# Provenance, Anonymisation and Data Environments: a Unifying Construction

Muhammad Aslam Jarwar[1], Adriane Chapman[2], Mark Elliot[1], and Fatemeh Raji[2]

[1] University of Manchester, Manchester M13 9PL, UK
{aslam.jarwar, mark.elliot}@manchester.ac.uk
[2] University of Southampton, Southampton, SO17 1BJ, UK
{adriane.chapman, f.raji}@soton.ac.uk

**Abstract.** The Anonymisation Decision-making Framework (ADF) operationalizes the risk management of data exchange between organizations, referred to as "data environments". The second edition of ADF has increased its emphasis on modeling data flows, highlighting a potential new use of provenance information to support anonymisation decision-making. In this paper, we provide a use case that showcases this functionality more. Based on this use case, we identify how provenance information could be utilized within the ADF framework, and identify a currently un-met requirement which is the modeling of *data environments*. We show how data environments can be implemented within the W3C PROV in four different ways. We analyze the costs and benefits of each approach, and consider another use case as a partial check for completeness. We then summarize our findings and suggest ways forward.

**Keywords:** Data Environment Representation · Anonymisation Decision-Making Framework · Data Provenance · W3C PROV-DM

## 1 Introduction

In the knowledge economy, large amounts of data are collected to support decision-making, policy analytics, service delivery etc. However, the usability of these data is constrained by the disclosure risks involved in data processing in general and data sharing in particular. One of the important tools used to mitigate this risk is anonymisation. The Anonymisation Decision-Marking Framework (ADF) operationalises the processes of functional anonymisation [1]. This conceptualisation originated in the work of the *data environment analysis service* [2]; a

support system for the 2011 UK census focused on data confidentiality and disclosure control [e.g. 3, 4, 5] and, in particular, re-identification risk assessment [e.g. 6, 7, 8]. The critical point underlying this concept is that disclosure risk resides not in the data themselves but in the relationship between the data and their environment. Mackey and Elliot define the data environment as "the set of formal and informal structures, processes, mechanisms and agents that either: (i) act on data; (ii) provide interpretable context for those data or (iii) define, control and/ or interact with those data" [9].

Data environments come in a variety of types. For example, the open data environment, an end-user license management data environment, restricted access secure data environments etc. Notwithstanding this variety, the ADF framework assumes that all data environments can be described through four descriptive features: other data, agents, infrastructure, and governance.

It follows from the foregoing that in order to apply the appropriate anonymisation processes, one needs to take account of both the data and their environment. Elliot et al. [10] developed the ADF to operationalise exactly such a process. The ADF emphasises that the appropriate anonymisation decisions for a given set of data are only be possible by considering the relationship between the data and their environment(s) which they call the *data situation*.

***Problem statement.*** *Data situations are often dynamic in that data move between environments for both processing and use. Thus, understanding contextual risk, and how to manage that risk through anonymisation, requires an awareness of, and capacity to map, the data flows between environments.*

Currently, capturing and mapping *data situations* for analysis within the ADF framework is done manually, which is labor intensive and prone to errors. In order to automate this mapping, we propose the use of data provenance - a concept that is already mentioned in an informal sense in the ADF. By integrating provenance with the ADF, we will be able to track the flows of data and recognise the upstream and downstream data situations - both existing and proposed. We note that, data provenance has already been applied in the modelling of similar problems such as situation awareness and decision making [11], controlling of direct and indirect data flows [12], big data security and privacy [13].

W3C PROV is a standard for provenance interoperability and for representing where data came from, and how it has been processed [14, 15]. PROV provides an abstract data model that includes agents, entities, activities, and relationship

properties and which enables the representation of the provenance of data and systems.

A critical element in the feasibility of linking provenance to the ADF is the representation of data environments. In the W3C PROV data model, two constructs *bundles* and *namespaces* might be considered to be candidates for such representation. In this paper, we examine the potential value of both of these solutions. We also consider how the elements of PROV (i.e. Entity, Bundle, Agent, Activity) could be used to represent data environment features (agents, other data, infrastructure, governance). We observe that there are limitations to representing data environments in this way and suggest some modifications which would enable full capture of the desired features.

The contributions of this work are as follows:

1. We outline – using an ADF use case – the requirements for provenance in the representation of data environments (in section 2).
2. Using these requirements, we propose four different approaches to apply and extend W3C PROV to enable the representation of data environments for machine enabled reasoning (in section 3).
3. We then analyse the four approaches (in section 4)

## 2 An ADF Use Case

A seemingly simple data flow between environments can in fact be complex depending on the nature of the data and the environment(s), the intended data use and the responsibilities of the data situation's stakeholders. When data moves between environments (called a *dynamic data situation* in ADF parlance), each environment produces a different risk profile, depending upon how the data interacts with the four defining features (other data, governance, infrastructure and agents). Below we describe an example use case. This example, drawn from [10], is an idealisation of a common data situation; the sharing of data held by a national statistics agency with a research data service.

---

**The set up:** the Government Office for National Data (GOND) collects several types of national level datasets. For example, national census data, public healthcare data, birth-death related data, pupil data from schools, traffic data from the smart cities sensors, etc.

