workflows and offers suggestions on how this work can be best utilised.

6.1. Design of DUO concepts

As we outlined in Section.3.1, the concepts within DUO have duplicity in semantics, and do not present the conditions they represent as explicit machine-readable code. This has an impact on the ability to use these for expression of policies and the implementation of automation in dataset discovery and request matching processes. In addition, the structuring of concepts requires clarity on their intended role without overlap (i.e. permissions, modifiers, and investigations), and should have separation of concerns (i.e. purposes from modifiers). Through this, the use of concepts becomes clearer and consistent, and provides the ability to introduce additional conditions and constraints without impact on existing concepts. We recommend following the ODRL model and concepts in terms of representing rules (permission, prohibition, duty), and constraints (purposes, scopes) separately from one another.

For further refinement of DUO terms and their interpretation, the textual descriptions provided should utilise controlled natural language (see survey on [26] for variety of approaches) that match the expression of rules (as in ODRL) so as to provide a reduced level of ambiguity and high-degree of specificity in the terms used. Through these, the descriptions can be made self-sufficient in terms of describing how they should be applied, or when (i.e. before or after data has been released), which can benefit the non-technical processes and stakeholders in understanding and using them. In addition, the specificity of descriptions will also assist approaches such as ours in constructing machine-readable rules that match the exact intention of that concept.

By specifying policies in ODRL (or other similar policy-based semantic models), the use of DUO gains additional automation potential where policies may encompass other requirements (e.g. legal), or have information about the provenance of the data access committees and other relevant processes. This would aid in maintaining documentation, further using automation to ensure it is correct, and perform follow-up actions periodically or as contextually required. In all of these, the benefits do not require everyone to adopt a large amount of technical debt, and adopters of DUO can choose what and how they wish to utilise - in our case the choice of adopting just the ODRL rules, or its matching algorithm, or also the connection to legal compliance using DPV. Our primary contribution is in demonstrating their usefulness and providing a path for their adoption.

6.2. Integration into Existing Implementations

We acknowledge that some of our proposed changes may break backwards or existing compatibility with DUO utilising systems, and therefore suggest any adopter to perform an assessment regarding whether the gains obtained from such changes outweighs the cost of making these changes. In our opinion, our changes do offer more advantages than disadvantages in longer term, and therefore they should be adopted gradually if not immediately. We recommend the adoption of equivalent ODRL policies for DUO concepts and the (re-)structuring of existing taxonomies and concepts as the first steps. After this, systems such as DUOS can take advantage of the increased availability of machine-readable data to enhance their data discovery and matching algorithms.

We also acknowledge the value of DUO concepts in being simple for stakeholders to understand and utilise, and their basis in 'textual clauses' such as those offered in informed consent or data donation/release forms. With this in mind, our modelling of ODRL policies ensures that there is no immediate need to replace the use of DUO concepts since the ODRL policies are complimentary to these i.e. the ODRL policy is linked to DUO concepts rather than replacing them entirely. Thus, stakeholders who lack or have limited technical expertise can continue to utilise DUO concepts as they have, with machine-based implementations taking advantage of the increased clarity and specificity of ODRL rules associated with those DUO concepts. An important advantage this provides, that is not possible in the current DUO based implementations, is from the underlying constraints or conditions being made explicit, thereby providing a larger avenue for where automation and logic-based reasoning can be applied. This makes it possible to scale the approach to larger and more diverse use-cases than is currently feasible with DUO.

It also offers the possibility to encode as machine-readable metadata what is currently external information i.e. (i) who: the data is about, requested access, was granted access; or (ii) followup duties once data has been released: checking whether it has been fulfilled, documenting fulfilment or violation; (iii) legal obligations associated with data use. All these information and factors are what DUO-utilising systems currently utilise (such as DUOS) and will do so in any practical use-case in the future. By providing a clear path for adopters to express this information, the use of DUO can be made more systematic and consistent - thereby also increasing the potential co-operation between adopters and facilitating cross-boundary data requests and access as envisioned by GA4GH.

