UNECE

CONFERENCE OF EUROPEAN STATISTICIANS

Expert Meeting on Statistical Data Confidentiality

1-3 December 2021, Poland

Modelling data environments within PROV to assist anonymisation decision-making.

Mark Elliot (University of Manchester)
mark.elliot@manchester.ac.uk

Abstract

The Anonymisation Decision-making Framework (ADF) operationalises the management of the risk of data exchange between organisations and environments. Despite providing clearer theoretical underpinnings for implementing functional anonymisation, the complexity of the framework means that it still in general needs to be operated by an expert. The medium term goal is to automate as much of the anonymisation decision-making process as possible. In its second edition, the ADF has increased its emphasis on modelling data flows, highlighting the potential value for formal provenance information in anonymisation decision-making. We provide a use case that showcases this functionality. Based on this use case, we identify the requirements for provenance information such that it can be utilised within the ADF framework, and identify a currently unmet requirement: the modelling of data environments. We show how data environments can be implemented using the W3C PROV standard in four different ways. We analyse each approach for costs and benefits, as well as checking them against a second use case for completeness. We summarize our findings and suggest ways forward for representing data environments within W3C PROV to underpin the automation of the ADF.

Modelling Data Environments Within PROV to Assist Anonymisation Decision-making

Muhammad Aslam Jarwar*, Adriane Chapman**, Mark Elliot*, Fatemeh Raji**

* University of Manchester, Manchester M13 9PL, UK,
{aslam.jarwar,mark.elliot}@manchester.ac.uk

** 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. Based on this use case, we identify how provenance information could be utilized within the ADF, 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.

1 Introduction

The Anonymisation Decision-Making 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 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 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 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 formal 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.

W3C PROV is a standard for provenance interoperability that represents where data came from, and how it has been processed [11, 12]. 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 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 (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 drawn from [10] that is an idealisation of a common data situation; the sharing of data by an NSI 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, pupil data from schools, traffic data from smart sensors and 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 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.
- GOND also releases aggregated data into the public domain, i.e., an open environment by definition.
- 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; represented as data environments 1, 2, and 2_a. The research labs and the open environment are labelled with data environments 3_n to 3_{n+1} and 4.

For the purposes of understanding this data situation, the origin of the data flow is the GOND data environment (1).[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).

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 ($3_n, 3_{n+1}, ...$) who want to access the data for substantive analyses.

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).

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.

[1] See [10, p 37-39] for a more detailed discussion of the concepts of responsibility and control.

[2] Questions of granularity and scope affect all uses of provenance information. Sometimes, one may want to push the flow back to the data subjects. For simplicity's sake here we are assuming that GOND is the origin.

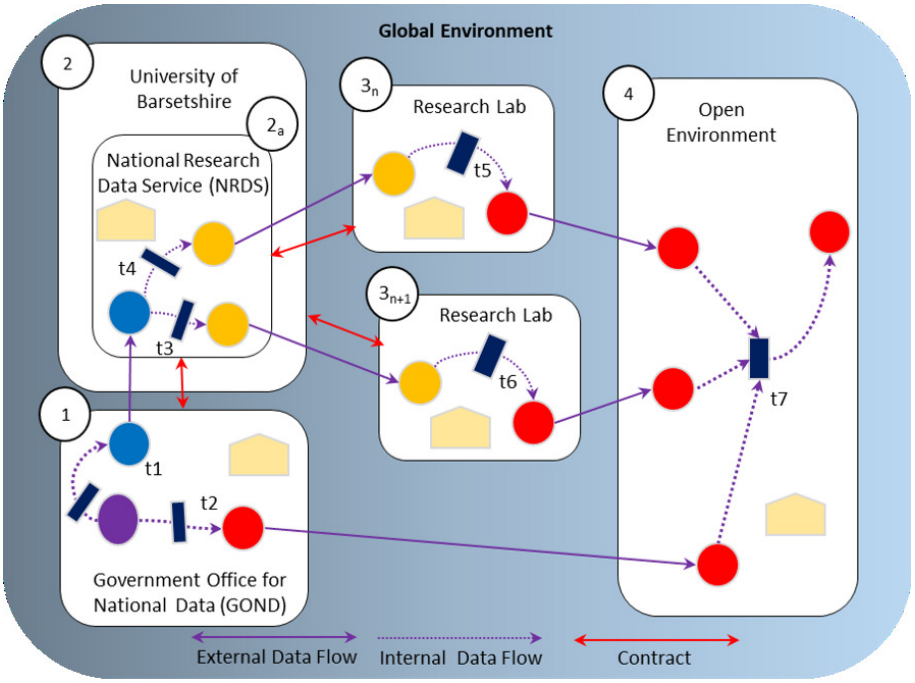

Figure 1: *A use case of data flows between and within multiple data environments. The red arrows indicate contractual agreements. The blue lines indicate data flow. Data environments are indicated by rounded rectangles, a circle represents a piece of data, a rectangle represent a process and a pentagon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

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: Nesting data environments within data environments
Often an environment will contain other environments. For example, data flows between an organisation's units for analysis, auditing and etc. In our example, the NRDS data environment is contained within the Barsetshire University data environment. In general, access control will be tighter in the sub-environment than the host environment.