- Part of GOND's remit is to make available some of those datasets for secondary research use. In service of this, it shares versions of the national datasets that it holds with the National Research Data Service (NRDS).
- The NRDS is part of University of Barsetshire. The NRDS's role is to acquire data from data holders, including GOND, under contract and then enable (and manage) access to those data under controlled conditions by researchers from research laboratories across the country.
- GOND also releases highly aggregated data into the public domain (by definition an open environment).
- The researchers carry out data analysis on GOND's data and then publish papers reporting on this analysis in the public domain.
- This data flow involves various loci of responsibility and control (key concepts in the ADF) over the data sharing in and from the different environments:
  - GOND has *indirect responsibility* and *strategic control* over the data released from the NRDS environment into the open environment (in the form of analytical output within publications). GOND also has direct responsibility and control over the data released from its own environment into the public domain (in the form of aggregate statistics).
  - NRDS's responsibility and control are different from GOND's, NRDS has *direct responsibility* and *operational control* over the data release from the output of publications.[1]

---

The sketch diagram of this use case is shown in Figure 1. Four focal data environments are part of global data environment. GOND, the University of Barsetshire, and NRDS are represented as data environments 1, 2, and $2_a$ respectively. The research labs and the open environment are labelled with data environments $3_n$ to $3_{n+1}$ and 4 respectively. As shown in Figure 1:

1. For the purposes of understanding this data situation the origin of the data flow is the GOND (1) data environment.[2]. At $t_1$, the data are processed to make them compliant for sharing with (2), according to contractual obligations. At $t_2$, in parallel, the data are processed more heavily for public release into the open environment (4).

[1] See [10] for a more detailed discussion of the concepts of responsibility and control.

[2] The same questions of granularity and scope affect the anonymisation use case as other uses of provenance information. In some instances, one may want to push the flow all the way back to the data subjects. For simplicity's sake here we are assuming that GOND are the origin.

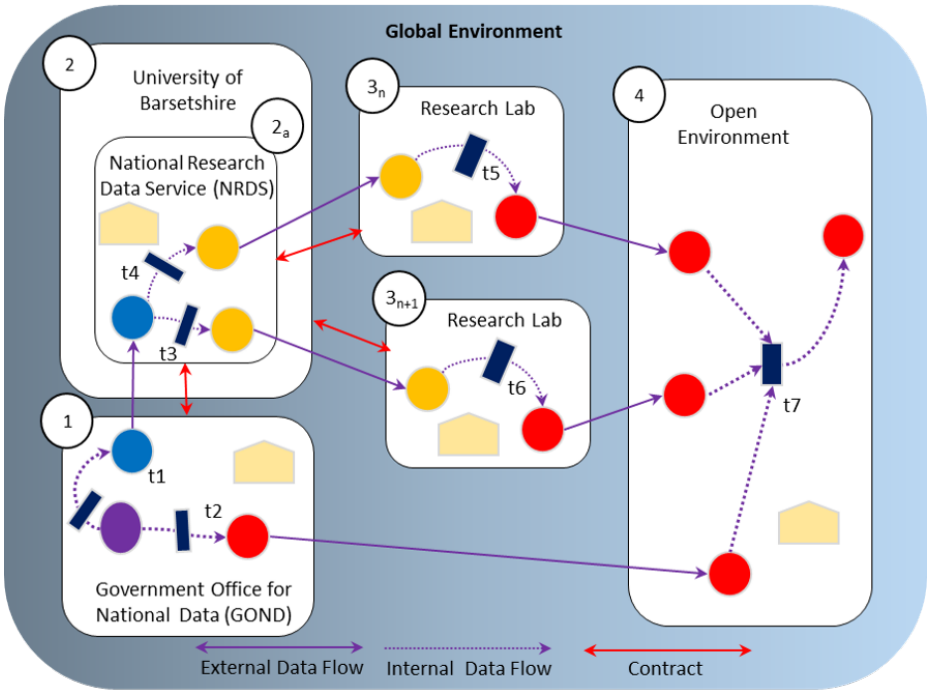


Fig. 1: A use case of data flows between and within multiple data environments. The red arrows indicate contractual agreements. The blue lines indicate data flow. Data environments are indicated by rounded rectangles, a circle represents a piece of data, a rectangle represent a process and a pentagon represents a user (in the data environment). The time for processing and sharing of data in the environments are labelled from $t_1$ to $t_7$.

2. The data that is shared from GOND to the NRDS (2a), might be subjected to additional processing (disclosure controls) so that they can be shared with the various research labs who want to access the data for substantive analyses $(3_n, 3_{n+1}, ....)$.
3. Each research lab analyses the data according to their particular needs and research questions. The research labs wish to produce publications and research datasets for public consumption (4).
4. One of the goals of the ADF is to ensure that when data that has been derived from the same original data, are released by different organisations (or indeed at different times by the same organisation), inadvertent disclosures of personal information do not happen as a consequence. This is an increasingly critical issue which this data situation epitomises.

### 2.1 The Provenance Requirements of the ADF (using the GOND-NRDS use case)

The next step in understanding the relationship between provenance information and anonymisation is to produce a set of representational requirements. Based on these requirements, a data environment formalism will be created using the W3C PROV data model (PROV-DM).[3] A specification of those requirements is as follows:

***R1: The data environment construct***
The data environment construct defines a boundary state that contains data. For example, GOND and NRDS are two closed data environments containing different data and within which different processing events occur.