6.3. Assisting with Legal Compliance

Currently DUO or GA4GH do not provide information on how the use of its efforts relate to legal interpretation and obligations, though they have ongoing discussions for the same. This is a particularly challenging task given the global scope of the work which encompasses different jurisdictions and their laws, and that laws such as GDPR are fairly recent in terms of how their obligations are understood to be applied. We suggest the use of domain-agnostic vocabularies such as ODRL and DPV to first provide a clear indication of how DUO and DUO-based systems relate to specific concepts within legal terminology. By using these within ODRL policies, DUO can provide what is effectively a digital contract.

Further specific jurisdictional applications can then be introduced as an extension of these. For example, the DPV-GDPR extension provides a convenient way to specify GDPR's legal bases and rights alongside DPV. This reduces the burden on adopters who do not want to express this information or do not want to express any jurisdiction-specific information. For example, a data depositor who only stipulates use of data should be based on consent without explicitly defining the conditions for that valid consent can be expressed as a policy using ODRL and DPV. The oversight committee or an ethics board can then evaluate this further based on their knowledge of the valid consenting requirements, and add additional restrictions or obligations to follow a specific regulation such as the GDPR before permitting use of that data by using DPV-GDPR.

This freedom also offers benefits for systems like DUOS that can explicitly denote datasets as requiring GDPR-level consenting or its applicability by adding relevant metadata to the dataset policy. Doing so assists the matching process to also check for legal obligations and compatibility, such as by requiring specific information about the requestor (e.g. a Data Protection Officer), or requiring additional legal bases and safeguards for transfer of that data (e.g. outside EU). Through this, DUO and its applications can gain a wider legal applicability across the globe and also have the means and mechanisms to address specific interpretations of the law. And given that all this information would be machine-readable and shareable with the dataset, it can be used by both provider and requesting entity for automation in identifying and checking fulfilment of legal obligations based on utilising the existing state of the art.

7. Conclusion

The Data Use Ontology (DUO) is an important initiative to enable wider data sharing towards the goal of progressing health and medical research. Its design and application is driven by the workflows and use-cases present in a socio-technical system consisting of a data repository, utilising a data access committee or approval board, and maintaining compatibility with textual clauses and machine-readable metadata.

We provide an argument for why the design of DUO concepts should be enhanced in terms of making its data use conditions explicit – also as machine-readable data and to utilise these in the matching of data use policies and requests. For this, we have demonstrated the applicability, suitability, and potential of ODRL as a standardised language to express all facets of DUO's applications. We provide: (i) ODRL rules for each DUO concept; (ii) Integration of DUO concepts into an ODRL policy for a dataset; (iii) ODRL policy representing a data use request; and (iv) Automated checking for compatibility between dataset and request policies. Through these, we provide a better mechanism for the use of machine-readable information and its use in automation as compared to the current DUO implementation.

In addition to the above, we also demonstrated how the use of DPV within ODRL policies enables connection with privacy and data protection laws without making it specific to a particular jurisdiction. For cases where a

specific law is needed, the DPV concepts can be easily extended, which we showed for GDPR. Along with the descriptions of our research, we also provided links to resources and a demonstration of its implementation to assist adopters of DUO in assessing and using our work.

Importantly, rather than suggesting a radical new method of doing things, we started with the goal of constructing a mechanism that complements DUO rather than replacing it. As we've shown, using ODRL and DPV alongside DUO is possible, and can be done with minimum disruption. Through this, we hope to have our work influence and improve existing DUO-related efforts, and in doing this to bring DUO and the GA4GH closer towards implementing the EU's Health Data Space vision.

Acknowledgements

Funding: Harshvardhan J. Pandit has received funding from the 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. Beatriz Esteves has received funding from European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 813497 (PROTECT).
Thanks: We thank Víctor Rodríguez-Doncel for valuable insight and inputs regarding the use of ODRL.

