Semantic Web -1 (2024) 1–26
DOI 10.3233/SW-243583
IOS Press

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, Dublin City University, Ireland*
E-mail: me@harshp.com
[b] *Ontology Engineering Group, Universidad Politécnica de Madrid, Spain*
E-mail: beatriz.gesteves@upm.es

Editor: Cogan Shimizu, Wright State University, USA
Solicited reviews: Jaime Delgado, Universitat Politècnica de Catalunya, Spain; three anonymous reviewers

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 encoded using OWL, only contain the textual descriptions of the conditions for data use they represent, and do not specify the intended permissions, prohibitions, and obligations explicitly – which limits their usefulness. We present an exploration of how the Open Digital Rights Language (ODRL) can be used to explicitly represent the information inherent in DUO concepts to create policies that are then used to represent conditions under which datasets are available for use, conditions in requests to use them, and to generate agreements based on a compatibility matching between the two. We also address a current limitation of DUO regarding specifying information relevant to privacy and data protection law by using the Data Privacy Vocabulary (DPV) which supports expressing legal concepts in a jurisdiction-agnostic manner as well as for specific laws like the GDPR. Our work supports the existing socio-technical governance processes involving use of DUO by providing a complementary rather than replacement approach. To support this and improve DUO, we provide a description of how our system can be deployed with a proof of concept demonstration that uses ODRL rules for all DUO concepts, and uses them to generate agreements through matching of requests to data offers. 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

*Corresponding author(s). E-mails: me@harshp.com, beatriz.gesteves@upm.es.

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 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) [16,23] 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, express data use requests, and identify or discover 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 the `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) [22] 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 GDPR. However, compatibility with existing regulations is an important and mandatory requirement that each adopter and data user must fulfil, and where the lack of support from the specification risks harming the interoperability through fragmentation in approaches. Additionally, the EU envisions a 'Health Data Space'[4] where machine-readability and automation will play an important role in facilitating the exchange of data without prejudice to existing regulations such as the GDPR.

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 data provider's and requestor's conditions for data use are expressed as policies, and are compared to evaluate whether the intended use is permissible. 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 as not all relevant information can be explicitly represented.

More importantly, to automate this process, a set of requirements should be taken into consideration when choosing a vocabulary to express dataset usage conditions, such as: (i) the expressiveness for defining specifics of rules and policies, i.e., the specification of actions, purposes, or other constraints as concepts that can be independently expressed and assessed, and their combinations to represent different categories of policies; (ii) the ability to associate and check their conformance and compliance with legal requirements; and (iii) the ability to specify requirements in

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

machine-readable form and use them to assess correctness and completeness of information. Such solutions have existed for a while now – for example, Answer Set Programming (ASP) and logic-based semantic reasoners have been utilised in a variety of domains – including for representing information and using it for checking legal compliance for GDPR (see Section 2.2).

With the above motivation, we present an approach for representing the inherent information and rules in DUO concepts explicitly in RDF through use of the Open Digital Rights Language[6] (ODRL) [11] it is the W3C standard developed explicitly to model rules and policies and also concerns the intended requirements for which DUO was created. We specifically chose ODRL because: (i) it uses RDF and is machine-readable; (ii) it provides concepts modelling the domain-specific and legally-relevant terms to represent constraints – e.g. spatial and temporal, and types of policies – e.g. offers, requests and agreements, along with the flexibility to use them in similar manner as the conventional contents and structures of legal agreements; (iii) the use of ODRL can be validated[7] and a formal semantics specification[8] is being actively developed by the W3C ODRL Community Group (ODRL CG)[9] to ensure correctness and consistency on the deployments of services that use ODRL; (iv) the specification provides the ability to develop extensions through ODRL profiles;[10] and

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 annotations to datasets and are used to request access to them can continue without hindrance, and DUO stakeholders 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, for which we use the Data Privacy Vocabulary[11] (DPV) [20], an output of the W3C Data Privacy Vocabularies and Controls Community Group[12][13] (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

In addition to these, a late contribution is the preliminary analysis of two articles providing improvements to DUO that were published while this article was under review. We provide a summary of these recent developments, compare it with the work presented in this paper, and discuss the continued relevance and benefits of our contributions.

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, the late contribution containing analysis of recent work in Section 7, and concluding statements in Section 8.

[6] https://www.w3.org/TR/odrl-vocab/ The prefix `odrl` has the IRI http://www.w3.org/ns/odrl/2/.
[7] Implementation of a ODRL validator using SHACL available at https://odrlapi.appspot.com/.
[8] ODRL Formal Semantics CG report available at https://w3c.github.io/odrl/formal-semantics/.
[9] Note: The author (Beatriz Esteves) is a contributing member of the ODRL CG's work on the development of a formal semantics specification.
[10] ODRL Profile Best Practices CG report available at https://w3c.github.io/odrl/profile-bp/.
[11] https://w3id.org/dpv The prefix `dpv` has the IRI https://w3id.org/dpv#.
[12] https://www.w3.org/community/dpvcg/
[13] Note: Both authors are active contributing members to DPV.

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[14] (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[15] (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[16] (ADA-M) [29] 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'.[17] A brief outline and summary of DUO and its use to streamline access to biomedical datasets is presented in [16], and a list of GA4GH initiatives and standards along with the relevance of DUO within those is presented in [23].

The Data Use Oversight System[18] (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 data access requests, which are then matched using an algorithm based on hierarchical compatibility i.e. permitted conditions identified based on 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 decision-making by human data access committees [4]. DUOS is currently being implemented in an ongoing large-scale pilot [26].

Other uses of DUO include specification of informed consent for health and genomics research in Africa [17], along with ADA-M for representing consent for health data sharing in a blockchain [12], and in CTRL [10] – 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) [1] 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 [25]. We found only one article that provided a machine-readable metadata representation of information using DUO – which used SWRL[19] to express the rules [2]. 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 [9].

