Anonymisation and Provenance Expressing Data Environments with PROV - Turing Pilot Project Final Report

ADRIANE CHAPMAN, Electronics and Computer Science University of Southampton, UK
MARK ELLIOT, The Cathie Marsh Institute (CMI), University of Manchester, UK
MUHAMMAD ASLAM JARWAR, The Cathie Marsh Institute (CMI), University of Manchester, UK
FATEMEH RAJI, Electronics and Computer Science University of Southampton, UK
TOM BLOUNT, Electronics and Computer Science University of Southampton, UK
DUNCAN SMITH, The Cathie Marsh Institute (CMI), University of Manchester, UK
KIERON O'HARA, Electronics and Computer Science University of Southampton, UK

ACM Reference Format:
Adriane Chapman, Mark Elliot, Muhammad Aslam Jarwar, Fatemeh Raji, Tom Blount, Duncan Smith, and Kieron O'Hara. 2022. Anonymisation and Provenance Expressing Data Environments with PROV - Turing Pilot Project Final Report. 1, 1 (October 2022), 34 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 PROJECT OVERVIEW

The knowledge economy needs data; datasets; data processes (streaming, sharing, linking). The dissemination and use of data carries privacy and confidentiality risks through the inadvertent disclosure of personal data. Without better tools to understand risk and choose anonymisation techniques, data science and artificial intelligence activities will lack the needed data. This project explored the overlap and interaction between the provenance interoperability standard, W3C PROV, and the information required to make data sharing and anonymisation decisions.

There are many forms of anonymisation, but none will remove the threat of an attacker recreating personal information. Choosing which technique to use requires understanding of the attacker threat, what the shared data is to be used for, and the context in which it was both gathered and released. Recent work [1] has introduced the concept of **Functional Anonymisation** which states that risk lies not in the properties of the data on their own, but in the relationship between data and their context, called the **data environment**, which can be characterised by four parameters: the **agents** with access to the data; the supplementary data which can be integrated with the data; the **infrastructure** in which the data is stored and processed; and the **governance** of the data. Anonymity is not therefore solely a property of the data, but a function of the data environment(s)

Authors' addresses: Adriane Chapman, adriane.chapman@soton.ac.uk, Electronics and Computer Science University of Southampton , Southampton, UK, SO17 1BJ; Mark Elliot, mark.elliot@manchester.ac.uk, The Cathie Marsh Institute (CMI), University of Manchester, Manchester, UK, M13 9PL; Muhammad Aslam Jarwar, aslam.jarwar@manchester.ac.uk, The Cathie Marsh Institute (CMI), University of Manchester, Manchester, UK, M13 9PL; Fatemeh Raji, f.raji@soton.ac.uk, Electronics and Computer Science University of Southampton , Southampton, UK, SO17 1BJ; Tom Blount, t.blount@soton.ac.uk, Electronics and Computer Science University of Southampton , Southampton, UK, SO17 1BJ; Duncan Smith, duncan.g.smith@manchester.ac.uk, The Cathie Marsh Institute (CMI), University of Manchester, Manchester, UK, M13 9PL; Kieron O'Hara, kmo@soton.ac.uk, Electronics and Computer Science University of Southampton, Southampton, UK, SO17 1SX.

XXXX-XXXX/2022/10-ART $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

in which it is held. Anonymisation can be reversed when someone with relevant auxiliary data can gain access to the confidential data and perform the necessary data integration and by doing so reidentify some or all people in the dataset.

However, understanding the complexity of data environments is challenging, and requires an awareness of data flows. In the absence of procedural rigour, some high-profile attempts to anonymise data had led to damaging data breaches [2]. The development of the Anonymisation Decision-Making Framework (ADF) [3] which captures risk by mapping data flows has been a step forward and has imposed procedural formality onto a process which was previously somewhat subjective. What is needed to move this on further is a representational formalism which can capture the necessary features of the environment to allow more principled analysis of risk. We proposed in this pilot project to investigate the use of provenance formalisms for that purpose.

Provenance is the record of creation and modification of data and processes. It has many uses, including: debugging, scientific reproducibility, and establishing trust in data. The W3C PROV is an interoperability standard for provenance that defines actors, entities, activities, and the relationships between them. Using provenance (as described by PROV), it is possible to trace where data came from, and how it was processed.

Thus, what has emerged from our investigation is a richer understanding of the multiple fields that are combined in a complex problem. This project tackled several aspects of the initial problem:

- Fundamental evaluation of the requirements of the information required by the ADF for decision making, and mapping this to the W3C PROV.
- Expansion of the W3C PROV model in order to meet the ADF requirements.
- Analysis of ADF-PROV across real, industrial use cases.

Key findings from the project which are set out in detail in the sections that follow are.

- Formulation of RP^4 and extended model of provenance: Retrospective, Prospective, Permitted, Prescriptive and Proscriptive provenance.
- Creation of a prototype that provides analysis for sharing concerns to decision makers using RP^4 provenance-encoded data and ADF rules.

Table 1 outlines the coverage in previous work of the core theoretical and methodological concepts we have been working with.

2 FOUNDATIONS

2.1 Anonymisation for Personal Data Protection

In order to protect personal data various approaches have been taken. To improve data utility while focusing on protecting data from de-anonymisation attacks, the anonymisation framework was developed [1]. This framework argues that the risk of personal data de-anonymisation lies in the relationship of data and the environments in which the data currently reside or to where it will be moved. The authors demonstrate that both the quality of data could be improved and the risk of data re-identification could be reduced when the decision about anonymisation takes into account the context and data flows.

Ulltveit-Moe et al. [9] developed a contextual anonymisation scheme. Their scheme supports real time anonymisation over the messages that are embedded in XML payloads. The proposed scheme was contextual in the sense that the anonymisation service anonymises the data to various levels by considering the receiving party. Additionally the anonymisation scheme sends the de-anonymisation key along the anonymised data if the contracts allows usage of de-anonymised data for a period of time. In this approach, the authors consider system logs as provenance to distinguish different types of data stakeholders.

Table 1. Comparison of related work (- symbol indicate "no")

Article	Contextual annonymisation	Data flows modeling	PROV extention	Real time provenance capturing	Provenance enabled disclosure control	Privacy focus
[1]	✓	✓	-	-	-	✓
[4]	-	✓	-	-	✓	✓
[5]	-	✓	✓	✓	-	-
[6]	✓	-	✓	-	✓	-
[7]	-	✓	✓	✓	-	-
[8]	-	✓	✓	-	-	✓
[9]	✓	-	-	✓	-	✓
[10]	✓	-	-	-	✓	✓
[11]	✓	-	-	-	-	✓
This work	✓	✓	✓	✓	✓	✓

Rumbold et al. [10] argue that current anonymisation standards use the same set of processes across all the situations and this increases the risk of successful de-anonymisation attacks. The main focus of their work was to protect the privacy of sensitive medical data that requires consent in order to use that data for secondary purposes. In their research, they have developed a set of anonymisation methods that work on the formulation of matrices which can be adjusted according to the sensitivity of the data, people, place and time involved in the environment. Their proposed formulation matrix for contextual anonymisation was based on four adjustable parameters: *"(i) research in safe heaven, (ii) research to which duty of confidentiality applies, (iii) Research to which no duty of confidentiality applied, (iv) information for public release"*. In their proposed framework, the authors did not explicitly acknowledge the usage of provenance metadata, but we note that the concepts within their matrix have obvious relationships to provenance information.