***R2: Data environments within data environments***
Sometimes an environment will contain other environments. For example, data flows between an organisation's sub-units for processing, auditing, etc. Another example is that the NRDS data environment is contained within the Barsetshire University data environment. In general, access control will be tighter in sub-environments than the host environment.

***R3: Attaching attributes to data environments***
To determine appropriate disclosure (control) practices, the purpose of data collection, type of data environment and any constraints and features (infrastructure and governance) of a data environment need to be recorded as attributes of that data environment. For instance, GOND collects data from its partners for use and onward sharing via a legal gateway; the processing occurs in a restricted access data environment the parameters of which may be defined by - for example - a data sharing agreement, GONDS own data policies, the enabling legislation itself etc.

***R4: Relationships between data environments***
This describes the possible relationship from data environment to another data environment. For example, Within the NRDS, a research lab might have a specialised, secure processing environment which is owned and maintained by NRDS, but hosted for and used by the research lab. This is an example of a data environment with more complex relationships between data environment constructs than containment.

***R5: Annotation of relational constructs***

[3] PROV-DM is the conceptual data model and core part of W3C PROV that defines each term used to represent provenance information [16].

In order to reason over data environment interactions, controllers, processors, subjects, users, etc., it is important to allow the attachment of semantic meaning to the relationships between the constructs. For example, NRDS receive data from GOND and store it for onward sharing with researchers. In PROV, this might be achieved by labelling with *prov:use* or *prov:generated* but these labels do not represent all of the required information needed for the ADF.

***R6: Representation of agents, data and processes within a data environment***
Agents might include data controllers, data processors, data users and data subjects. Data includes datasets, reports, etc. Processes include data extractions, sharing, storing, sampling aggregating, etc. In our use case the research labs contain agents (users), a process (analysis), input data for the analysis and output data (e.g. tables, models, graphs).

***R7: Data governance instruments: contracts***
There are numerous types of data governance instruments that affect what can and can't be done with data. One important type is the contract; typically a data sharing agreement to share, exchange and use data between the environments. For example, in our use case GOND share data with NRDS based on the contract between them.

***R8: Access and control (direct and indirect)***
A record of the access and control mechanisms over the data and services. For instance, GOND has a data dissemination function that can be used by NRDS (based on some contract). GOND also has indirect control over data releases from the NRDS environment (in that the output disclosure control policy of NRDS will be defined by GOND).

## 3 Supporting Data Environments with W3C PROV

In this section, we will explain how PROV can be applied to support the data environment representation requirements outlined in section 2. We will show that the existing W3C PROV data model does already support some of the data environment representation requirements in that some of them can be mapped onto PROV elements. However, there are some data environment specific requirements that need extensions in PROV. We will describe four possible mechanisms: namespaces with or without supporting structures and bundles with or without an extension.

### 3.1 Namespaces and support structures

The namespace concept was inspired by the World Wide Web architecture and was designed to make objects interoperable across technologies and platforms [16]. In PROV-DM, Namespaces are a Uniform Resource Identifier (URI); a provenance graph can contain multiple - possibly many - namespaces. The namespace is a candidate for use as an identifier to capture the idea of multiple data environments (including data environments within data environments) and their associated entities, activities, agents, etc. By using Namespaces and prefixes we could differentiate the representation of nested data environments and can access related elements information through namespace concatenating and de-concatenating.

For example, we might refer the University of Barsetshire and NRDS data environments as *http://global-env.com/bu/* and *http://global-env.com/bu/nrds/* respectively. (Note: In the example use case, the NRDS data environment is a part of University of Barsetshire environment). We can also express the control mechanism over the data environments and its elements with namespace features. The visual representation of the GOND-NRDS use case with the support of Namespaces and PROV constructs is shown in Figure 2.

In Figure 2, there are five main data environments with separate namespace. For instance, the GOND data environment can be recognised with namespace *http://global-env.com/gond/*. The elements of GOND such as entity_001 can be accessed with *http://global-env.com/gond/entity_001#*. Similarly the agent with an id "agent_controller_001" from NRDS data environment can be recognised with a *http://global-env.com/bu/bu/nrds/agent_controller_001#*. Additionally, as illustrated Figure 2, the data provenance for research labs can be tracked in forward and backward called as forward and backward chaining. The forward chaining informs how the research labs data will be utilised and backward chaining tracks the sources of data and the contracts between research labs with the data providers. Moreover, the right hand side of Figure 2 shows the pseudo code of attributes attachment with the data environment through namespaces' support. In the pseudo code, the global data environment has two attributes with values *foo* and *bar*.