R3: Attaching attributes to data environments

[3]PROV-DM is the conceptual data model and core part of W3C PROV that defines each term used to represent provenance information [13].

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. 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. Its parameters may be defined by - for example - a data sharing agreement, GONDS own data policies, enabling legislation etc.

R4: Relationships between data environments
This describes the possible relationship of one data environment with another. 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
In order to reason over data environment interactions, it is important to allow the attachment of semantic meaning to the relationships between the constructs. For example, NRDS receive data from GOND 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 information needed for the ADF.

R6: Representation of agents, data and processes in data environments
Agents might include data controllers, processors, users and subjects [4]. Data includes datasets, reports, analytical outputs 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. 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. In our use case, GOND has a data dissemination function that enables the sharing with by NRDS (based on some contract). GOND also has indirect control over data released 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

[4] following the terminology of the General Data Protection Regulation (GDPR), see for example: [14] for definitions

existing W3C PROV data model does already support some of these requirements in that they can be mapped onto existing PROV elements. However, there are some data environment specific requirements that need extensions in PROV. In the following sections, we will describe four possible mechanisms: namespaces with or without supporting structures and bundles with or without an extension.

3.1 Namespaces

The namespace concept was inspired by the World Wide Web architecture and was designed to make objects interoperable across technologies and platforms [13]. In PROV-DM, Namespaces are a Uniform Resource Identifier (URI) and 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 the information pertaining to related elements through Namespace concatenating and de-concatenating.

In Figure 2, there are five main data environments each with a 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" in the NRDS data environment can be recognised with a *http://global-env.com/bu/bu/nrds/agent_controller_001#*. Additionally, as illustrated in Figure 2, the data provenance for research labs can be tracked forward and backward through 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.

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, on their own they are not sufficient to satisfy all of the requirements identified in section 2. For instance, the attachment of additional attributes to the data environment itself and the contractual relationships between data environments cannot be accommodated.[5] Additionally, relationships among namespaces beyond containment cannot be captured. So whilst it is possible using namespaces to represent a data environment such as *http://www.nytimes.com* data environment that contains a sub-data environments eg *http://www.nytimes.com/ads*, within our use case, we need to be able to represent more complex relationships than this strict-hierarchical containment. For example, researchers from one of the Research Labs might have a specialised data 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.

[5]In related work we are analysing - and developing a grammar for - *data governance instruments* a concept which covers everything from participant information sheets through organisational policies, data sharing agreements to national legislation but this lies beyond the scope of this paper.