Hasanzadeh et al. [11] created a context sensitive anonymisation approach to protect the privacy in spatial datasets which is developed from and used by the Public Participation Geographic Information System (PPGIS). Their scheme anonymised the individuals home locations when the spatial datasets were used for mapping purposes. They utilise several techniques including: bimodal Gaussian displacement algorithm to deviate the locations spatial attributes, and K-anonymity for non-spatial attributes. Their approach focused on privacy preservation without considering a de-anonymisation attack.

2.2 The Anonymisation Decision Framework (ADF)

Elliot et al. developed the Anonymisation Decision-Making Framework (ADF) to provide practical and operational guidance about anonymisation in order to prevent unintended disclosure of personal information [3, 12].

The core underpinning concept of the ADF *functional anonymisation*[1, 12] revolves around four core principles. For our current purposes the most important of these is the **Comprehensiveness Principle** which states: You cannot decide whether or not data are safe to share/release by looking at the data alone. This principle encapsulates the *data situation approach* where risk is seen as arising from the interaction between data and their environment and where the environment includes other data, agents and (the soft and hard) structures that shape the interaction (such as

national policies on data sharing and access, the legal framework, IT systems, governance practices, cultural attitudes to data sharing and privacy, etc.).

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 defining features

- **Other Data:** Auxillary information that is or could be co-present in the environment and is in principle linkable to the data in question.
- **Governance:** Policies, Procedures and Processes that control data access and processes.
- **Infrastructure:** Technical-ware which increases security by constraining access.
- **Agents:** Humans and other intelligent systems which are (potentially) present in the environment.

When data moves between environments (a *dynamic data situation*), each environment produces a different risk profile, depending upon how the data interacts with the four defining features.

Using a risk profile which is based on the interaction between the data in question and these four features allows a data controller to make an informed decision in whether and how the data should be protected and shared.

2.3 W3C PROV

The W3C PROV is a standard for provenance interoperability that represents where data came from, and how it has been processed [13, 14]. 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.

To protect provenance of mutable values, [7] developed a PROV extension for time-versioning entities by adding reference sharing and checkpoints. 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 the folding and unfolding required to add the checkpoints.

A W3C PROV based provenance model has been proposed by Benjamin et al. [8] that uses the PROV data model and data protection ontologies to express the provenance for the purposes of 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 using PROV properties. Both of the ADF examples presented in this work fall under GDPR regulations, and the extensions introduced in Benjamin et al. [8] would facilitate some of the more general requirements of **representation of agents, data and processes** and **contracts** within data environments.

Missier et al. [6] worked on provenance abstraction operator for managing the access to and control of provenance graphs. The motivation of their approach was that the disclosure control could be applied over the sensitive provenance content. The operator they developed rewrites

provenance graph G_1 to a new abstract version G_2. Their approach also focused on the dependencies between the two versions of the provenance graph through various constraints and relationship properties. Other work that looks to protect the disclosure of provenance information includes [15–17].

The PROV data model has also been extended with new relationship properties in order to supervise the security of data streaming [5]. These extensions focus on collecting the provenance information about data operations inside and outside of big data clusters and representing the data interaction flow between the clusters. The harvested provenance information is analysed for the detection of anomalies in the data. (checking for inconsistency between the graphs in nodes and edges).

Recently Jung et al. [4] utlised provenance to analyse the risk involved in sharing data. The authors argued that the risk of a datasets' disclosure might be increased if someone released another related dataset in the public environment because the attacker could link both datasets in order to discover personal information. According to their methodology when tables T_1 and T_2 are co-related, then the data owner or controller should check the provenance of T_1 from T_o (master dataset) in order to identify all the dependencies of T_2. Then the tables with dependencies were clustered and modelled into several groups. So, if T_1 and T_2 are co-related, and T_2 and T_3 are co-related, then these could be clustered as T_1, T_2 and T_3 in one group.

3 FUNDAMENTAL EVALUATION OF THE REQUIREMENTS OF THE INFORMATION REQUIRED BY THE ADF FOR DECISION MAKING

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.*

Based on the core principles of functional anonymisation, and the definition of the data environment described above, the necessary requirements to model the ADF concepts for automated processing include:

E: The Entity construct The data, documents, or other artefacts that are affected by some process. These can be digital, physical, or conceptual, and can be described by a set of attributes or metadata.

E1: The Data Governance Instrument construct defines the Permissions (what **may** be shared), Prescriptions (what — or how it — **must** be shared), and Proscriptions (what **must not** be shared) that bind Actors in a contract describing how they are able to share Entities between one-another or data environments.

P: The Process construct The set of activities that act on entities; this might include creating, transforming, deleting those entities.

A: The Actor construct The owner or controller of a Process or Entity, such as a person or organisation. Actors might include data controllers, processors, users, and subjects.[1]

N: The Environment construct A container and associated boundaries that contains entities, processes and agents associated with a particular set of governance.

N1: The Nested Environment construct An Environment partially or fully contained within another Environment. This is often the case with institutional or other hierarchical structures (such as the NRDS within the UoB in the example above). The data within the child Environment is subject to the union of constraints imposed by both child and parent.

N2: The Annotated Environment construct A record of the properties and relationships of the access and control mechanisms, relationships, and risk and disclosure levels of each environment. This construct is associated with an Environment or Nested Environment construct. In order

[1] following the terminology of the General Data Protection Regulation (GDPR), see for example: [18] for definitions

Table 2. Requirements that can be completely (✓) or partially (—) fulfilled by the standard PROV model

Requirement		Fulfilled?
Entities	(E1)	✓
Processes	(P1)	✓
Actors	(A1)	✓
Environments	(N1)	✓
Nested Environments	(N2)	—
Environment Annotations	(N3)	—
Relationships	(R1)	—
Relationship Annotations	(R2)	—
Data Governance Instrument	(D1)	—

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 needs to be recorded.

R: The Relationship construct defines the relationships between one data environment with another more complex containment (N2).

R1: Relationship Annotations describe, in order to reason over data environment interactions, the semantic meaning of the relationships between the constructs.

4 MAPPING ADF REQUIREMENTS TO PROV MODEL

Table 2 summarises the requirements from Section 4 that can be met at present using standard PROV.

4.1 Ability to model the ADF Tool

The goal of this work is to allow a more standardised capture and representation of information that is necessary for the ADF to operate. As such, we utilise the questionnaire based tool provided by the ADF and determine which information can be fully modelled within provenance.

The ADF tool poses around thirty questions to data controllers in order to determine the data situation sensitivity and summary risk, and from this how the risk can be managed. In Tables 3- 7 we itemise these questions, and describe whether they can be sufficiently modelled in a machine-readable manner with our approach according to the following definitions:

- **Natively (✓):** If the concepts required by a given ADF question can be represented in W3C PROV with no adaptations or requirements on how the PROV model is deployed.
- **Model (•):** If the concepts required by a given ADF question can be represented in W3C PROV with no extensions, as long as the PROV model is deployed under particular constraints. These constraints include requirements of the types of relationships to capture, or entities that must be represented.
- **Extensions (○):** If the concepts required by the ADF question can be represented in W3C PROV that has been extended (see section 5).
- **Not Applicable (-):** Indicates an ADF question that does not focus on the data handling or environments, and thus is not germane to PROV.