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 [1] 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 of 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*.

[14] http://obi-ontology.org/
[15] https://environmentontology.github.io/gaz/
[16] https://www.ga4gh.org/genomic-data-toolkit/regulatory-ethics-toolkit/
[17] https://www.ga4gh.org/wp-content/uploads/Machine-readable-Consent-Guidance_6JUL2020-1.pdf
[18] https://duos.broadinstitute.org/
[19] https://www.w3.org/Submission/SWRL/

2.2. Expression of machine-readable information and policies for GDPR

Given that health data is 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[20] (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 meeting their compliance requirements [28].

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. [24]). Of note regarding the matching of DUO dataset annotations is the policy checking algorithm for GDPR developed by SPECIAL H2020 project [3] 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 [27]. 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 [19] provides a vocabulary of concepts, and GConsent [18] 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) [20] provides a taxonomy of concepts which can be used as jurisdiction-agnostic terms with an extension for specific concepts from GDPR[21] 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.

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. [15], describes the approaches associated with consent, and the second, by Esteves and Rodrigues-Doncel [8], analyses ontologies and policy languages for modelling information flows. Both highlight the variety of approaches available, and offer opinionated suggestions regarding the use of ODRL and DPV – which we have incorporated in our choice of implementations.[22] Based 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 descriptions 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[18]. 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 conditions that can be checked, verified, and consumed in an automated manner to perform tasks associated with validation of dataset policies,

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

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

[22] 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.

querying to discover suitable datasets, and to aid in matching requests with available data and its usage conditions. We also considered the contextual cases for representing additional conditions or information requirements such as records-keeping by institutions or for legal compliance, for which ODRL is also suitable as highlighted in our motivation.

Our methodology in this was to first analysedeletedd DUO concepts and textual information to identify their relevant representation as ODRL concepts. We then constructed rules expressing identified conditions and expressed them using ODRL along with identifying three categories of 'policies' from GA4GH's cases reflecting data usage, data request, and an agreement based on compatibility between the two. We then constructed a matching algorithm that utilised developed ODRL policies to compare a request policy with a dataset's policy to determine compatibility and to create an agreement where both were found to be compatible.

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 use 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. For example, the DUO concept for code `HMB` should be expressed in terms of it being a *permission* for *purpose* of type *HMB* which is also not a purpose of type *POA*. In this manner, the concept (*HMB*) is better expressed and applied by exposing its underlying concepts (*purpose*) and rules over it (*permission*). See Section 3.3 for more examples of using semantic concepts to represent implicit information in DUO codes, and DPV [20] for additional taxonomies and concepts available to further specify relevant information.

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.

We faced challenges in interpreting specific phrases such as "*is limited to*" which imply that usage is permitted 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 explicitly express the underlying concepts and how they are applied to create permissions, prohibitions, and obligations, and then using existing methods for ODRL [21,27] and OWL [3] to reason over them.

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 the recommendation to use the MONDO ontology[23] 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.

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

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 and 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 is :CollaborateWithStudyPI	
DUO0000021	IRB	Duty	Action is :ProvideEthicalApproval	
DUO0000016	GSO	Permission	Purpose is :GS or :GSG	
DUO0000016	GSO	Prohibition	Purpose is :GS and not :GSG	
DUO0000022	GS	Permission	Spatial is equal to specified :Location	:TemplateLocation
DUO0000022	GS	Prohibition	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 :NonProfitOrganisation and Purpose is :NCU	
DUO0000018	NPUNCU	Prohibition	Assignee is :ForProfitOrganisation and Purpose is :NCU	
DUO0000018	NPUNCU	Prohibition	Assignee is :NonProfitOrganisation and Purpose is not :NCU	
DUO0000046	NCU	Permission	Purpose is :NCU	
DUO0000046	NCU	Prohibition	Purpose is not :NCU	
DUO0000045	NPU	Permission	Assignee is :NonProfitOrganisation	
DUO0000045	NPU	Prohibition	Assignee is :ForProfitOrganisation	
DUO0000044	NPOA	Prohibition	Purpose is :POA	
DUO0000027	PS	Permission	Project is :ApprovedProject	:TemplateProject
DUO0000027	PS	Prohibition	Project is not :ApprovedProject	:TemplateProject
DUO0000024	MOR	Duty	Action is odrl:distribute :ResultsOfStudies with odrl:dateTime	:TemplateDateTime
DUO0000019	PUB	Duty	Action is odrl:distribute :ResultsOfStudies	
DUO0000012	RS	Permission	Purpose is specified :Research	:TemplateResearch
DUO0000012	RS	Prohibition	Purpose is not specified :Research	:TemplateResearch
DUO0000029	RTN	Duty	Action is :ReturnDerivedOrEnrichedData	
DUO0000025	TS	Permission	Time is less than specified :TemplateDateTime	:TemplateDateTime
DUO0000026	US	Permission	Assignee is :ApprovedUser	:TemplateUser
DUO0000026	US	Prohibition	Assignee is not :ApprovedUser	:TemplateUser
OBI0000066	**Data Use Permission**			
DUO0000034		Permission	Purpose is :AgeCategoryResearch	
DUO0000034		Permission	Age is specified :Age	:TemplateAgeCategory
DUO0000033		Permission	Purpose is :POA	
DUO0000037		Permission	Purpose is :HMB	
DUO0000040		Permission	Purpose is :DS and mondo:0000001	:TemplateDisease
DUO0000039		Permission	Purpose is :DrugDevelopment	
DUO0000038		Permission	Purpose is :GS	
DUO0000035		Permission	Purpose is :GenderCategoryResearch	
DUO0000035		Permission	Gender is specified :Gender	:TemplateGender