References

[1] J. Lawson, M.N. Cabili, G. Kerry, T. Boughtwood, A. Thorogood, P. Alper, S.R. Bowers, R.R. Boyles, A.J. Brookes, M. Brush, T. Burdett, H. Clissold, S. Donnelly, S.O.M. Dyke, M.A. Freeberg, M.A. Haendel, C. Hata, P. Holub, F. Jeanson, A. Jene, M. Kawashima, S. Kawashima, M. Konopko, I. Kyomugisha, H. Li, M. Linden, L.L. Rodriguez, M. Morita, N. Mulder, J. Muller, S. Nagaie, J. Nasir, S. Ogishima, V. Ota Wang, L.D. Paglione, R.N. Pandya, H. Parkinson, A.A. Philippakis, F. Prasser, J. Rambla, K. Reinold, G.A. Rushton, A. Saltzman, G. Saunders, H.J. Sofia, J.D. Spalding, M.A. Swertz, I. Tulchinsky, E.J. van Enckevort, S. Varma, C. Voisin, N. Yamamoto, C. Yamasaki, L. Zass, J.M. Guidry Auvil, T.H. Nyrönen and M. Courtot, The Data Use Ontology to Streamline Responsible Access to Human Biomedical Datasets, *Cell Genomics* **1**(2) (2021), 100028. doi:10.1016/j.xgen.2021.100028.

[2] H.L. Rehm, A.J.H. Page, L. Smith, J.B. Adams, G. Alterovitz, L.J. Babb, M.P. Barkley, M. Baudis, M.J.S. Beauvais, T. Beck, J.S. Beckmann, S. Beltran, D. Bernick, A. Bernier, J.K. Bonfield, T.F. Boughtwood, G. Bourque, S.R. Bowers, A.J. Brookes, M. Brudno, M.H. Brush, D. Bujold, T. Burdett, O.J. Buske, M.N. Cabili, D.L. Cameron, R.J. Carroll, E. Casas-Silva, D. Chakravarty, B.P. Chaudhari, S.H. Chen, J.M. Cherry, J. Chung, M. Cline, H.L. Clissold, R.M. Cook-Deegan, M. Courtot, F. Cunningham, M. Cupak, R.M. Davies, D. Denisko, M.J. Doerr, L.I. Dolman, E.S. Dove, L.J. Dursi, S.O.M. Dyke, J.A. Eddy, K. Eilbeck, K.P. Ellrott, S. Fairley, K.A. Fakhro, H.V. Firth, M.S. Fitzsimons, M. Fiume, P. Flicek, I.M. Fore, M.A. Freeberg, R.R. Freimuth, L.A. Fromont, J. Fuerth, C.L. Gaff, W. Gan, E.M. Ghanaim, D. Glazer, R.C. Green, M. Griffith, O.L. Griffith, R.L. Grossman, T. Groza, J.M. Guidry Auvil, R. Guigó, D. Gupta, M.A. Haendel, A. Hamosh, D.P. Hansen, R.K. Hart, D.M. Hartley, D. Haussler, R.M. Hendricks-Sturrup, C.W.L. Ho, A.E. Hobb, M.M. Hoffman, O.M. Hofmann, P. Holub, J.S. Hsu, J.-P. Hubaux, S.E. Hunt, A. Husami, J.O. Jacobsen, S.S. Jamuar, E.L. Janes, F. Jeanson, A. Jené, A.L. Johns, Y. Joly, S.J.M. Jones, A. Kanitz, K. Kato, T.M. Keane, K. Kekesi-Lafrance, J. Kelleher, G. Kerry, S.-S. Khor, B.M. Knoppers, M.A. Konopko, K. Kosaki, M. Kuba, J. Lawson, R. Leinonen, S. Li, M.F. Lin, M. Linden, X. Liu, I.U. Liyanage, J. Lopez, A.M. Lucassen, M. Lukowski, A.L. Mann, J. Marshall, M. Mattioni, A. Metke-Jimenez, A. Middleton, R.J. Milne, F. Molnár-Gábor, N. Mulder, M.C. Munoz-Torres, R. Nag, H. Nakagawa, J. Nasir, A. Navarro, T.H. Nelson, A. Niewielska, A. Nisselle, J. Niu, T.H. Nyrönen, B.D. O'Connor, S. Oesterle, S. Ogishima, V. Ota Wang, L.A.D. Paglione, E. Palumbo, H.E. Parkinson, A.A. Philippakis, A.D. Pizarro, A. Prlic, J. Rambla, A. Rendon, R.A. Rider, P.N. Robinson, K.W. Rodarmer, L.L. Rodriguez, A.F. Rubin, M. Rueda, G.A. Rushton, R.S. Ryan, G.I. Saunders, H. Schuilenburg, T. Schwede, S. Scollen, A. Senf, N.C. Sheffield, N. Skantharajah, A.V. Smith, H.J. Sofia, D. Spalding, A.B. Spurdle, Z. Stark, L.D. Stein, M. Suematsu, P. Tan, J.A. Tedds, A.A. Thomson, A. Thorogood, T.L. Tickle, K. Tokunaga, J. Törnroos, D. Torrents, S. Upchurch, A. Valencia, R.V. Guimera, J. Vamathevan, S. Varma, D.F. Vears, C. Viner, C. Voisin, A.H. Wagner, S.E. Wallace, B.P. Walsh, M.S. Williams, E.C. Winkler, B.J. Wold, G.M. Wood, J.P. Woolley, C. Yamasaki, A.D. Yates, C.K. Yung, L.J. Zass, K. Zaytseva, J. Zhang, P. Goodhand, K. North and E. Birney, GA4GH: International Policies and Standards for Data Sharing across Genomic Research and Healthcare, *Cell Genomics* **1**(2) (2021), 100029. doi:10.1016/j.xgen.2021.100029.