Table 3. Assessment of ADF questions on the theme of Agreement Sensitivity

ADF Question	W3C PROV
Are the data subjects aware that their data have been collected in the first place?	●
Have the data subjects agreed (explicitly or implicitly) to the collection of their data?	●
Were the data subjects completely free to agree to the collection of their data (or have they agreed to collection because they want something (a good or service) for which are required to hand over some data before they can obtain it)?	-
Are the data subjects aware of the original use of their data?	-
Have the data subjects agreed (explicitly or implicitly) to the original use of their data?	●
Have the data subjects agreed in general to the sharing of a functionally anonymised version of their data?	●
Are the data subjects aware of the specific organisations that you are sharing a functionally anonymised version of their data with?	○
Have they agreed to your sharing their data with those organisations?	●
Are the data subjects aware of the particular use to which their functionally anonymised data are being put?	●
Have they agreed to those uses?	●

Table 4. Assessment of ADF questions on the theme of Expectation Sensitivity

ADF Question	W3C PROV
Do you (the sending organisation) have a relationship with the data subjects such that a reasonable data subject would expect you to have access to their data?	-
Does the receiving organisation have a relationship with the data subjects such that a reasonable data subject would expect them to have access to their data?	-
Is the receiving organisation a government or commercial entity?	●
Is your organisation's area of work one where trust is operationally important (e.g. health or education)?	●
Will you receive financial or commercial benefit from the data share?	-
Is there an actual or likely perceived imbalance of benefit arising from the proposed share or release? E.g. is the data controller benefiting but the data subjects not?	-

Table 5. Assessment of ADF questions on the theme of Data Sensitivity

ADF Question	W3C PROV
Are some of the variables sensitive?	•
Are the data about a vulnerable population?	•
Are the data about a sensitive topic?	•
Is the use of the data likely to be considered sensitive?	•
Do you have reason to believe that the intended use of the data might lead to discrimination against the data subjects or a group of which they are members?	•

Table 6. Assessment of ADF questions on the theme of Desensitising Factors

ADF Question	W3C PROV
Will there be some public benefit arising from the downstream use of the data?	-
Have you carried consultations with groups of stakeholders (particularly the general public and/or data subjects)	-
Have you carried consultations with groups of stakeholders (particularly the general public and/or data subjects) and implemented the recommendations arising there from?	-
Does your communication plan engender trustworthiness through transparency (sufficient to offset adverse responses in the expectation sensitivity section)?	-

4.2 The Fit of PROV

Looking at Tables 3- 7, only 1 field is covered with a (✓), indicating that the information required by the ADF is natively supported in PROV. While this may seem discouraging at first blush, the fundamental building-blocks required by the ADF which are used to reason over the questions in Tables 3- 7, but not specifically stated as a core question are fundamental to provenance. These include: Actors/Agents, Activities/Processes, Entities/Data, Relationships among them. Moreover, all questions marked with a (•) are completely supported natively within PROV, but require a specific modelling choice when capturing and recording provenance. We expand on this topic in Section 5. Only the items marked as (○) are currently impossible to model within W3C PROV; we discuss the required(and W3C PROV allowable) extensions in Section 5.

5 EXTENDING PROV (○)

In order to meet the unfulfilled requirements for representing the information required for reasoning using the ADF (i.e. N2, N3, R1, D1) with the PROV standard, we must extend the model. Below, we consider how two constructs built into the PROV model, *bundles* and *namespaces*, might be re-used as a starting point. We briefly examine the potential value of of these solutions either as they are or

Table 7. Assessment of ADF questions on the theme of Summary Risk

ADF Question	W3C PROV
Are the data of high quality?	●
How old are the data?	✓
Do the data constitute a whole population or a sample?	●
How many variables are there that fall within the standard key variable sets?	●
Which of the following best describes the data?	●
Do the data include any data types that present particular reidentifiability challenges (e.g. genomics data, photographs, significant text narratives, timestamped location data or other timestamped sequences?	●
Now considering the details of the focal environment, which of the following best describes that environment?	○
Are there data in — or which could be moved into — the focal environment that could be used to re-identify any data subjects in the data?	○

with extension and 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).

5.0.1 Namespaces. The data environment concept is to support reasoning about the data, processes, governance at play within an organisation or sub-organisation's boundaries. In the semantic web, notions of organisation containment are often expressed via namespaces. The Namespace concept was inspired by the World Wide Web architecture and was designed to make objects interoperable across technologies and platforms [19]. In PROV-DM, Namespaces are a Uniform Resource Identifier (URI) and a provenance graph can contain multiple Namespaces. The Namespace is a candidate for use as an identifier to capture the idea of data environments, including data environments within data environments, and their associated entities, activities, agents, etc.

Assigning a Namespace to a data environment, and using prefixes[2] to concatenate nested data environments, it is possible to specify data environments within PROV. For example, we note that in the example use case, the NRDS data environment is a part of the University of Barsetshire environment and so, using Namespaces, we might refer to the University of Barsetshire and NRDS data environments as *http://global-env.com/bu/* and *http://global-env.com/bu/nrds/* respectively. We can also express the control mechanism over the data environments and its elements with namespaces. The visual representation of the GOND-NRDS use case with the support of Namespaces and PROV constructs is shown in Figure 1.

In Figure 1, 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#*. The right hand side of Figure 1 shows the pseudo code of attributes attachment with the data environment through namespaces' support.

[2]Prefixes are an alias to a namespace.