Table 1
(Continued)

Concept	Code	Rule Type	Constraint	Placeholder
DUO0000031		Permission	Purpose is :MDS	
DUO0000032		Permission	Purpose is :PopulationGroupResearch	
DUO0000032		Permission	Population is specified :Population	:TemplatePopulation
DUO0000036		Permission	Purpose is :ResearchControl	

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 the 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 by information about the entities involved, provenance associated with the 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 that are indicated as being applicable 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. DUO also does not specify how to indicate or identify values associated with conditions such as specific diseases or temporal duration. To ensure ODRL policies are always valid, and to clearly indicate how to later apply or *instantiate* them for a dataset, we created the class `TemplateQuery` whose instances represent a placeholder to be substituted with the actual value(s) retrieved by executing a SPARQL query associated with it through the property `sparqlExpression`. In Table 1 these are indicated as *Placeholder*. Examples of this can be seen in Listing 3 for a `odrl:Set` representing a DUO permission which is then used in a request. The placeholders are used to indicate datasets and assignees

```
: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

which are not known ahead of time, and which are substituted with actual instances in real policies by using the SPARQL queries associated with each placeholder.

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 the specific values or instances, e.g. diseases or locations, 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;[24] (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

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

```
: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`

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.[25]
5. Add provenance information or other additional documentation, e.g. `dct:dateSubmitted` for when the dataset was added to a system.

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.

3.5. *Data use decisions as `odrl:Agreement`*

Instances of `odrl:Agreement` are recorded outcomes of decisions resulting from matching processes where access to the data has been granted or denied. In this, the ODRL terms assist in specifying who has granted or denied the access (`odrl:assigner`), to whom (`odrl:assignee`), for what resources (`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

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

```
: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`

```
:Agreement a odrl:Agreement ;
    dct:references <https://example.com/Offer>, <https://example.com/Request_for_GS> ;
    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

data use decision as an `odrl:Agreement` between a data depositor and a data requestor for the purpose of genetic studies is presented in Listing 4.

The following algorithm is used to create the `odrl:Agreement` instance:

1. Retrieve the `odrl:Request` and dataset's `odrl:Offer`.
2. Match the `odrl:Offer` with the `odrl:Request` (the algorithm is defined in Section 3.6).
3. Record the result where `odrl:target` property specifies the dataset, and `odrl:assignee` and `odrl:assigner` identify the data provider and the requestor respectively.
4. If the matching result shows a compatibility between the request and the offer, then access is expressed as permissible by using a permission with a constraint on the requested purpose for access, as well as any other additional constraints, e.g., spatial, temporal, or duties on the `odrl:assigner`. If the access is denied, similar information is added to policy as a prohibition.
5. `dct:references` is used to associate the agreement with the `odrl:Offer` and `odrl:Request` that are being matched.

6. Provenance and other relevant information, e.g., `dct:dateAccepted` is added to document the agreement's creation and acceptance amongst the parties.

In the matching algorithm, we only considered the case where a request is matched with a dataset's offer. In a practical situation, there may be a single broad request that could have potential matches with several datasets, and it may be undesirable to run the matching against all possible combinations of requests and datasets. To select relevant datasets, a filtering mechanism can be used, such as based on the request's specified purpose, and the dataset's policies could be indexed in a database to enable efficient retrieval and matching. We note these as candidates for future improvements in the progression of this work.

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 uses the agreement to represent a use case 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 [3] and for granting access to resources in Solid [7].

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 criterion, 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., [3] and [7]) 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 satisfied 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, the request 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 [7]. 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, the result is that conditions are non-compatible. If no prohibitions are found and all permissions are satisfied, then the result is that the conditions are compatible. Note that here the matching only asserts the compatibility of the dataset usage policy and request, whose result is then used to make a decision on whether to grant or refuse access.

Algorithm 1 provides a pseudo-code representing the steps to be performed for policy matching. Please note that the algorithm only represents a broad indication of actions and that the DUO documentation lacks specifics for correctly interpreting aspects of semantics. Given that this interpretation has a significant impact on the decision-making

Algorithm 1 Pseudo-code of the matching algorithm

for *prohibition* ← *odrl*:*Offer* **do**
  if *odrl*:*assignee* ∈ *offer*:*prohibition* **then**
    if *offer*:*assignee* ≡ *request*:*assignee* **then** *decision* ← *DENY*
  for *constraint* ← *prohibition* **do**
    if *odr*:*spatial* ∈ *constraint* **then**
      if *offer*:*spatial* ∩ *request*:*spatial* ≠ ∅ **then** *decision* ← *DENY*
    else if *duodrl*:*Project* ← *constraint* **then**
      if *request*:*project* ∩ *offer*:*project* ≠ ∅ **then** *decision* ← *DENY*
    else if *odrl*:*dateTime* ← *constraint* **then**
      if *timeNow* < *moratoriumDate* **then** *decision* ← *DENY*
    else if *offer*:*purpose* ∩ *request*:*purpose* ≠ ∅ **then** *decision* ← *DENY*
for *permission* ← *odrl*: *Offer* **do**
  if *odrl*:*assignee* ∈ *offer*:*permission* **then**
    if *offer*:*assignee* ≢ *request*: *assignee* **then** *decision* ← *DENY*
  for *constraint* ← *permission* **do**
    if *odrl*:*dateTime* ∈ *constraint* **then**
      if *timeNow* > *timeLimit* **then** *decision* ← *DENY*
    else if *request*:*purpose* ∈ *groupResearchPurposes* **then**
      if *request*:*purpose* ⊈ *offer*:*purpose* ⋁ *request*:*group* ⊈ *offer*:*group* **then** *decision* ← *DENY*
    else if *request*:*purpose* ⊈ *offer*:*purpose* **then** *decision* ← *DENY*
if ∄*DENY* **then** *decision* ← *GRANT*

within DUO's processes and that DUO only specifies interpretation of hierarchical concepts only for purposes but not others (e.g. location, users, projects) – we explicitly identify this as a topic that requires further consideration and investigation in terms of better understanding and expressing the interpretation of DUO's conditions in approval decision-making processes. To remedy this lack of information, except for the purposes in DUO's concepts, we followed existing implementations for legally relevant interpretation of hierarchical concepts [3] where a narrower concept or a sub-class cannot be considered compatible with a request for a broader concept or parent-class. For example, a permission for city as a location cannot be satisfied by a request for a region containing that city.