While namespaces have potential for representing the bounded nature of data environments, and what has occurred within a given data environment and it's sub-environments, namespaces alone are not enough to satisfy all of the requirements identified in section 2. For instance, the attachment of additional attributes to the

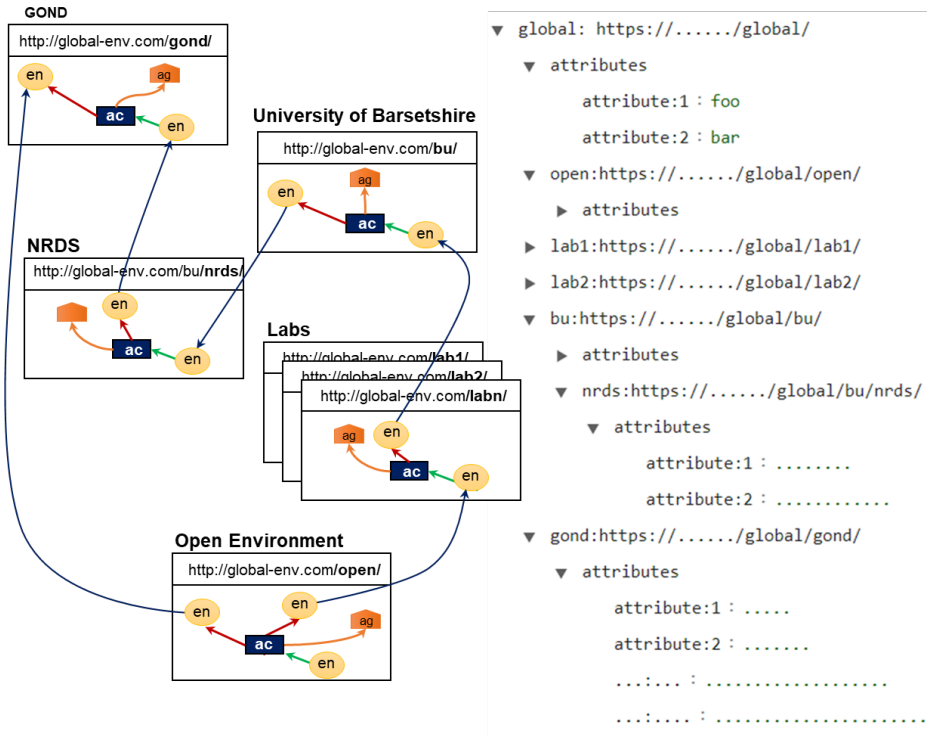


Fig. 2: Illustration of the use of namespaces to represent Data Environments: ag, ac, and en indicates agent, activity, and entity respectively; the right-hand part shows data environments with attribute attachment using namespaces. Relationships across namespaces could be captured in the same manner.

data environment itself and contracts between the data environments cannot be accommodated. Additionally, relationships among namespaces beyond containment cannot be captured. For instance, it is possible in namespaces to distinguish that *http://www.nytimes.com* data environment that contains a sub-data environments related to advertising functions, *http://www.nytimes.com/ads*. However, within our use case, there is more than strict-hierarchical containment. For example, researchers from one of the Research Labs might have a specialised data analysis environment built-by, hosted-by and managed-by NRDS, but considered an enclave of both NRDS and the Research Lab. In this case, namespaces do not capture enough information to represent this relationship.

To solve these issues, an additional set of structures would need to be created. For instance, a separate document which extends namespaces and allows attachment of attributes, could be used.

### 3.2 Bundles and Extended Bundles

In PROV, the bundle concept has some similarities to the data environment construct. The bundle is itself an entity which provides provenance information regarding the creation and modification of a group of entities [17]. For example, a bundle can contain the entities, activities, agents, and the relationships between them. Within a given bundle, the data, and data processes can be represented with entities and activities respectively. Bundles can also support entities with attributes. This can help us to add necessary metadata to the entities. The excerpt view of the GOND-NRDS use case representation as supported by PROV bundles is shown in Figure 3.

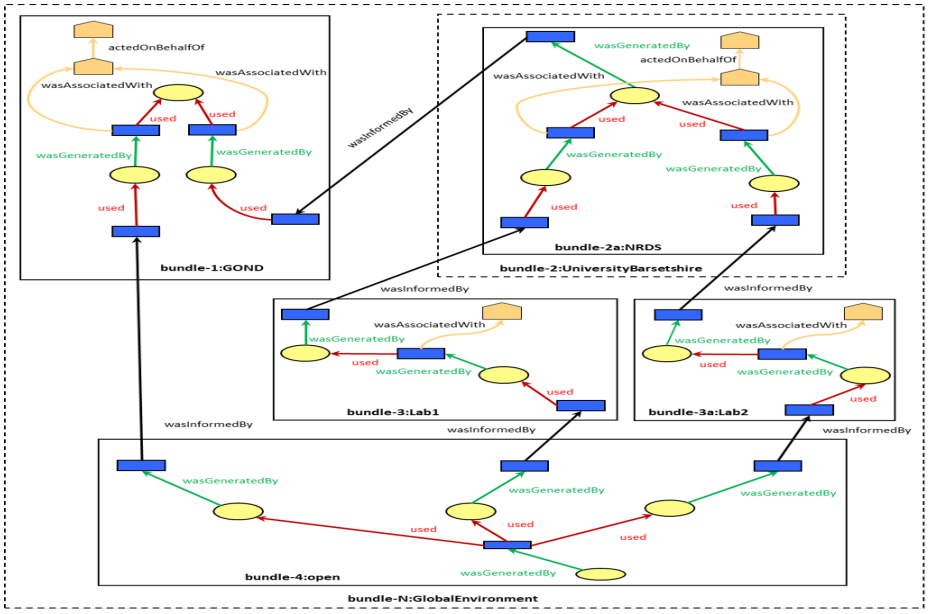


Fig. 3: A representation of the GOND-NRDS use case supported with PROV bundles. Please note that the nested data environments are shown with dotted lines are for illustration of use case and currently these are not supported in PROV.