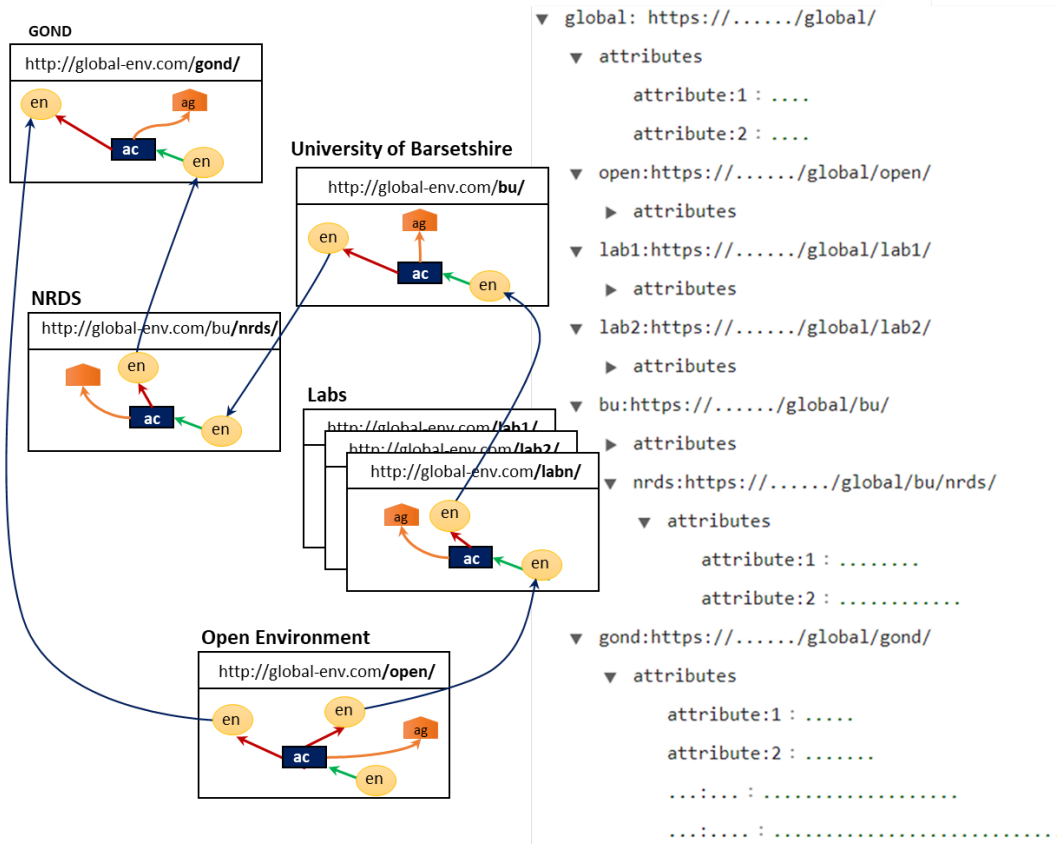

Fig. 1. Illustration of the use of namespaces to represent data environments: ag, ac, and en indicates agent, activity, and entity respectively; the right-hand part shows data environments with attribute attachment using namespaces. Relationships across namespaces could be captured in the same manner.

5.0.2 Namespaces with additional support structures. While namespaces have potential for representing the bounded nature of data environments, and what has occurred within a given data environment and its sub-environments, they are not sufficient to satisfy all of the requirements identified in Section 3. For instance, the attachment of additional attributes to the data environment itself and the contractual relationships between data environments cannot be accommodated. Additionally, relationships among namespaces beyond containment cannot be captured. 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 complex relationship. To solve these issues, an additional set of structures are required that tracks the relationships between namespaces and attached attributes.

5.0.3 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 [20]. For example, a bundle can contain the entities, activities, agents, and the relationships between them. Bundles can also support entities with attributes.

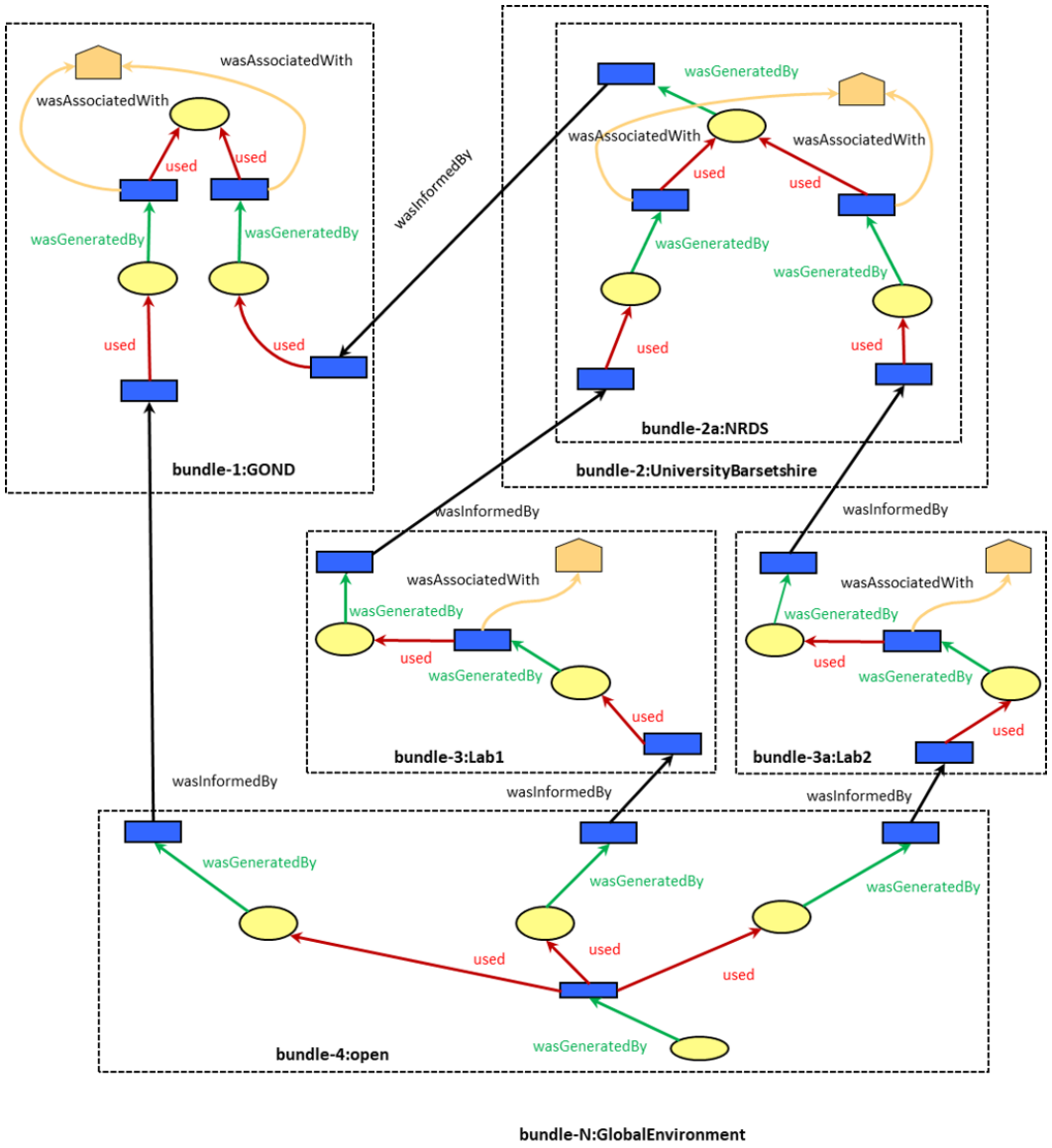

Fig. 2. A representation of the GOND-NRDS use case supported with PROV bundles. Please note that the nested data environments are shown with dotted lines are for illustration of use case and currently these are not supported in PROV.

This can help us to attach necessary metadata to the entities. A view of our use case as we might conceive of them in PROV bundles is shown in Figure 2.

In Figure 2, 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 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 denoted with the "wasInfromedBy" property.

We can also see in Figure 2 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 environment. At present, the bundles construct in PROV does not support this nesting and so an extension is required.

5.0.4 Extended Bundles. W3C PROV constructs were designed to be extensible [19]. In previous work, PROV has been extended to express the provenance of big data security supervision [5], provenance access control [6], data privacy protection based on GDPR using provenance [21] and managing mutable entities by adding reference derivations and checkpoints [7]. Looking at our use case, the existing structure of PROV bundles could be extended to support nested data environments. For instance, PROV could be extended to making the data environment a first class object, *dataEnvironment*, by creating a new layer over the bundle structure. As a second step, we could then build a mechanism to include attributes, entities, activities, agents, etc. in each *dataEnvironment*. In this approach, we can reuse some of the existing features of bundles and entities.

5.1 Comparison of Extension Approaches

Table 8 shows how well each implementation option discussed in Section 2.3 meets the requirements for modelling ADF concepts outlined in Section 3.