The algorithm reflects DUO's prohibitive interpretation in matching where the offer's prohibitions are checked and ensured to be satisfied before any permissions are checked. The prohibition checking will deny the request if any of the following constraints in the offer are incompatible with the request:

1. offer assignee matches[26] (≡) the request;
2. offer has a spatial constraint matching or not satisfying (∩ ≠ ∅) the request;
3. request has a project matching (∩ ≠ ∅) the project in offer;
4. there is a moratorium with a date in the future; and
5. request has a purpose matching (∩ ≠ ∅) the purpose in offer.

If no prohibitions are found, the permissions are checked next. The permission checking will deny the request if any of the following constraints in the offer are incompatible with the request:

1. offer assignee does not match (≢) the assignee of the request;
2. offer time limit on use has lapsed;
3. offer has a group-related research purpose, e.g., `PopulationGroupResearch`, `AgeCategoryResearch` or `GenderCategoryResearch` and the request purpose does not match (⊈) it or the request

[26] Permissions and prohibitions for complex legal structures such as subsidiaries or group of companies cannot be accurately represented using equality (=) or subset (⊆) relations. We, therefore, use the equivalence relation (≡) to indicate the request entity should satisfy the legal interpretation of equality – defining which is outside the scope of this article.

Table 2
Examples demonstrating the matching process

DUO term	Constraint	Offer	Rule	Request	Decision	Reason
GS	Location	Spain	Permission	Europe	DENY	*Europe* $\nsubseteq$ *Spain*
GS	Location	Europe	Permission	Spain	GRANT	*Spain* $\subseteq$ *Europe*
GS	Location	Spain	Prohibition	Europe	DENY	*Europe* $\cap$ *Spain* $\neq \emptyset$
GS	Location	Europe	Prohibition	Spain	DENY	*Spain* $\cap$ *Europe* $\neq \emptyset$
GS	Location	UK	Prohibition	Spain	GRANT	*Spain* $\cap$ *UK* $= \emptyset$
GRU	Purpose	HMB	Permission	DS-Cancer	GRANT	*DS-Cancer* $\subseteq$ *HMB*
GRU	Purpose	DS-Cancer	Prohibition	HMB	DENY	*HMB* $\cap$ *DS-Cancer* $\neq \emptyset$

purpose matches it but the group does not ($\nsubseteq$), e.g. the `PopulationGroup`, `Age` or `Gender` in the request are different from the one in the offer; and

4. offer purpose does not match ($\nsubseteq$) request purpose, e.g., DUO's general research use purpose `GRU` in a request does not match a health, medical or biomedical research purpose `HMB` in an offer as `GRU` is a superclass of `HMB`.

These steps are checked for all prohibitions and permissions of the dataset's offer and if all permissions and prohibitions are satisfied without violations, access to the dataset can be granted. The proof-of-concept demonstration described in Section 5 uses these steps to match an offer with a request policies.

Table 2 presents examples of how the matching process works for permissions and prohibitions in offer with constraints for location and purpose. In a semantic web implementation, the processes for checking equivalence ($\equiv$), intersection ($\cap$), and subset ($\subseteq$) require additional considerations beyond simply using `owl:sameAs` or `rdfs:subClassOf` inferences. For example, to compare *location Spain* with a request for *location Europe* using subset ($\subseteq$) for permissions or intersection ($\cap$) for prohibition requires both locations to be expressed in a manner where such 'hierarchical' or 'set-based' interpretations are possible. In this case, the matching requires interpreting *Spain* is a *narrower concept* or a *subset* of *Europe* – which can be indicated using various relations such as `rdfs:subClassOf`, `skos:broader`, `dct:isPartOf`, or even a property such as `ex:inContinent`. Further complications arise when legal jurisdictions are to be represented, such as *EU* which *Spain* is a member of.

Therefore, an implementation of the matching process has to be cognisant of such cases and be careful when implementing the equivalence, intersection, and subset processes using conventional semantic web interpretations (e.g. `rdf:type` and `rdfs:subClassOf`). We strongly recommend using a standardised vocabulary such as the DPV when declaring both offer and request terms so as to ensure the matching process is accurate and produces the expected correct result. To support consistent application and interpretation of the standardised vocabulary, a specification is needed that clarifies the expression of concepts and their interpretation within the matching process – for example to indicate that any location term in an offer or request *MUST* be an instance of `Purpose` and *MUST* be related to at least one concept in the purpose vocabulary using `rdf:type` or `rdfs:subClassOf`. Using such a specification, the matching process can then function by relying on these assertions to interpret the constraints.

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.

Table 3
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

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` is the subclass of `odrl:Party`). This utilised the approach from existing work regarding extending ODRL concepts for GDPR [27]. 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 3 describes the performed alignment[27] between ODRL and DPV concepts to define DUO concepts.

Using DPV enables 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 the 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[28] extension which provides these concepts. Through this separation (between DPV and DPV-GDPR), the policies can be declared in a 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 the EU. This translates to knowing the locations of people whose data is being offered for use as well as the requesting entity location.