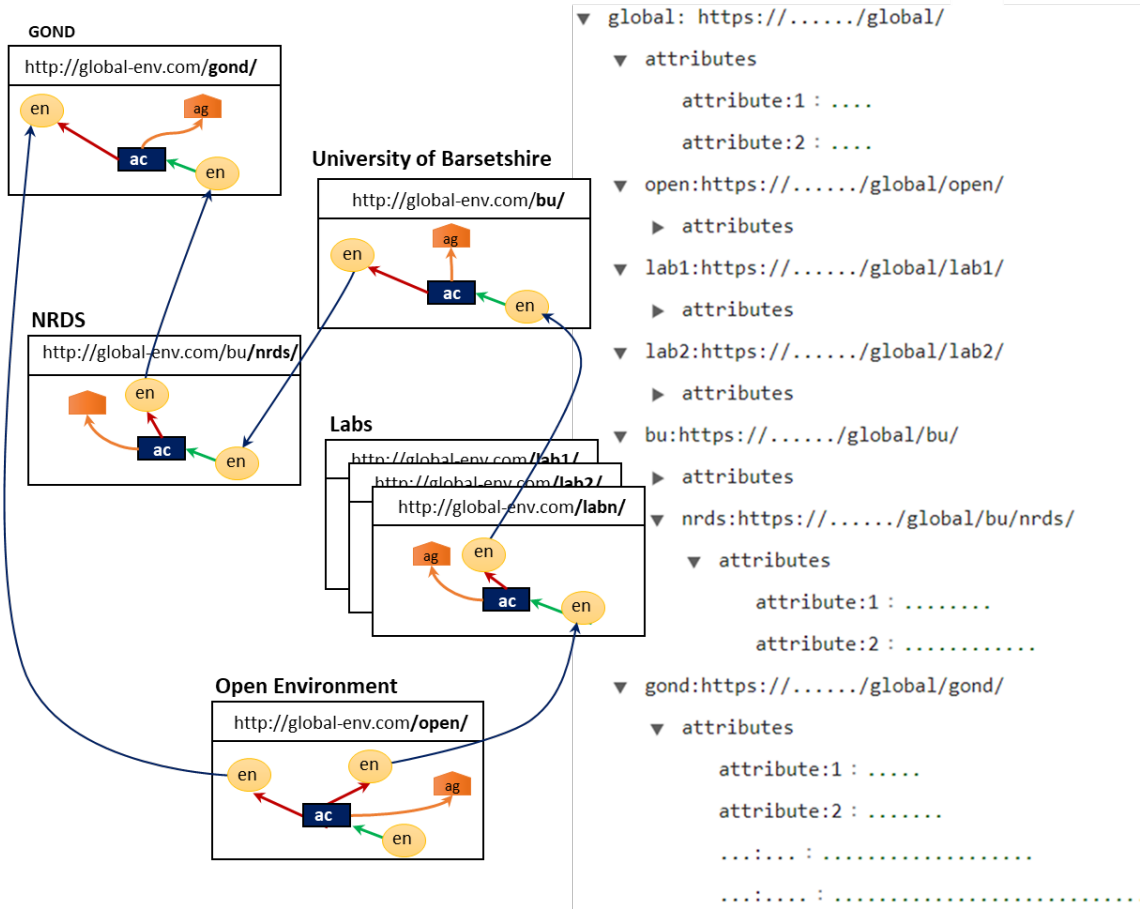

Figure 2: *Illustration of the use of namespaces to represent data environments: ag, ac, and en indicates agent, activity, and entity respectively; the right-hand part shows data environments with attribute attachment using namespaces. Relationships across namespaces could be captured in the same manner.*

To solve these issues, an additional set of structures are required. For instance, a separate document could be used to extend namespaces and allow the attachment of attributes.

3.2 Bundles and Extended Bundles

In PROV, the *bundle* 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 [15]. 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. A view of our use case as we might conceive of them in PROV bundles is shown in Figure 3.

In Figure 3, the large rectangles delineate data environments each represented as a PROV bundle. Each bundle contains data environment elements (represented as