Table 8. Concepts required by ADF and ability to meet them in different PROV extensions

Representation requirements	W3C PROV implementation			
	Bundle	Namespace	Namespace+	Bundles+
Data Environment Construct	✓	✓	✓	✓
Nested 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	✓	✓	✓	✓

Both the bundles and namespaces solutions, and their extensions, 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.

The nesting of data environments is one of the important features for reasoning within the ADF. The use of namespaces and namespaces+ can support nesting with no additional extension of

PROV. Using bundles we cannot represent this nesting because PROV does not allow the nesting of bundles [19]. This gap is one of the drivers for bundles+.

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 as would an extension of Bundles, (Bundles+). Attaching annotation is important to help identify data governance instruments that specify particular disclosure control processes. Bundles+ supports this requirement.

6 ANALYSIS OF ADF-PROV ACROSS REAL, INDUSTRIAL USE CASES

6.1 Government Office for National Data (GOND)

A seemingly simple data flow between environments can in fact be complex depending on the nature of the data, the data environment(s), the intended data use and the responsibilities of the data stakeholders. Below we describe an example use case drawn from [12] that is an idealisation of a common data situation; the sharing of data by a National Statistical Institute with a research data service.

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, part of University of Barsetshire, has a role 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, (by definition an open environment). Additionally, researchers carry out data analysis on GOND's data and publish papers that report on this analysis in the public domain. One of the goals of the anonymisation decision-making is to ensure that when data that has been derived from the same original data, are released (either by different organisations or 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 example data situation epitomises.

Figure 3 shows the four data environments described above. These in turn are all part of the *global data environment*; the sum of all data environments. In effect the global data environment is layered and partitioned into an inestimable number of sub-environments to create a complex ecosystem through which data (and agents) move. Boundaries between sub environments are defined by the infrastructure and governance properties. The complexity of and uncertainty about the global data environment is one of the main reasons why the anonymisation problem is so operationally difficult and why utilising data provenance to enhance anonymisation decision making is an emerging and vital step forward.

For the purposes of understanding our example data situation, the origin of the data flow is deemed to be the GOND data environment (1).[3] 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.

[3]Questions of granularity and scope affect all uses of provenance information. Sometimes, one may want to push the origin back to the data subjects. However, here, for simplicity of exposition, we are assuming that GOND is the origin.

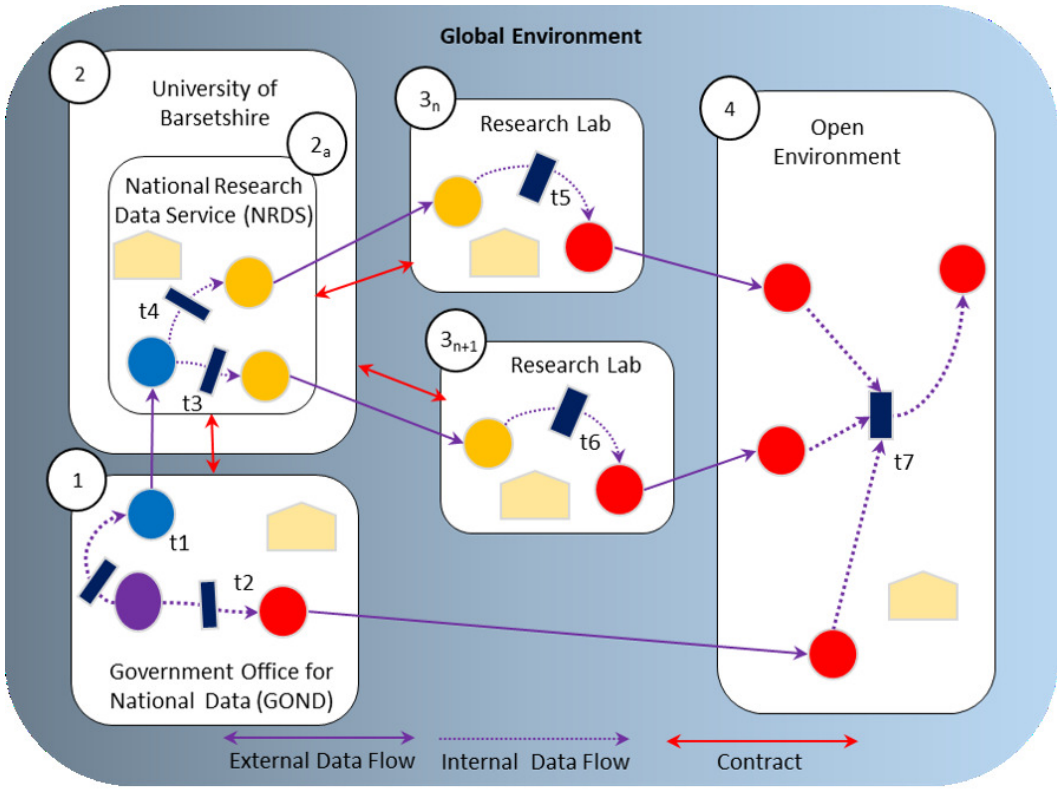

Fig. 3. A use case of data flows between and within multiple data environments. The red arrows indicate contractual agreements. The blue lines indicate data flow. Data environments are indicated by rounded rectangles, a circle represents a piece of data, a rectangle represent a process and a pentagon represents a user (in the data environment). The time for processing events is labelled from t_1 to t_7.

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).

6.2 Pharma-based Use Case

In our second scenario, data from clinical trials are generated at multiple 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. Explicit consent has been given by trial participants for secondary research using of anonymised versions of their data. 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.

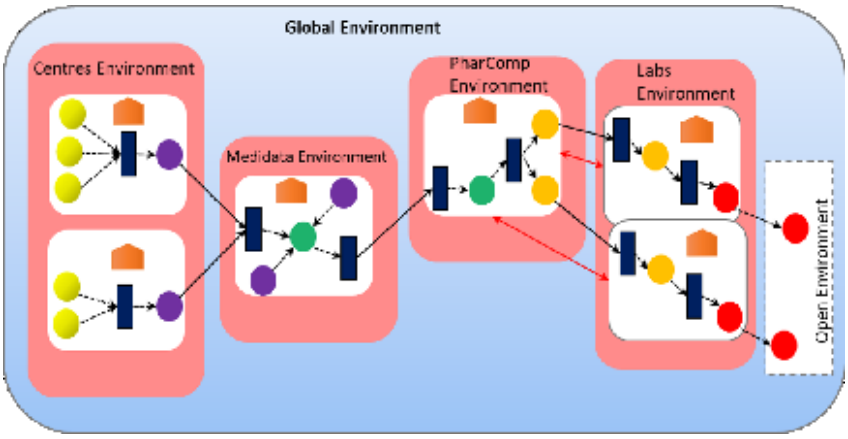

Fig. 4. A Second ADF use case with data environments: The red arrows indicate contractual agreements. The black lines indicate data flow. Data environments are indicated by rounded rectangles, A circle represents a piece of data, a rectangle represent a process and a pentagon represents a user in the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

7 FORMULATION OF RP^4 PROVENANCE: RETROSPECTIVE, PROSPECTIVE, PERMITTED, PRESCRIPTIVE AND PROSCRIPTIVE PROVENANCE

Typically, previous work has divided provenance into retrospective provenance (which relates to previous processing of data) and prospective provenance (which relates to planned future processing). In this paper we identify extensions to the semantics of provenance in order to enable the use of formal provenance for the ADF reasoning process. We expand the notions of provenance beyond workflow settings adding the concepts of *permitted, prescriptive and proscriptive provenance.*

7.1 Theoretical Foundations

The concepts of retrospective and prospective provenance have been mostly employed in scientific workflows at the system level. Provenance of scientific workflows captures executed tasks and their parameter settings, inputs and outputs. In other words, provenance records the runtime executions to achieve better understanding of the history and lineage of results [22, 23].