[3] 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* **L119** (2016).

[4] R. Iannella, M. Steidl, S. Myles and V. Rodriguez-Doncel, ODRL Vocabulary & Expression 2.2, 2018.

[5] H.J. Pandit, A. Polleres, B. Bos, R. Brennan, B. Bruegger, F.J. Ekaputra, J.D. Fernández, R.G. Hamed, M. Lizar, E. Schlehahn, S. Steyskal and R. Wenning, Creating A Vocabulary for Data Privacy, in: *The 18th International Conference on Ontologies, DataBases, and Applications of Semantics (ODBASE2019)*, Rhodes, Greece, 2019, p. 17. doi:10/ggwx7x.

[6] S.O.M. Dyke, A.A. Philippakis, J.R.D. Argila, D.N. Paltoo, E.S. Luetkemeier, B.M. Knoppers, A.J. Brookes, J.D. Spalding, M. Thompson, M. Roos, K.M. Boycott, M. Brudno, M. Hurles, H.L. Rehm, A. Matern, M. Fiume and S.T. Sherry, Consent Codes: Upholding Standard Data Use Conditions, *PLOS Genetics* **12**(1) (2016), e1005772. doi:10.1371/journal.pgen.1005772.

[7] J.P. Woolley, E. Kirby, J. Leslie, F. Jeanson, M.N. Cabili, G. Rushton, J.G. Hazard, V. Ladas, C.D. Veal, S.J. Gibson, A.-M. Tassé, S.O.M. Dyke, C. Gaff, A. Thorogood, B.M. Knoppers, J. Wilbanks and A.J. Brookes, Responsible Sharing of Biomedical Data and Biospecimens via the "Automatable Discovery and Access Matrix" (ADA-M), *npj Genomic Medicine* **3**(1) (2018), 1–6. doi:10/gg2mpc.

[8] M.N. Cabili, J. Lawson, A. Saltzman, G. Rushton, P. O'Rourke, J. Wilbanks, L.L. Rodriguez, T. Nyronen, M. Courtot, S. Donnelly and A.A. Philippakis, Empirical Validation of an Automated Approach to Data Use Oversight, *Cell Genomics* **1**(2) (2021), 100031. doi:10.1016/j.xgen.2021.100031.