In Figure 3, the large rectangles delineate data environments each represented as a PROV bundle. Each bundle contains data environment elements (represented as nodes) and relationships between those elements (represented as edges). For example, in the "bundle-1:GOND" data environment, the processes (small blue rectangles) are using a piece of data for generating another piece of data. For these processes a data processor (agent expressed with pentagon) is responsible (the responsibility relationship is shown with "wasAssociatedWith"). The relationship between the data controller and data processor[4] is shown with "actedOnBehalf" property.

The data flow between one data environment and another environment (we can say that from bundle to bundle) is shown with "wasInfromed" property. For instance the data flows in direction of bundle-1:GOND->bundle-2a:NRDS-> (bundle-3:Lab1, bundle-3:Lab2), etc is represented with "wasInfromed" property.

We can also see in Figure 3 that the NRDS data environment ("bundle-2a:NRDS") is a sub environment of University of Barsetshire (bundle-2:UniversityBarsetshire). We note that in ADF terms, NRDS is said to have *direct control* over the labs environment for releasing of data, whereas GOND has *indirect control*. To support the representation of control (and its companion concept of responsibility) would need additional mechanisms to be added to PROV but this lies outside of the immediate scope of this paper.

W3C PROV constructs were designed to be extensible [16]. In previous work, PROV has been extended to express the provenance of big data security supervision [13], provenance access control [18], data privacy protection based on GDPR using provenance [19] and managing mutable entities by adding reference derivations and checkpoints [20]. Likewise, we can extend the existing structure of PROV bundles in order to support and express the requirements of ADF with more flexibility. For example, by extending we can attach additional metadata to the bundle construct, which would enable us to define the different types of data environments. Another extension that we would need in PROV Bundles, is support for nested data environments.

[4] The terms data processor and controller are key terms in the General Data Protection Regulation (GDPR). See https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/controllers-and-processors/what-are-controllers-and-processors/ for definitions.

# 4 Comparative Analysis

## 4.1 Requirements Completeness

To test the specification of the provenance needed for the ADF define in section 2, we analyse a second ADF use-case (please refer Figure 4) in which:

- Data from clinical trials are generated at several participating centres.
- The data are uploaded electronically by the participating centres to a company called Capturedata which offers an electronic data capture and management system for the pharmaceutical industry.
- PharComp, a European pharmaceutical business, extracts and downloads the clinical trial data from the Capturedata database onto PharComp systems for analysis.
- PharComp shares some of the data with researchers, for use in public health research.
- Researchers publish their analysis in journal articles in the public domain. These data will not include information that directly identifies the patients, and additional steps are taken to safeguard the patients confidentiality.
- Explicit consent has been given by trial participants for secondary research using of anonymised versions of their data.

We use this use case to confirm that the requirements identified in Section 2 are correct (they are sufficient in description to cover equivalent elements of this use case) and complete (there are no additional requirements identified in this use case).

The **data environment construct** is required to fully capture the data situation as encompassing the environments of collection centers, Capturedata, PharComp, research labs and the open environment. The data collection centers and the research labs contain sub-environments for various types of data collection, processing and to meet research protocol requirements, confirming the **data environments within data environments** requirement. In this use case, the Capturedata and PharComp are types of restricted access data environments and can be accessible to only authorised users and researchers. To express this type of and restriction to these data environments we need to **represent data, agents and processes**.

As with the GOND-NRDS use case, here we need to **annotate the relationship constructs** between the data environments of collection centers and Capture-

data, where the data is collected and stored instead of data derivation/usage or generation. This will support the semantic meaning of relationship construct in the data provenance. Given the nature of the data collected for the pharmaceutical company, contracts specifying data collection, exchange and control exist (**contracts**). PharComp has indirect control over the data release in the open environment in the form of publications and the researcher must follow the code and conduct given by the PharComp, this requirement can be represented with **access and control** category as described in section 2.

One of the requirements presented in section 2 the **relationship between data environments** is not found in this use case, as there are no data environments with multiple institutional ownership and use. However, if the specific example contained an enclave in PharComp in which regulatory employees from a government could review specific data, this requirement would be needed. There are no additional requirements that seem necessary to capture the provenance of data situation for the purposes of the ADF.