For instance, noWorkflow [24] automatically captures retrospective provenance of Python script executions. It reflects the processing history of a script without requiring any modifications to the script itself. However, the amount of information generated by noWorkflow may become overwhelming for the users. YesWorkflow [25] provides an annotation language in script-based workflows in such a way that the users are allowed to embed annotations within their scripts. Afterwards, the steps that compose the script and their data flow dependencies can be visualised and queried to expose the prospective provenance. More recent work has integrated noWorkflow and YesWorkflow [26, 27]. This line of research generally combines the provenance traces captured by noWorkflow (retrospective provenance) with the more abstract workflow view defined by YesWorkflow (prospective provenance).

7.1.1 The Application of Provenance in Data Protection. The ADF is concerned with identifying how data should be protected as they move from one environment to another. There is some existing work on the intersection of provenance and data protection. A large portion of this concerns protecting data within the provenance information. An overview of different strategies is provided

by [28], which includes path modification [29] and customisation for provenance disclosure [30]. The results of [28] have been extended into a framework of obfuscation and disclosure strategies [31]. Another secure provenance management technique is illustrated in [32] which develops an access control language model to support specification of policies for provenance.

There are relationships here with the statistical disclosure control tools that the ADF utilises in one of its components [33, 34]. Undoubtedly, the ADF could be applied to provenance data just as it can first order data forms. However, we are not primarily concerned with protecting provenance information, but rather with using provenance to assist in protecting other data and systems. For instance, CamFlow [35] captures system-level provenance in the Linux operating system for the purpose of intrusion detection. CamFlow tracks the interactions of the applications with the operating system, and with each other. Additionally, [36] contributes provenance for query results from database to evaluate the trustworthiness of data. Another example is ProFact [37], which employs provenance to support the analysis of the quality of access control policies.

There are then some existing techniques at the interface between provenance and data protection and related issues [29–32] [35–38]. For example, [38] has looked at creating provenance statements from terms and conditions and other types of privacy policy documents. However, as far as we are aware there is no existing work, which examines the application of provenance to anonymisation decision-making – the concern of this paper.

7.1.2 Novel Expansion. As mentioned above, in this paper we expand the notions of provenance beyond workflow settings adding the concepts of *permitted, prescriptive and proscriptive provenance*[4]. So, a system of five types of provenance, called **RP4**, is proposed:

(1) *Retrospective provenance:* To map the existing processing of the data. Retrospective provenance is useful for determining the elements of the existing data situation that present risk and sensitivity by providing the lineage of the data in question. In other words, retrospective provenance captures information about data processing that has happened (or is already happening).
(2) *Prospective provenance:* To map the intended processing of data. Prospective provenance comes into play as a plan of the steps to be applied to the data in question.[5].
(3) *Permitted provenance:* To specify the set of activities and processes that **could** happen when the data in question are processed.
(4) *Prescriptive provenance:* To specify the set of activities and processes that **should** happen when the data in question are processed.[6].
(5) *Proscriptive provenance:* To specify the set of activities or processes that **should not** happen when data in question are processed.

The combination of retrospective and prospective provenance provides a complete description of data situation (i.e. the existing and intended/proposed processing). The results of this combination can be compared with the P3 provenance to highlight whether particular processing is allowed, required or prohibited according to the ADF. RP4 provides the conceptual architecture that enables

[4]The term "policy" was also considered here to represent these concepts. However, this is potentially confusing as one of the inputs into permitted, prescriptive and proscriptive provenance are organisational policies (as one type of DGI). We wanted to keep the provenance ontology separate from the specific instruments that might comprise it.
[5]Note that, the two core terms retrospective and prospective provenance are effectively the same as the equivalently named concepts in the workflow systems literature [25–27]. The application is different here but the underlying meaning is the same.
[6]We note that the term prescriptive provenance has been used within a streaming context, where it denotes a prescribed collection of specific provenance components to reduce overall capture size [39]; we acknowledge this slight term overload but here we use prescriptive provenance to mean statements of obligation.

the user to address the essential questions (WHO-WHAT-HOW-WHEN-WHERE) about the data and their environments that the ADF requires:

- WHO refers to the parties (individuals, groups, organisations, etc) involved in the data situation.
- WHAT refers to the nature of the data being processed in the data situation.
- WHERE refers to the environments in which the data are stored, processed, etc. during the lifetime of the data situation.
- HOW refers to the data processing activities such as: collect, capture, store, share, etc.
- WHEN refers to the timing of the activities.

7.1.3 Data Governance Instruments (DGIs). P3 provenance could be created in a similar manner to [38, 40]. However, we anticipate that P3 provenance statements can be generated using a DGI template to capture the details that are integral to the decision to share or not.

There are potentially multiple DGIs that are applicable to any given data situation. For instance, legislation such as GDPR, the corresponding parties' contracts and unilateral documents such as data sharing policies can all be interpreted as a set of permitted, prescriptive and proscriptive provenance statements.

Other DGI documents that effectively govern organisational behaviour and data management may also contribute to P3 provenance statements. For instance, consent forms that may well have been signed by the data subjects, who are the source of the data, would become the basis of P3 provenance statements since they clearly identify how the data subjects' data should be used and handled.

To formalise this: a Data Governance Instrument (DGI) is a document that constrains, enables and manages a given data situation (and specifically the relationships between the parties involved and the behaviour of the parties with respect of the data) [12]. Through the DGIs, the parties involved agree on and to policies and procedures that permit, prescribe and proscribe WHO carries out WHAT processes on WHAT data in WHICH data environments. So, rather than restricting the data, a DGI aims to specify the set of activities that could, should or should not happen during the lifetime of the data situation.

The term DGI encompasses three main categories of documents: (i) multilateral agreements between parties such as Terms of Use/Service, User Agreements, licences and data sharing agreements; (ii) unilateral (institutional) statements such as Privacy Policies and operating procedures and (iii) Jurisdictional instruments, such as laws, regulations or conventions (such as USA's Health Insurance Portability and Accountability Act (HIPAA) or the EU's GDPR) and also codes of practice and guidance documents produced by regulators.

Note that, DGIs signed between parties about data exchange could align with or may override any already existing P3 provenance, and may also be used to create additional P3 provenance statements. Additionally, in many data situations there exist multiple DGIs and these may even conflict with one another in some circumstances.