[9] M.C. Schatz, A.A. Philippakis, E. Afgan, E. Banks, V.J. Carey, R.J. Carroll, A. Culotti, K. Ellrott, J. Goecks, R.L. Grossman, I.M. Hall, K.D. Hansen, J. Lawson, J.T. Leek, A.O. Luria, S. Mosher, M. Morgan, A. Nekrutenko, B.D. O'Connor, K. Osborn, B. Paten, C. Patterson, F.J. Tan, C.O. Taylor, J. Vessio, L. Waldron, T. Wang, K. Wuichet, A. Baumann, A. Rula, A. Kovalsy, C. Bernard, D. Caetano-Anollés, G.A. Van der Auwera, J. Canas, K. Yuksel, K. Herman, M.M. Taylor, M. Simeon, M. Baumann, Q. Wang, R. Title, R. Munshi, S. Chaluvadi, V. Reeves, W. Disman, S. Thomas, A. Hajian, E. Kiernan, N. Gupta, T. Vosburg, L. Geistlinger, M. Ramos, S. Oh, D. Rogers, F. McDade, M. Hastie, N. Turaga, A. Ostrovsky, A. Mahmoud, D. Baker, D. Clements, K.E.L. Cox, K. Suderman, N. Kucher, S. Golitsynskiy, S. Zarate, S.J. Wheelan, K. Kammers, A. Stevens, C. Hutter, C. Wellington, E.M. Ghanaim, K.L. Wiley, S.K. Sen, V. Di Francesco, D. s Yuen, B. Walsh, L. Sargent, V. Jalili, J. Chilton, L. Shepherd, B.J. Stubbs, A. O'Farrell, B.A. Vizzier, C. Overbeck, C. Reid, D.C. Steinberg, E.A. Sheets, J. Lucas, L. Blauvelt, L. Cabansay, N. Warren, B. Hannafious, T. Harris, R. Reddy, E. Torstenson, M.K. Banasiewicz, H.J. Abel and J. Walker, Inverting the Model of Genomics Data Sharing with the NHGRI Genomic Data Science Analysis, Visualization, and Informatics Lab-space, *Cell Genomics* **2**(1) (2022), 100085. doi:10.1016/j.xgen.2021.100085.

[10] V. Nembaware, K. Johnston, A.A. Diallo, M.J. Kotze, A. Matimba, K. Moodley, G.B. Tangwa, R. Torrorey-Sawe and N. Tiffin, A Framework for Tiered Informed Consent for Health Genomic Research in Africa, *Nature Genetics* **51**(11) (2019), 1566–1571. doi:10.1038/s41588-019-0520-x.

[11] V. Jaiman and V. Urovi, A Consent Model for Blockchain-Based Health Data Sharing Platforms, *IEEE Access* **8** (2020), 143734–143745. doi:10.1109/ACCESS.2020.3014565.

[12] M.A. Haas, H. Teare, M. Prictor, G. Ceregra, M.E. Vidgen, D. Bunker, J. Kaye and T. Boughtwood, 'CTRL': An Online, Dynamic Consent and Participant Engagement Platform Working towards Solving the Complexities of Consent in Genomic Research, *European Journal of Human Genetics* **29**(4) (2021), 687–698. doi:10.1038/s41431-020-00782-w.

[13] G. Alter, A. Gonzalez-Beltran, L. Ohno-Machado and P. Rocca-Serra, The Data Tags Suite (DATS) Model for Discovering Data Access and Use Requirements, *GigaScience* **9**(2) (2020), giz165. doi:10.1093/gigascience/giz165.