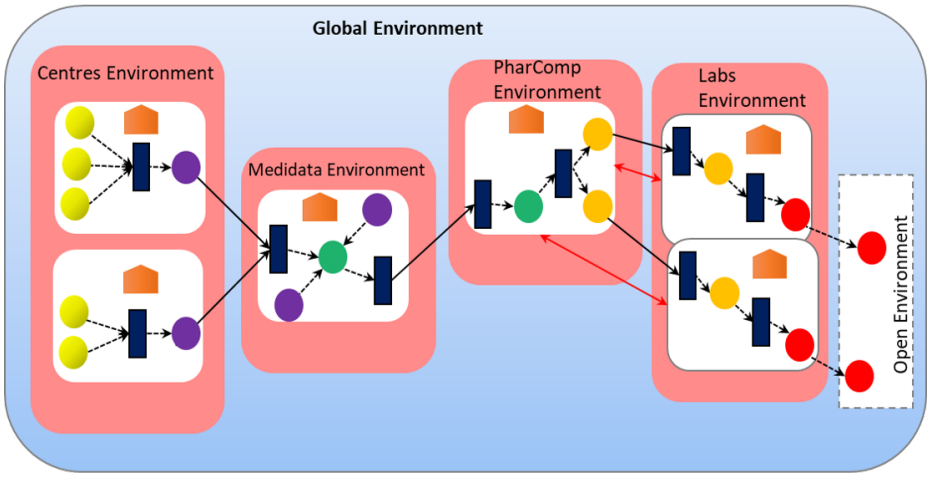


Fig. 4: A Second ADF use case with data environments: The red arrows indicate contractual agreements. The black lines indicate data flow. Data environments are indicated by rounded rectangles, A circle represents a piece of data, a rectangle represent a process and a pentagon represents a user in the respective data environment

### 4.2 Analysis of Implementation Approaches

Table 1 shows the data environment representation requirements outlined in section 2 and the ability of each of the implementation options discussed in Section 3 to meet those requirements.

Table 1: Use case requirements analysis (NB: here Namespace + includes attributes and PROV constructs

| Representation requirements | Support | | | |
|---|---|---|---|---|
| | Bundle | Namespace | Namespace+ | Bundles+ |
| Data Environment Construct | ✓ | ✓ | ✓ | ✓ |
| Data Environments within Data Environments | - | ✓ | ✓ | ✓ |
| Attaching Attributes to Data Environments | - | - | ✓ | ✓ |
| Relationship between Data Environments | ✓ | - | ✓ | ✓ |
| Annotation to relational constructs | - | - | ✓ | ✓ |
| Representation of agents, data and processes within DE | ✓ | ✓ | ✓ | ✓ |
| Data governance instruments: contracts | - | - | ✓ | ✓ |
| Access and control | ✓ | ✓ | ✓ | ✓ |

The nesting of data environment (i.e. data environments within data environments) is one of the important features. However, with bundles we cannot represent nested data environments because PROV does not allow the nesting of bundles [16].This gap is one of the drivers for bundles+. This requirement is supported by namespaces (and so namespaces+ can also support nesting).

The ability to attach attributes to a data environment is also an important feature for the ADF. Neither bundles or namespaces support attachment of attributes. For example, currently, we cannot express following using PROV bundles:

*bundle (EX-A:GOND,*
*prov:envType ="Government",*
*prov:governance-accessType="Restricted",*
*prov:governance-userdefinition="TrainedLevel2",*

*prov:infrastructure="ISO27001").*

The additional structures provided in Namespace+ do allow attributes to be maintained with the namespace information. However, Bundles+ is a more elegant option; using the W3C PROVs standard of attaching attributes to other object types, but expanding that notion to bundles+.

As we observed in the GOND-NRDS use case (see Figure 1), the GOND data environment contains the representation of collected, processed and shared data along with the data processes, agents, and contracts (i.e. contract with the NRDS), and IT infrastructure and services. In order to create the provenance graph for this data situation, the relationships between these elements would need to be supported with PROV properties. For example *wasGeneratedBy (entity_id , activity_id)*, and *used(activity_id, entity_id)* properties could be used to represent the relationship between the GOND collected data and processing of the data to generate the new dataset for NRDS.

Both the bundles and namespaces solutions could naturally support the representation of agents, processes and entities using native W3C PROV concepts. On the other hand, supporting additional metadata such as annotation with the relationship constructs is not fully supported in PROV. However, this could be managed by attaching additional attributes with the relationship construct (this approach was used by [20] for the purpose of tracking changes in entities over time). Attaching annotation will also be helpful here in selecting an appropriate disclosure control processes. Bundles+ supports this requirement. The GOND-NRDS contracts are supported by both Bundles+ and Namespaces+. The representation of access control requirement is supported by all four constructs.

### 4.3 Validation of Data Environment Representation

Currently, the source code for PROV validation is not openly available. However, the source code for the SEIS-PROVs[5] document validation is available at [21]. The SEIS-PROV validation mechanism is implemented in python. Using this validation tool as an exemplar, to validate the representation that includes the data

[5] SEIS-PROV is a domain specific extension based on the W3C PROV data model, used in the seismological data processing. This extension defines a new namespace with entities, activities and attributes in the context of seismology.

environments within data environments feature, the PROV document should include a formalism for data environments: $[d_i = I_i \cup d_1, ...., [d_n = I_n \cup d_{n-1}]$ where n is number of data environments and PROV elements instances, I is the top level prov element instance and d is the data environment instance.The value of i will be between 0 and n.

The PROV validation mechanism has two components: inference and constraints. The inference component deals with the fixing of missing information based on the definition of the element defined in the PROV data model. The constraint component includes a checking mechanism that deals with uniqueness, ordering, impossibility and typing. Impossibility checks for prohibited patterns, while the typing constraints check the type of identifier when it is used in relations. Inferencing should be performed over the document, and the elements should be categorised as per the definition. For example, similar entities in two different data environments might be categorised according to the prefixes of definition or prefixes over the data environment.