7.2 Study Design

With a goal of enabling automated reasoning for data sharing, as depicted in Figure 5, we propose that P3 provenance should be created through analysis of the relevant DGIs. To develop a framework for this process we investigated twelve DGIs to extract the features emphasised in them as well as categorising and organising them as the DGI template. Then, we tested the validity and reliability of this template on a separate sample of ten DGIs. Four of the authors provided separate functions in the process. ME selected the DGIs. FR analysed the DGIs in order to develop the framework and

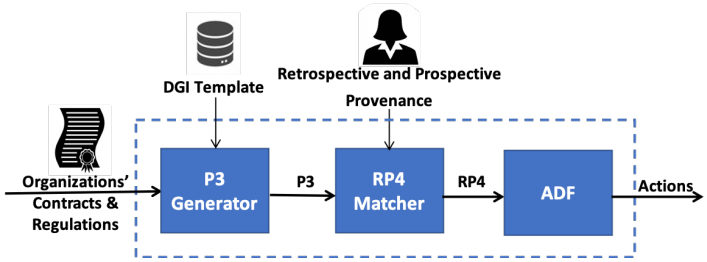

Fig. 5. The concept of operations for using Data Governance Instruments (DGI) with RP4 provenance to facilitate release and anonymisation decisions.

the tool for subsequent extraction. AC and TB independently tested the template against the fresh sample of DGIs.

7.2.1 Data. Our data for this study were two sets of DGIs. In common with much document analysis research, obtaining a representative sample of such instruments is clearly infeasible. In selecting the DGIs for the study, we were instead motivated to ensure a diversity of functions. The function types that we have covered are as follows:

- User agreements
- Data use licenses
- Privacy Policies
- Data Processing Policies
- Confidentiality Agreements
- Information Governance Checklists
- Guidance documentation for the interpretation of regulations
- Data Sharing Agreements

The goal in setting this up is not to produce an unequivocal ontology of DGIs (although that may be a valid line of enquiry for future work), rather we aimed to giving good coverage of the domain.

The DGIs used in the development set were as follows:

- ◇ **DGI_1:**The guidance published by the UKs Information Commissioner's Office (ICO) for DGI's between data controllers and data processors based on the UK's GDPR[7]. This DGI discusses the responsibilities and liabilities of controllers and processors to help them understand the required statements in their agreed DGI.
- ◇ **DGI_2:**A guidance document called "External guidance on the implementation of the European Medicines Agency policy on the publication of clinical data for medicinal products for human use" produced by the European Medicines Agency (EMA) [8].
- ◇ **DGI_3:**A user agreement published by PayPal[9] related to legally using the PayPal services.
- ◇ **DGI_4:**The privacy notice published by the University of Manchester for registered students[10]. This DGI informs students of how the University of Manchester collects, maintains and uses the students' personal information during their education and afterwards.

[7]https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/contracts-and-liabilities-between-controllers-and-processors-multi/
[8]https://www.ema.europa.eu/documents/regulatory-procedural-guideline/external-guidance-implementation-european-medicines-agency-policy-publication-clinical-data_en-3.pdf
[9]https://www.paypalobjects.com/marketing/ua/pdf/GB/en/ua-073021-v1.pdf
[10]http://documents.manchester.ac.uk/display.aspx?DocID=37237

- **DGI_5:**A research data management policy published by the University of Manchester for the research data that specifies the responsibilities of the university, its staff and students with regard to management of research data[11].
- **DGI_6:**The Information Governance Checklist (including Data Protection Impact Assessment) organised by the Information Governance Office at the University of Manchester to facilitate an assessment of the risks associated with the processing of university information[12].
- **DGI_7:**The UK's Open Government Licence for parties in the public sector to enable the re-use of their data under a common open licence[13].
- **DGI_8:**The terms of Use of Administrative Data Research UK (ADRN) services that is applicable to the researchers and institutional guarantors[14].
- **DGI_9:**UK Anonymisation Network Policy on Data Handling and Management that entails the UK Anonymisation Network (UKAN) policy for consultancy projects[15].
- **DGI_{10}:**The confidentiality agreement between Covidien LP, a Medtronic company, on behalf of itself and its worldwide affiliates, having a place of business[16].

Other than the open access DGIs, we also explored several restricted access DGIs agreed between two parties such as some agreed DGIs between UKAN and its contracting parties which we specifically refer to two of them in this paper as **DGI_{11}** and **DGI_{12}**.

7.2.2 Creating the DGI Template. Using an iterative process on the selected DGIs, the DGI template are created.

In the initial step, the DGIs were explored for common features (that were relevant to the P3 framework and/or anonymisation decision making). In this way, themes were noted and keywords where relevant were identified. In the follow on step, the features that were identified in the first step were organised into a functional hierarchy which we refer to as *features* and *feature categories*.

7.2.3 Testing the reliability DGI Template. In order to test the reliability of the DGI template, two of our team members applied DGI template to a DGI test set comprising ten DGIs. In other words, each tester worked independently through the test DGIs taking the relevant text and pasting it into the template. So for each of the ten test DGIs we had two completed templates. Afterwards, we assessed the correspondence between the completed template pairs.

7.3 Results

7.3.1 Exploratory Step. We noticed that the first part of the DGIs commonly describe the agents (individuals, organisations, etc.) involved in the DGI. It mostly contains the keywords *company*, *party*, *disclosing party*, *sending party*, *data controller*, *data processor*, *data subject*, *user*, *individual*, *entity* followed with the keywords *contact*, *email address* and their derivations or synonyms.

After the agent introduction, the next part of DGIs is often related to the data that will be shared (or otherwise processed). We realised that this part usually contains the keywords *personal information*, *personal data*, *sensitive data*, *data*, *information*, *identifier* and their derivations or synonyms.

Occasionally, some DGIs entail the specification about the infrastructure that contains the keywords *store*, *safekeeping*, *safeguard*, *archive*, *storage*, *repository*, *computer*, *database*, *IT*, *system*,

[11]https://documents.manchester.ac.uk/display.aspx?DocID=33802
[12]https://www.staffnet.manchester.ac.uk/igo/
[13]https://www.nationalarchives.gov.uk/doc/open-government-licence/version/3/
[14]https://www.adruk.org
[15]https://ukanon.net
[16]https://www.medtronic.com/covidien/

computer, *environment*, *technology*, *infrastructure*, *platform* followed with the keywords *anti-virus*, *protection*, *secure*, *safe*, *confidential*.

Some DGIs explain explicitly about the data capturing that usually contains the keywords *obtain*, *capture* followed with the keywords *data*, *information* and their derivations or synonyms.

In some DGIs, there may be a clear statement about data transference. Usually, the related part contains the keywords *transfer*, *collect*, *move*, *download*, *upload*, *receive* followed with the keywords *data*, *information* and their derivations or synonyms.

One of the important part of DGIs explains the list of activities that aims to take place on the data. This part contains the keywords *store*, *share*, *receive*, *upload*, *download*, *use* and *analyse*, *process*, *reverse engineer* followed with the keywords *purpose* and their derivations or synonyms.

In some DGIs, there may be an explanation about the requirements or expectations about data retention or deletion in the receiving party's databases. The related part usually contains the keywords *retain*, *keep*, *destroy*, *remove* followed with *database* or *years*, *months*, *period*, *end of the project*, *permanently* and their derivations/synonyms.

More specifically, some DGIs explain about the probable activity not listed before but may need to be happened on the data. Usually, the related part contains the keywords *secondary use*, *re-use* or the words *use the data for new purposes*, *use the data for other purposes* followed with the keywords *consent*, *notice*, *refuse*, *legal basis* and their derivations/synonyms.