[14] G. Saunders, M. Baudis, R. Becker, S. Beltran, C. Béroud, E. Birney, C. Brooksbank, S. Brunak, M. Van den Bulcke, R. Drysdale, S. Capella-Gutierrez, P. Flicek, F. Florindi, P. Goodhand, I. Gut, J. Heringa, P. Holub, J. Hooyberghs, N. Juty, T.M. Keane, J.O. Korbel, I. Lappalainen, B. Leskosek, G. Matthijs, M.T. Mayrhofer, A. Metspalu, A. Navarro, S. Newhouse, T. Nyrönen, A. Page, B. Persson, A. Palotie, H. Parkinson, J. Rambla, D. Salgado, E. Steinfelder, M.A. Swertz, A. Valencia, S. Varma, N. Blomberg and S. Scollen, Leveraging European Infrastructures to Access 1 Million Human Genomes by 2022, *Nature Reviews Genetics* **20**(11) (2019), 693–701. doi:10.1038/s41576-019-0156-9.

[15] M. Amith, M.R. Harris, C. Stansbury, K. Ford, F.J. Manion and C. Tao, Expressing and Executing Informed Consent Permissions Using SWRL: The All of Us Use Case, *AMIA Annual Symposium Proceedings* **2021** (2022), 197–206.

[16] S. Grabus and J. Greenberg, The Landscape of Rights and Licensing Initiatives for Data Sharing, *Data Science Journal* **18**(1) (2019), 29. doi:10.5334/dsj-2019-029.

[17] R. Wenning and S. Kirrane, Compliance Using Metadata, in: *Semantic Applications: Methodology, Technology, Corporate Use*, T. Hoppe, B. Humm and A. Reibold, eds, Springer Berlin Heidelberg, Berlin, Heidelberg, 2018, pp. 31–45. ISBN 978-3-662-55433-3.

[18] C.M.d.O. Rodrigues, F.L.G. de Freitas, E.F.S. Barreiros, R.R. de Azevedo and A.T. de Almeida Filho, Legal Ontologies over Time: A Systematic Mapping Study, *Expert Systems with Applications* **130** (2019), 12–30. doi:10/gf223z.

[19] P.A. Bonatti, L. Ioffredo, I.M. Petrova, L. Sauro and I.R. Siahaan, Real-Time Reasoning in OWL2 for GDPR Compliance, *Artificial Intelligence* **289** (2020), 103389. doi:10.1016/j.artint.2020.103389.

[20] M.D. Vos, S. Kirrane, J. Padget and K. Satoh, ODRL Policy Modelling and Compliance Checking, in: *3rd International Joint Conference on Rules and Reasoning (RuleML+RR 2019)*, Bolzano, Italy, 2019, p. 16.

[21] H.J. Pandit, K. Fatema, D. O'Sullivan and D. Lewis, GDPRtEXT - GDPR as a Linked Data Resource, in: *The Semantic Web - European Semantic Web Conference*, Lecture Notes in Computer Science, Springer, Cham, 2018, pp. 481–495. ISBN 978-3-319-93416-7 978-3-319-93417-4. doi:10/c3n4.

[22] H.J. Pandit, C. Debruyne, D. O'Sullivan and D. Lewis, GConsent - A Consent Ontology Based on the GDPR, in: *The Semantic Web*, P. Hitzler, M. Fernández, K. Janowicz, A. Zaveri, A.J.G. Gray, V. Lopez, A. Haller and K. Hammar, eds, Lecture Notes in Computer Science, Springer International Publishing, 2019, pp. 270–282. ISBN 978-3-030-21348-0.

[23] A. Kurteva, T.R. Chhetri, H.J. Pandit and A. Fensel, Consent through the Lens of Semantics: State of the Art Survey and Best Practices, *Semantic Web* **Preprint**(Preprint) (2021), 1–27. doi:10/gmsjzn.

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

[25] B. Esteves, H.J. Pandit and V. Rodríguez-Doncel, ODRL Profile for Expressing Consent through Granular Access Control Policies in Solid, in: *2021 IEEE European Symposium on Security and Privacy Workshops (EuroS PW)*, 2021, pp. 298–306. ISSN 2768-0657. doi:10/gnck5x.

[26] T. Kuhn, A Survey and Classification of Controlled Natural Languages, *Computational Linguistics* **40**(1) (2014), 121–170. doi:10/gf6grx.