### 4.4 Translation and Visualisation

In order to share the data environment representation with other stakeholders we may need to support the translation from PROV-N to other formats (e.g. json, provx, turtle, trig, svg, rdf, xml) and vice versa. Therefore, to accommodate the proposed extension to PROV bundles, the existing translation and visualisation mechanism would also need to be updated.

Our goal would be to incorporate the support for the data environment representation the in PROV python implementation. The PROV python implementation provides a PROV serialisation module [22] that provides various classes to transform PROV document from one format to another format. For example, *ProvJSONSerializer, ProvRDFSerializer, ProvNSerializer, ProvXMLSerializer* provides the implementation to translate PROV in JSON, RDF, prov notation and XML formats respectively. All of these serialisation classes would need consequential changes to support the bundle extension.

For graphical visualisation of provenance statements, the PROV python implementation uses three open source libraries pydot [23], Graphviz [24], and DOT language [25] in *prov.dot* module. The *prov.dot* module also needs substantial changes in *prov_to_dot()*, *_bundle_to_dot()*, *_attach_attribute _annotation()*, etc methods that translate the provenance statements into visualisation

graphs. These methods would also need updates along with additional methods to support the graphical visualisation of the complexity of data situations.

## 5 Related Work

W3C PROV has been used elsewhere to capture provenance for the protection of data subjects' confidentiality, and the security of data.

A W3C PROV based provenance model has been proposed by Benjamin et al. [26] that uses the PROV data model ontology and data protection ontology to express the provenance for compliance with the European Union (EU) General Data Protection Regulation (GDPR). The Agent, Activity, and Entity classes from the PROV ontology were extended with sub-classes to express the provenance of GDPR compliance. For example, *Subject*, *Controller*, *Processor*, and *Supervising-Authority* sub-classes were introduced within the agent class. The Activity class was extended with two additional sub-classes: *Process* and *Justify*. Similarly, the Entity class was extended with three sub-classes: *PersonalData*, *Request* and *Justification*. The relationships among the classes were expressed with PROV properties. Both of the ADF examples presented in this work fall under GDPR regulations, and the extensions introduced in Benjamin et al. [26] would facilitate some of the more general requirements of **representation of agents, data and processes** and **contracts** within data environments.

To support provenance of mutable values by time-versioning entities, a PROV extension has been developed by adding the reference sharing and checkpoints feature [20]. These features were built on top of PROV events that track a version of and object or entity through change or generation events (i.e. *prov:Generation*) and access or usage events (i.e. *prov:Usage*). The checkpoint attributes were used with the PROV entities, activities, relationship properties for tagging and tracking of changes in the entities over the time period. For this purpose, two namespaces (i.e. version and script) were created to support the checkpoints mechanism. These were used for both general PROV extension concepts and specific script concepts. However, this approach increases the overhead for querying the provenance graph due to folding and unfolding for adding the checkpoints.

The PROV data model has also been extended with new relationship properties in order to supervise the security of data streaming [13]. These extensions focus on collecting the provenance information about data operations inside and outside of big data clusters; representing the data interaction flow between the

clusters. The harvested relationship provenance information of the graph is analyzed for the detection of anomalies in the data. The anomaly detection and reasoning mechanism checks for inconsistency between the nodes and edges.

Pahl et al. have used the PROV data model along with blockchain technology to implement a trust analysis platform. PROV was used to capture the features for verifying the originality and source of data received from sensors and edge cloud devices.[27],

Finally, PROV has been used to protect data provenance content that is sensitive and subject to disclosure control [18], modelling the threat of attack to supply chain electronic management system [28], and detection of bottlenecks in the system by analyzing the patterns in the provenance graph [29].

## 6 Conclusions and Future Work

In this paper, we have considered a new application of provenance: to support anonymisation of data exchanged across organisations and environments. To this end, we introduce the Anonymisation Decision Framework (ADF) which is used to reason about data flows and anonymisation. Through analysis of the ADF, how it is applied, and the information required to make such decisions, we have identified how provenance might be utilised more formally.

In order to do this effectively, we need to be able to represent one of the core components of the ADF approach the *data environment*, an organising concept constituted from other data, agents, governance processes and infrastructure. We identified the key properties of such environments from an idealisation of a real world use case which can be mapped with W3C PROV elements: entities, bundles, activities, and agents.

We analysed how data environments can be represented within the W3C PROV. We observed that, in order to fully express the features of data environments, the existing PROV constructs are not sufficient and would need extending. We identified four different candidate mechanisms within the W3C PROV, and evaluated each with respect to trade-offs of cost, maintenance and suitability for the problem. While two obviously do not pass muster, the other two are viable solutions, with one *Namespaces+* that utilises existing W3C PROV structures but requires an additional management, and the second *Bundles+* which requires an extension to PROV.

# Bibliography


[1] Mark Elliot, Kieron O'hara, Charles Raab, Christine M O'Keefe, Elaine Mackey, Chris Dibben, Heather Gowans, Kingsley Purdam, and Karen McCullagh. Functional anonymisation: Personal data and the data environment. *Computer Law & Security Review*, 34(2):204–221, 2018.

[2] Mark Elliot, Susan Lomax, Elaine Mackey, and Kingsley Purdam. Data environment analysis and the key variable mapping system. In *International Conference on Privacy in Statistical Databases*, pages 138–147. Springer, 2010.