Using DPV, both of these can be expressed using the appropriate Entity concepts and *dpv:hasLocation*. This enables expressing using ODRL 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 these as permissions or prohibitions to be satisfied when matching a request with a dataset by using a

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

```
PREFIX dpv: <https://w3id.org/dpv#>
PREFIX dpv-legal: <https://www.w3id.org/dpv/dpv-legal#>
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-legal:EU-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

matching algorithm as in Section 3.6 along with an encoding of GDPR's requirements such as those from Vos et al. [27]. The DPV-LEGAL[29] extension providing Jurisdictions, Laws, and Authorities for DPV is helpful in representing these conditions.

5. Demonstration and evaluation using a proof-of-concept

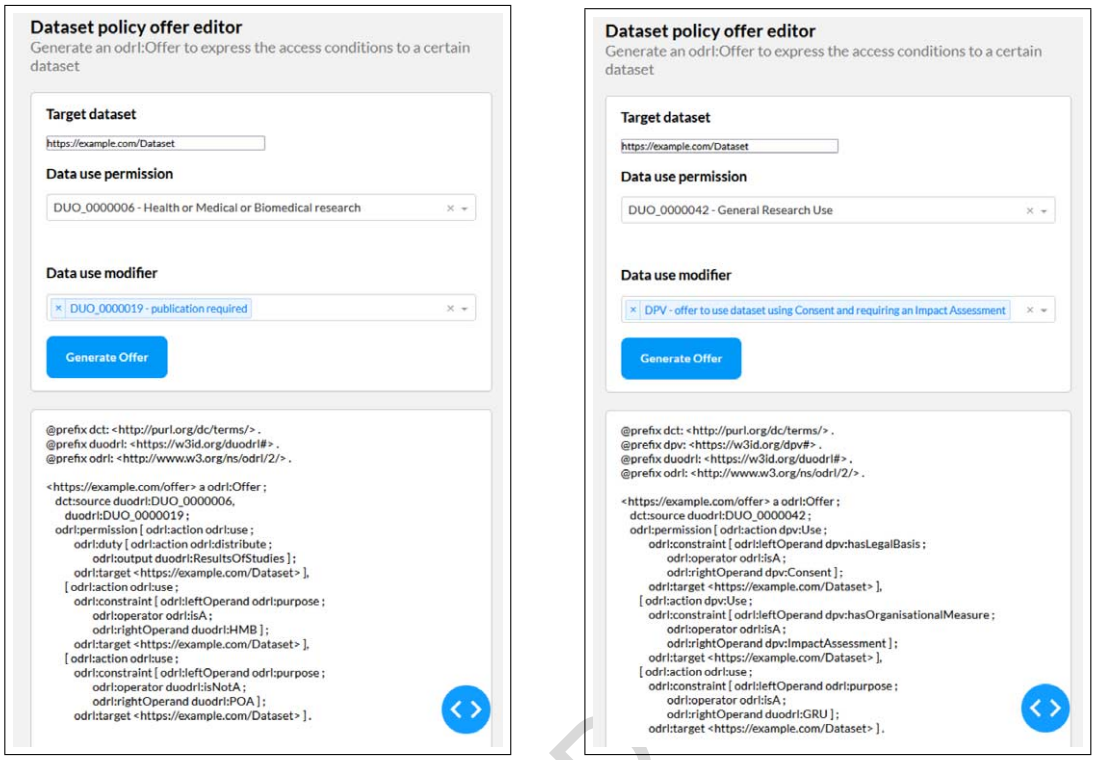
In this section, we describe the implementation of a User Interface to generate dataset policies and a prototype implementation of the matching algorithm is available at https://w3id.org/duodrl/demo/. The prototype can be installed and used locally and an online demonstration is available at https://w3id.org/duodrl/app/.

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.[30] 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 on screen. The code and data used in this are available online at https://w3id.org/duodrl/repo. The matching algorithm used here was adapted from prior work in utilising ODRL for GDPR-based policy matching [7]. We modified it as per the requirements elicited in Section 3.

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

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



(a) from DUO concepts (b) from DUO and DPV concepts

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

For the matching process, the conditions represented in an `odrl:Request` instance should be compatible with 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 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 the policies of every dataset. This process can be made faster and more 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.

To record the result of the matching algorithm, an `odrl:Agreement` is created with a permissive or prohibitive rule to indicate the case where `odrl:Request` is allowed or denied. This agreement policy also includes the date where the agreement was created and/or reached, using the `dct:dateAccepted` property, and the `dct:references` property to indicate the identifiers of the `odrl:Offer` and `odrl:Request` used to reach the result of the matching agreement. The depositor of the dataset is added as the `odrl:assignee` and the requestor as

the `odrl:assigner` of the agreement. When the request is allowed/denied due to a particular purpose, this purpose is recorded as a constraint of the permission/prohibition in the agreement. If further constraints are specified in the associated `odrl:Offer` and are allowed/denied, these are also included in the permission/prohibition of said agreement. In addition to this, if the agreement in question has a permissive result, any duties that might be present in the related `odrl:Offer` are copied to the agreement as duties to be fulfilled for the allowed access to the dataset. In this manner, the `odrl:Agreement` instance represents and can be used to create suitable legal documents to document and communicate the agreement based on the issued request. The computational cost to generate such an agreement is not in the scope of analysis of this work as it is highly dependent on the use case and on the interpretation of the DUO concepts/rules, which DUO does not explicitly provide, hence reflect the authors' own interpretation of such concepts. Therefore, this work focuses on representing the information in an explicit manner through the usage of ODRL policies, which can be made more efficient by using existing reasoning approaches, such as by converting it to specific OWL forms that permit efficient reasoning [3].

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 the expression of policies and the implementation of automation in dataset discovery and request matching processes, as well as the inability to further use this information in other processes such as to keep records and create documentation. 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 [14] 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), DUO gains additional 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, using validation and other forms of automation to ensure it is complete and 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 the extent of what and how they wish to utilise our suggestions – such as 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 development and 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 outweigh the cost of making these changes. In our opinion, our changes do offer more advantages than disadvantages in the 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 complementary 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 further research into the use of automation and logic-based reasoning can be investigated 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) follow-up 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 cooperation between adopters and facilitating cross-boundary data requests and access as envisioned by GA4GH as well as the EU's Health Data Space ambitions.

6.3. Assisting with legal compliance

Currently, DUO or GA4GH do not provide information on how the use of its efforts relates 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 requester (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 the fulfilment of legal obligations based on utilising the existing state of the art.

7. Analysis of recent developments

Two articles [5,13] relevant to DUO were made available online during the reviewing of this article. To better position our work, we provide an informative preliminary analysis of their contributions and discuss the (continued) relevance of our work given these new developments.

7.1. Summary of new articles