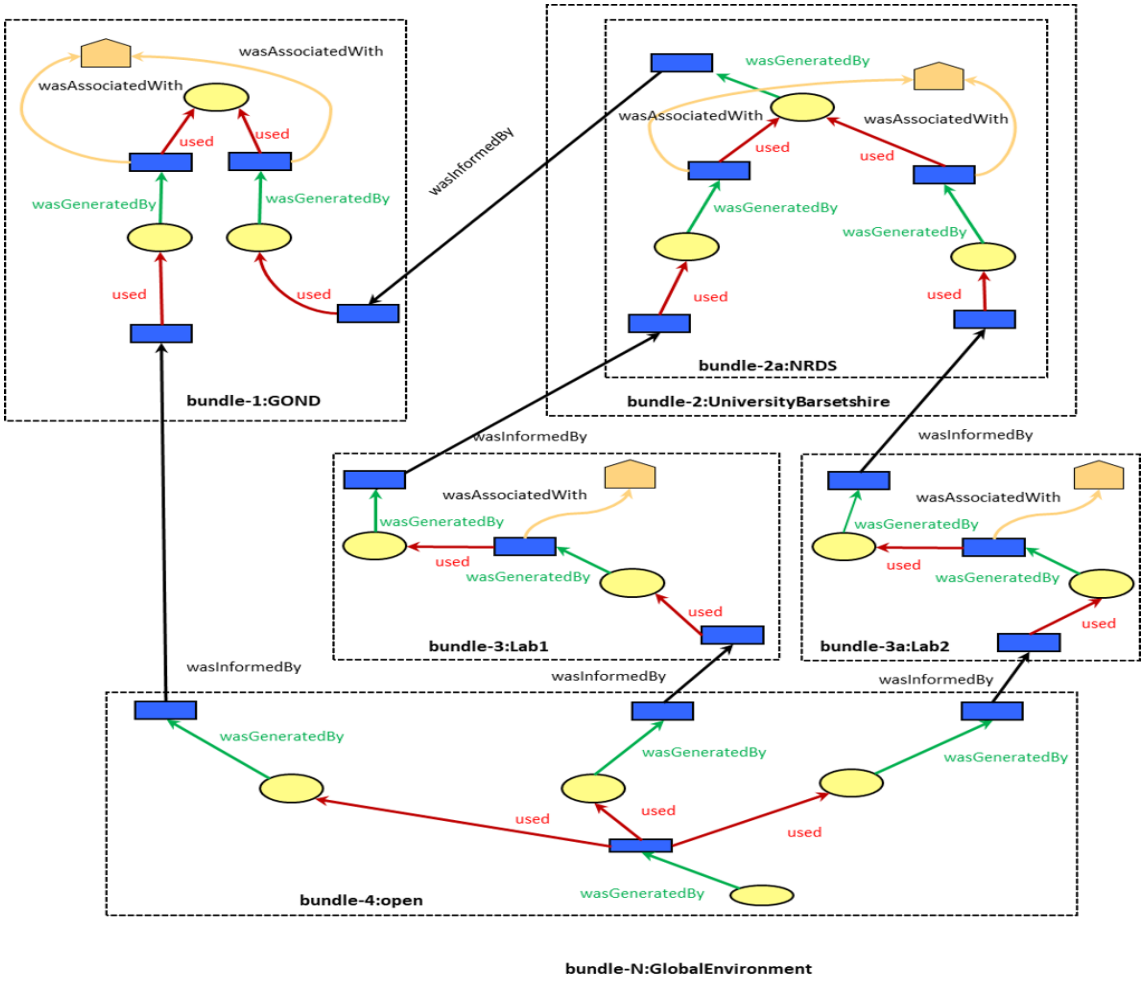

Figure 3: *A representation of the GOND-NRDS use case supported with PROV bundles. Please note that the nested data environments are shown with dotted lines are for illustration of use case and currently these are not supported in PROV.*

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 is shown with "actedOnBehalf" property. The data flow between one data environment and another (in the representation from bundle to bundle; for instance bundle-1:GOND -> bundle-2a:NRDS) is shown 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) and indeed all of the environments are nested within the global data environemnt. At present the bundles construct in PROV does not support this nesting.[6] However,

[6]We also note that in ADF terms, NRDS is said to have *direct control* over the labs environment

it is worth noting that W3C PROV constructs were designed to be extensible [13]. In previous work, PROV has been extended to express the provenance of big data security supervision [16], provenance access control [17], data privacy protection based on GDPR using provenance [18] and managing mutable entities by adding reference derivations and checkpoints [19]. Similarly, the existing structure of PROV bundles could be extended in order to support and express the requirements of ADF with more flexibility. For example, additional metadata to the bundle construct could be attached to define different types of data environments and another extension that we would need in PROV Bundles, is support for nested data environments.

4 Comparative Analysis

Table 1 shows how well each implementation option discussed in Section 3 meets the representation requirements for data environments outlined in section 2.[7]

Table 1: Use case requirements analysis for representing data environments

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

The nesting of data environments (i.e. data environments within data environments) is one of the important features. Using bundles we cannot represent this nesting because PROV does not allow the nesting of bundles [13]. This gap is one of the drivers for bundles+. This requirement is supported by namespaces (and so namespaces+ can also support nesting).

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.

[7]To construct this, we also analysed the completeness of the requirements mapping through an additional example scenario. This can be found in an earlier but longer version of this paper [20]. No additional requirements were found to exist.

The ability to attach attributes to a data environment is also an important requirement. Neither bundles nor namespaces support this feature. The additional structures provided in Namespace+ do allow attributes to be maintained using the namespace information.

As we observed in the GOND-NRDS use case in Figure 1, the GOND data environment contains the representation of collected, processed and shared data along with the associated 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.

As it can be seen from Table 1, the contracts are supported by both Bundles+ and Namespaces+. The representation of access control requirement is supported by all four constructs.

Both the bundles and namespaces solutions could naturally support the representation of agents, processes and entities using native W3C PROV concepts. Additional granularity can be added to the representation through developing functions that enable annotations to the relationships of agents, processes, etc

5 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 based on the Anonymisation Decision-making Framework (ADF) which is a now well established approach for supporting reasoning 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 formal provenance might be utilised.

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 required 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 and suitability for the problem. While two obviously do not pass muster, the other two are viable solutions. The first is *Namespaces+* that utilises existing W3C PROV *Namespaces* structures but requires an additional management, and the second is *Bundles+* which requires an extension to the existing W3C PROV *Bundles*.

Acknowledgement

This work was supported by the Alan Turing Institute (grant No. R-SOU-008).

References

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

[12] 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.

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

[14] EDPS Guidelines on the concepts of controller, processor and joint controllership under Regulation (EU) 2018/1725. `https://edps.europa.eu/sites/edp/files/publication/19-11-07_edps_guidelines_on_controller_processor_and_jc_reg_2018_1725_en.pdf`.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] Muhammad Aslam Jarwar, Adriane Chapman, Mark Elliot, and Fatemeh Raji. Provenance, anonymisation and data environments: a unifying construction. *arXiv preprint arXiv:2107.09966*, 2021.