[3] Leon Willenborg and Ton De Waal. *Elements of statistical disclosure control*, volume 155. Springer Science & Business Media, 2012.

[4] George T Duncan, Mark Elliot, and Juan-José Salazar-González. Concepts of statistical disclosure limitation. In *Statistical Confidentiality*, pages 27–47. Springer, 2011.

[5] Anco Hundepool, Josep Domingo-Ferrer, Luisa Franconi, Sarah Giessing, Eric Schulte Nordholt, Keith Spicer, and Peter-Paul De Wolf. *Statistical disclosure control*. John Wiley & Sons, 2012.

[6] Guang Chen and Sallie Keller-McNulty. Estimation of identification disclosure risk in microdata. *Journal of Official Statistics*, 14(1):79, 1998.

[7] CJ Skinner and David J Holmes. Estimating the re-identification risk per record in microdata. *Journal of Official Statistics*, 14(4):361, 1998.

[8] Chris J Skinner and MJ Elliot. A measure of disclosure risk for microdata. *Journal of the Royal Statistical Society: series B (statistical methodology)*, 64(4):855–867, 2002.

[9] Elaine Mackey and Mark Elliot. Understanding the data environment. *XRDS: Crossroads, The ACM Magazine for Students*, 20(1):36–39, 2013.

[10] Mark Elliot, Elaine Mackey, and Kieron O'Hara. The anonymisation decision-making framework 2nd edition: European practitioners' guide. 2020.

[11] Kenneth Baclawski, Eric S Chan, Dieter Gawlick, Adel Ghoneimy, Kenny Gross, Zhen Hua Liu, and Xing Zhang. Framework for ontology-driven decision making. *Applied Ontology*, 12(3-4):245–273, 2017.

[12] Xie Rong-na, Li Hui, Shi Guo-zhen, Guo Yun-chuan, Niu Ben, and Su Mang. Provenance-based data flow control mechanism for internet of things. *Transactions on Emerging Telecommunications Technologies*, page e3934, 2020.

[13] Yuanzhao Gao, Xingyuan Chen, and Xuehui Du. A big data provenance model for data security supervision based on prov-dm model. *IEEE Access*, 8:38742–38752, 2020.

[14] PROV Data Model. https://www.w3.org/TR/prov-dm/. Last Accessed: 2020-04-20.

[15] Paolo Missier, Khalid Belhajjame, and James Cheney. The w3c prov family of specifications for modelling provenance metadata. In *Proceedings of the 16th International Conference on Extending Database Technology*, pages 773–776, 2013.

[16] Luc Moreau, Paul Groth, James Cheney, Timothy Lebo, and Simon Miles. The rationale of prov. *Journal of Web Semantics*, 35:235–257, 2015.

[17] Lucy McKenna, Christophe Debruyne, and Declan O'Sullivan. Modelling the provenance of linked data interlinks for the library domain. In *Companion Proceedings of the 2019 World Wide Web Conference*, pages 954–958, 2019.

[18] P Missier, J Bryans, C Gamble, and V Curcin. Abstracting prov provenance graphs: A validity-preserving approach. *Future Generation Computer Systems*, 111:352–367, 2020.

[19] Maryam Davari and Elisa Bertino. Access control model extensions to support data privacy protection based on gdpr. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 4017–4024. IEEE, 2019.

[20] João Felipe N Pimentel, Paolo Missier, Leonardo Murta, and Vanessa Braganholo. Versioned-prov: A prov extension to support mutable data entities. In *International Provenance and Annotation Workshop*, pages 87–100. Springer, 2018.

[21] SEIS-PROV validator. https://github.com/SeismicData/SEIS-PROV/tree/master/validator/seis_prov_validate. Last Accessed: 2021-01-06.

[22] PROV python documentation. https://prov.readthedocs.io/en/latest/prov.html#. Last Accessed: 2021-02-22.

[23] Python interface to Graphviz's Dot. https://pypi.org/project/pydot/. Last Accessed: 2021-02-22.

[24] John Ellson, Emden Gansner, Lefteris Koutsofios, Stephen C North, and Gordon Woodhull. Graphvizopen source graph drawing tools. In *International Symposium on Graph Drawing*, pages 483–484. Springer, 2001.

[25] Emden Gansner, Eleftherios Koutsofios, and Stephen North. Drawing graphs with dot, 2006.

[26] Benjamin E Ujcich, Adam Bates, and William H Sanders. A provenance model for the european union general data protection regulation. In *International Provenance and Annotation Workshop*, pages 45–57. Springer, 2018.

[27] Claus Pahl, Nabil El Ioini, Sven Helmer, and Brian Lee. An architecture pattern for trusted orchestration in iot edge clouds. In *2018 Third Inter-*

*national Conference on Fog and Mobile Edge Computing (FMEC)*, pages 63–70. IEEE, 2018.

[28] Basel Halak. Cist: A threat modelling approach for hardware supply chain security. In *Hardware Supply Chain Security*, pages 3–65. Springer, 2021.

[29] Sara Boutamina, James DA Millington, and Simon Miles. Bottleneck patterns in provenance. In *International Provenance and Annotation Workshop*, pages 212–216. Springer, 2018.