The two articles [5,13] together represent improvements to the way information is expressed and encoded as rules based on DUO terms. The first [5] presents *Common Conditions of Use Elements* (CCE) – a controlled vocabulary representing *concepts* for use in data sharing policies. The second [13] presents *Digital Use Conditions* (DUC) – a policy expression mechanism to specify *rules* regarding conditions for sharing and reuse of datasets.

Where DUO terms singularly represent both concepts and rules, CCE and DUC distinguish between information and rules, which provides flexibility in their use, enables granularity in their respective uses, and provides a mechanism to extend them via *profiles* to suit specific use-cases and requirements. An online tool demonstrating this is provided at https://ducejprd.le.ac.uk/.

The CCE vocabulary consists of 20 concepts that were identified from an analysis of requirements and conducted user studies. Its motivation is to provide "*flexible ontologies that can capture complex and conditional permissions in data in a manner that enables logical computer-based reasoning*" [5]. The article describes four requirements for CCE concepts:

1. atomic i.e. each term should represent a single *concept* as opposed to representing a complex or combination of several concepts;
2. no directionality i.e. the term by itself should not specify whether its usage means data reuse or sharing is allowed, forbidden, or obligatory;
3. generalised i.e. the term should be a modular category without any customisation, conditionality, or dependencies; and
4. the term should be "widely applicable and relevant".

The DUC specification [13] defines the expression of policies where each policy contains an optional header section providing metadata regarding the policy and a necessary core section containing one or more statements. The header section also provides references to the datasets associated with the policy, and information on the interpretation of 'unstated conditions' as either 'forbidden' or as 'permitted'. Each DUC statement contains four components:

1. a condition term, which is a CCE concept;
2. a rule, which is one of *Obligatory, Permitted, Forbidden*, and *No Requirement*;
3. a scope, which is either '*Whole of asset*' by default or '*Part of asset*'; and
4. optionally a condition parameter with an optional value.

An example mentioned in the article [13] of a DUC statement is *Country* (condition) with *Permitted* (rule) for *Whole of Asset* (scope) with *UK* (parameter). Another example mentioned is *Time limit* (condition) with *Obligatory* (rule) for *Whole of Asset* (scope) with *Month* (parameter) as *12* (value). Additional examples are available within the online tool documentation.[31] In addition to concepts, DUC statements can also contain free text descriptions to represent concepts, rules, scopes, and parameters. The serialisation of DUC policies is expressed using JSON in both the article and the tool website and documentation.

[31] https://ducejprd.le.ac.uk/assets/documents/Examples_of_DUC_profiles.pdf

The article [5] provides a mapping between the 20 CCE concepts and DUO terms and indicates whether the CCE term is exactly equivalent to a DUO term, or requires additional use of rules (e.g. obligations using DUC) to be considered equivalent, or is a combination of multiple rules (e.g. permission with one concept and 'forbidden' with another) to match a DUO term, or has no corresponding term in DUO.

7.2. Comparisons with work presented in this article

The advancements presented in [5,13] address similar motivations as those we discussed in Sections 1.2 and 2 regarding making information inherent within DUO terms explicit and machine-readable form. The key difference in the approaches is that while we focused on the reuse of existing approaches (ODRL as a standard for rules and DPV for vocabulary), the CCE/DUC approach creates a new vocabulary (CCE) and rule expression mechanism (DUC). In this section, we compare the two approaches based primarily on the distinctions between DUC and ODRL for expressing rules and policies, and between CCE and DPV for providing vocabularies.

Without a formal specification, it is difficult to compare DUC with ODRL. The way DUC is described and used in the examples and within the tool provides the perception that DUC can be a simplified subset of ODRL. This is because the structure of a DUC statement can be expressed in the form of RDF triples which can be grouped together within an ODRL rule. In this, the DUC *concept* is mapped to ODRL's *action*, DUC *rule* to ODRL's *Rule*, DUC *scope* to ODRL's *target*, and DUC parameter and value as ODRL *leftOperand* and *rightOperand*, respectively. For example, the two examples of DUC statements stated earlier are equivalent to the ODRL rules presented in Listing 6.

In this mapping, DUC concepts being mapped to *odrl:action* is inaccurate regarding the context of information as some of these concepts do not represent actual *actions*. For example, while concepts such as *Collaboration* and *Research Use* can be considered actions, others such as *Time Period* and *Regulatory Jurisdiction* are not compatible with the definition of an action. This can be reconciled by treating these concepts as a *constraint* rather than *action*, or by treating all concepts as a *rightOperand* in a constraint with the *leftOperand* being their defining context, such as *Purpose* for *Research Use*.

The DUC rules map exactly to ODRL rules as follows: DUC Obligatory is an ODRL Obligation, DUC Permitted is an ODRL Permission, DUC Forbidden is an ODRL Prohibition, and DUC No Requirement does not have an ODRL equivalent. Of these, the mapping of 'No Requirement' is problematic since it is not possible to interpret "no requirement" in a deontic sense by itself. In DUC while the header can specify a default interpretation which is equivalent to an ODRL permission or prohibition, ODRL does not support such interpretation frameworks and the article also does not mention such use of the DUC header. Further, an example from policies specified in [5] mentions "Collaboration with No Requirement" with the interpretation that "The collaboration is evaluated when appropriate". We find this interpretation to be unclear in terms of whether collaboration is permitted, or prohibited, or its interpretation is *deferred* and cannot be clearly stated. In the last case, it may be possible to express this in

```
# DUC statement: Country with Permitted for Whole of Asset for UK
ex:Policy odrl:permission [
    odrl:target :WholeAsset ; # using DUC terminology
    odrl:constraint [
        # odrl:spatial or dpv:hasJurisdiction offer specific interpretation of "location"
        odrl:leftOperand dpv:hasLocation ;
        odrl:operator odrl:eq ;
        odrl:rightOperand :UK ] ].

# DUC statement: Time limit with Obligatory for Whole of Asset for Month as 12
ex:Policy odrl:obligation [
    odrl:target :WholeAsset ; # using DUC terminology
    odrl:constraint [
        odrl:leftOperand odrl:dateTime ;
        odrl:operator odrl:eq ;
        odrl:rightOperand "P12M"^^xsd:duration ] ]. # can also use Time ontology here
```

Listing 6. ODRL Rules for DUC statements regarding country permission and time limits

ODRL as a Permission with a Duty to obtain prior approval to express that collaboration is a possibility. In any case, we highlight the need to provide clarity regarding the implications of such rules as this is necessary in request matching algorithms.

The CCE vocabulary, which contains 20 concepts, specifies the *atomicity* of its concepts as a core requirement. However, we find that the concepts can be further generalised and structured into a taxonomy based on their implied context – such as DPV's distinction between purposes, processing, or technical measures – as terms that have *legal meaning*. For example, we utilised the DPV concepts which match the terminology of regulations and GDPR by expressing DUO concepts as purposes, processing operations over data, location of operations, entities, and technical measures to safeguard data. With DPV we provided a rich taxonomy for each of these concepts, and also gained the ability to express jurisdictional concepts such as GDPR-defined explicit consent instead of just 'consent'.

In comparison, CCE concepts are still similar to DUO concepts in that they contain hidden implicit information (e.g. Clinical Care Use and Research Use are both Purposes), are not 'complete' in terms of balancing the concepts (e.g. Commercial Entity is defined but Non-Commercial Entity is not), and have not incorporated GDPR requirements sufficiently (e.g. CCE concepts do not address information security). In contrast, our use of ODRL and DPV addresses each of these limitations. For example, Section 4 shows the use of GDPR terminology within an ODRL policy. A mapping of CCE terms to DUO concepts is provided in [5].

Comparing it with our mapping of DUO concepts to DPV/ODRL concepts in Table 1, there are similarities in the approach and analysis in that both express DUO concepts as a combination of *Concept* + *Rule*. In continuation of this, we provide a mapping of DUC terms with relevant DPV/ODRL concepts in Table 4. In this, we focused on identifying the core DPV concepts for each CCE term and how it is applied with an ODRL rule (where relevant). From this, we can state that the work presented in this article is compatible with the development of CCE terms and that the use of DPV/ODRL terms to represent policies and matching processes based on DUO concepts is therefore also applicable to the use of CCE and DUC.

Table 4
Mapping of CCE terms to DPV/ODRL concepts

CCE Term	DUO Term	DPV/ODRL Mapping
Use As Control	Research Control	dpv:Purpose
Clinical Research Use	Biomedical Research	dpv:Purpose
Disease Specific Use	Disease Category Research	dpv:Purpose
Geographical Area + Permitted	Geographical restriction	dpv:Location
Research Use + Permitted	General research	dpv:Purpose
Clinical Care Use + Permitted	Clinical Care Use	dpv:Purpose
Return Of Results + Obligated	Return to database or resource	dpv:Data + dpv:Recipient + odrl:Obligation
Collaboration + Obligated	Collaboration required	dpv:Purpose + odrl:Obligation
Time Period + Obligated	Time limit on use	dpv:Duration + odrl:Obligation
Publication Moratorium + Obligated	Publication moratorium	dpv:Purpose + dpv:Duration + odrl:Obligation
Publication + Obligated	Publication required	dpv:Purpose + odrl:Obligation
User Authentication + Obligated	User specific restriction	dpv:TechnicalMeasure + odrl:Obligation
Ethics Approval + Obligated	Ethics approval required	dpv:OrganisationalMeasure + Obnligation
(Commercial Entity + Permitted) AND (Profit Motivated Use + Forbidden)	Non-commercial use only	dpv:Purpose + odrl:Rule
Fees	None	odrl:compensate
Regulatory Jurisdiction	None	dpv:Jurisdiction
Return Of Incidental Findings	None	dpv:Data + dpv:Recipient
(Re-)Identification Of Individuals Without Involvement Of The Resource Provider	None	dpv:Processing + dpv:isImplementedBy + dpv:Entity + odrl:Constraint
(Re-)Identification Of Individuals Mediated By The Resource Provider	None	dpv:Processing + dpv:isImplementedBy + dpv:Entity + odrl:Constraint

7.3. Concluding remarks

Given that ODRL is an established standard, is extensible through the *profiles* mechanism, and is also under active development – we posit that utilising ODRL and reorienting DUC as a "syntactic sugar" over ODRL can be a better alternative than the development and maintenance of a completely separate rules language. The benefits of basing DUC on ODRL also include access to ODRL's distinction between policies representing offers, requests, and agreements – as demonstrated in Section 3 – which is not mentioned in either CCE/DUC articles. The use of ODRL as we have demonstrated in this article also provides the necessary clarity and alignment with legal compliance (via DPV) – which is necessary for health data reuse, and which the CCE/DUC approach only addresses at a superficial level. Further, we have demonstrated (in Section 3.6) how our approach leads to a matching process that supports all three stages of Offer, Request, and Agreement – in a manner that can be automated for checking completeness (e.g. using SHACL) and correctness (e.g. using reasoners). Neither of the CCE/DUC articles demonstrate how their developments improve (or even clarify) the matching process first mentioned by DUO. Finally, with CCE being the vocabulary for DUC concepts, we find it limited in comparison to DPV based on the lack of a structured taxonomy for organising the concepts, fewer terms, no representation of jurisdictional or GDPR-specific concepts, and lack of an extension mechanism.

Based on these, our preliminary analysis concludes that while the work presented in [5,13] is an advancement over DUO regarding separation of information from rule expression, we believe several limitations still exist regarding the modelling of this information. Specifically, the issues regarding how CCE concepts are structured and their limited vocabulary, the development of DUC as yet another rule language without compatibility with existing standards such as ODRL, and difficulty in extending this language beyond its current capabilities – such as to other health use-cases. We find the contributions provided in this article are still of relevance and have valid contributions that can be applied to further develop DUO, CCE, and DUC while maintaining the existing adoptions.

8. 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 are 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) Demonstrating their use in 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 the automation of tasks regarding matching requests with offers and creating documentation 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 feasible, 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. We also thank the reviewers – both named (Jaime Delgado, Arianna Rossi, Visara Urovi) and anonymous – for assisting us in refining this work and its presentation.

Both authors have contributed equally to this work.

References

[1] 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.

[2] 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, in: *AMIA Annual Symposium Proceedings 2021*, 2022, pp. 197–206, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8861693/.

[3] 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.

[4] 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.

[5] M. del Carmen Sanchez Gonzalez, P. Kamerling, M. Iermito, S. Casati, U. Riaz, C.D. Veal, M. Maini, F. Jeanson, O.M. Benhamed, E. van Enckevort, A. Landi, Y. Mimouni, C. Le Cornec, D.D. Coviello, T. Franchin, F. Fusco, J.A. Ramírez García, L. van der Zanden, A. Bernier, M. Wilkinson, H. Mueller, S.J. Gibson and A.J. Brookes, Common conditions of use elements. Atomic concepts for consistent and effective information governance, *Scientific Data* (2023), (in peer review), https://zenodo.org/record/8200079.

[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] 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:10.1109/EuroSPW54576.2021.00038.

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

[9] 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.

[10] 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.

[11] R. Iannella, M. Steidl, S. Myles and V. Rodriguez-Doncel, *ODRL Vocabulary & Expression 2.2* (2018), https://www.w3.org/TR/odrl-vocab/.

[12] 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.

[13] F. Jeanson, S. Gibson, P. Alper, A. Bernier, P. Woolley, D. Mietchen, A. Strug, R. Becker, P. Kamerling, M.d.C. Sanchez Gonzalez, N. Lynne-Mah, A. Novakowski, M. Wilkinson, O. Benhamed, A. Landi, G.P. Krog, H. Müller, U. Riaz, C. Veal, P. Holub, E. van Enckevort and A.J. Brookes, *Getting Your DUCs in a Row – Standardising the Representation of Digital Use Conditions*, 2023, https://zenodo.org/record/8200044.

[14] T. Kuhn, A survey and classification of controlled natural languages, *Computational Linguistics* **40**(1) (2014), 121–170. doi:10.1162/COLI_a_00168.

[15] 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.

[16] 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.

[17] 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.

[18] 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, Cham, 2019, pp. 270–282. ISBN 978-3-030-21348-0. doi:10.1007/978-3-030-21348-0_18.

[19] H.J. Pandit, K. Fatema, D. O'Sullivan and D. Lewis, GDPRtEXT – GDPR as a linked data resource, in: *European Semantic Web Conference*, LNCS, Springer, Cham, 2018, pp. 481–495, 978-3-319-93417-4. ISBN 978-3-319-93416-7. doi:10.1007/978-3-319-93417-4_31.

[20] 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.

[21] T. Pellegrini, V. Mireles, S. Steyskal, O. Panasiuk, A. Fensel and S. Kirrane, Automated rights clearance using Semantic Web technologies: The DALICC framework, in: *Semantic Applications*, T. Hoppe, B. Humm and A. Reibold, eds, Springer, Berlin, 2018, pp. 203–218, 978-3-662-55433-3. ISBN 978-3-662-55432-6. doi:10.1007/978-3-662-55433-3_14.

[22] E.U. Regulation, 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 L* **119** (2016), http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:L:2016:119:TOC.

[23] 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.

[24] 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:10.1016/j.eswa.2019.04.009.

[25] 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), giz165. 693–701. doi:10.1038/s41576-019-0156-9.

[26] 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.

[27] M.D. Vos, S. Kirrane, J. Padget and K. Satoh, in: *ODRL Policy Modelling and Compliance Checking, in: 3rd International Joint Conference on Rules and Reasoning (RuleML+RR 2019)*, Bolzano, Italy, 2019, p. 16. doi:10.1007/978-3-030-31095-0_3.

[28] 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, 2018, pp. 31–45. ISBN 978-3-662-55433-3. doi:10.1007/978-3-662-55433-3_3.

[29] 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:10.1038/s41525-017-0040-5.