Another important part that could be in included in the DGI is about sharing data with third parties. Usually, the related part contains the keywords *disclose*, *disclosure*, *share*, *use* followed with the keyword *third party* and their derivations/synonyms.

In some DGIs, there may be an explicit statement about public release of data. Usually, this part contains the keywords *publish*, *broadcast*, *public release*, *public*, *open* and their derivations/ synonyms.

Occasionally, the receiving party is explicitly given permission for disclosing data to the court (if required). Usually, the related part contains the keywords *legal*, *court*, *jurisdiction*, *administrative agency* and their derivations/synonyms.

More specifically, some DGIs give information relating to the receiving party's actions in event of the accidental or inadvertent disclosure of data. Usually, the related part of the DGI contains the keywords *inadvertently*, *accidently*, *incident* followed with the keywords *disclose*, *breach*, *identify*, *link*, *unauthorised access*, *unauthorised use*, *unauthorised process* and their derivations/synonyms.

The DGIs may include the obligations of one or both parties in the future. Usually, the related part of the DGI contains the keywords *roles*, *responsibilities*, *obligation*, *agree*, *bound* followed with the agent's names mentioned in Feature 1 or the keywords *disclosing party*, *data controller*, *receiving party*, *data processor*, *company* and their derivations/synonyms.

Some DGIs entail the rights of each party in some detail. Usually, the related part of the DGI contains the keyword *rights* followed with the agent's names mentioned in Feature 1 or the keywords *disclosing party*, *data controller*, *receiving party*, *data processor*, *company* and their derivations/synonyms.

More specifically, some DGIs clarify the rights of Data Subjects. Usually, the related part contains the keyword *rights* followed with the agent's names for data subjects mentioned in Feature 1 followed with the keywords *access*, *correct*, *restrict*, *suspend*, *stop* and their derivations/synonyms.

Most DGIs state their compliance with regulations. Usually, the related part of the DGI contains the keywords *legislation*, *liable*, *law*, *regulation*, *compliance* and their derivations/synonyms) followed with a legislation name such as *Data Protection Act* (DPA), GDPR etc.).

The last part of most DGIs usually state the period during which the DGI is effective. The related part will usually contain the keywords *effective date*, *expiry* followed with the keywords *years*, *months*, *period* and their derivations/synonyms.

The last part of DGIs could also include the conditions for termination of each party from the agreed DGI. Usually, containing the keyword *terminate* and its derivations/synonyms.

DGIs may also contain miscellaneous not covered by the categories above statements. Typical keywords include *miscellaneous*, *read*, *understood* and *agree* and their derivations/synonyms.

DGIs may finish with details of the current version. Usually, the related part contains the keywords *version*, *changes*, *revise*, *review*, *waive*, *modify*, *amend* followed with the keywords *document*, *agreement*, *DGI* and their derivations/synonyms.

7.3.2 Feature Extraction Step. Below is a list of the Features Extracted during the Feature extraction stage. All of these features were found in at least one of the base set of DGIs. All of the features correspond to information within a DGI that is either directly pertinent to provenance or is relevant to the ADF.

The Features are grouped into five categories:

- **Agents:** Information about the people involved in or affected by the DGI or the related data processing.
- **Entities:** The data and their environment.
- **Activities:** Different activities that will, could, should or should not be carried out on the data
- **Rights and Obligations:** Secondary considerations, impacts and allowances arising from the DGI or the data processing.
- **Metadata:** Information relating to the DGI itself rather than to the data (processing).

Feature category Agents (Ag)

Feature Ag1 (Main Agents): Information related to the main individuals, groups and/or organisations involved in or affected by the DGI, this would include but is not restricted to any parties to the DGI. This might include:

- Data subject(s) whom the data is about. The categories of data subject such as the children, disabled individuals, etc., could also be considered here.
- The disclosing party, which could be the data controller(s) in the GDPR context, who determines the essential means and purposes of processing of data subjects' personal data.
- The receiving party, which could be the data processor(s) in the GDPR context, who processes personal data in a manner decided by the disclosing party.
- Data users who process anonymised data but do not determine the means or purposes of processing.

For instance, feature Ag1 is stated in DGI_1 as *"This notice provides information about the use of personal information while you are a registered student of the University of Manchester, including a student enrolled on a programme at a University of Manchester Worldwide global centre."*.

Note that in DGI_1, the disclosing party and the data subject are one and the same.

Feature category Entities (E)

Feature E1 (Data): This will usually be in the form of quasi-metadata related to the data that the disclosing party wants to share with the receiving party. This feature typically contains the following properties:

- The high-level properties of the data such as their type (statistics, text, etc.), categories (racial, ethnic origin, political opinions, etc), age, quality, etc.

- ◇ The low-level properties of the data items, which could be the data subject's name, email address, date of birth, behavioral information, etc.

The specification of the data might be general. For example, feature E1 is stated in DGI_2 as *"Research data is the evidence that underpins the answer to the research question and can be used to validate findings. This might be quantitative information or qualitative statements collected by researchers in the course of their work by experimentation, observation, modelling, interview or other methods, or information derived from existing evidence. Research data may take the form of numbers, symbols, text, images or sounds, including computer code, annotated fieldwork observations, or a descriptive record of a physical sample."*. The data specification might also be specific in terms of the data involved.

Feature E2 (Infrastructure): Details of the infrastructure that the receiving party shall (be required to) use to safeguard the data which may include the hardware, operating systems, database type, physical location and security etc.

For example, feature E2 is stated in DGI_3 as *"The data will be stored on a non-network computer. The computer will be password protected. The room in which the computer is stationed will be locked when unoccupied."*.

Feature category Activities (Ac)

Feature Ac1 (Data Collection/Capture): Detailed information related to activity that has been or will be carried out (often by the disclosing party) in order to obtain the data. Such data capture processes could either be direct from the data subjects (eg a survey or administrative process) or indirect (eg web scraping).

For example, feature Ac1 is stated in DGI_1 as *"5.1. We obtain personal data from you ..., as a student, including annually at registration and during the course of your relationship with us when accessing or using any of our services such as financial support, careers advice or counselling services. We also receive personal data about you from other organisations when you make an application to study at the University and this information will form the basis of your student record when you become a registered student of the University e.g. from UCAS and/or from individual referees..."*

Feature Ac2 (Data Transfer): The information related to the how the data will be transferred between agents (usually the disclosing and receiving parties). This may cover hardware and software infrastructure but also governance arrangements and may refer to actions taken by either or both parties involved in the transfer.

For instance, feature Ac2 is stated in DGI_{12} as *"The data will be uploaded to PharmComp secure access platform. It will be downloaded directly onto an external hard drive."*

Note that, if the disclosing party is the data subject, data capture and data transfer will be isomorphic. As a result, feature 6 and 7 would have the same meaning.

Feature Ac3 (Data Process): The list of the intended data processes. A non-exhaustive list of these is: store, share, receive, upload, download, publish (share the data in the public domain), delete (deleting the data permanently), use (analyse or "process" in the GDPR context, which generates a new version of data from the original data) and apply disclosure control [17]. Note that, from the above atomic activities, more complex data activities might be derivable using standard forms

[17]The term "disclosure control" is used here rather than "anonymise" as the ADF employs functional anonymisation which considers both the data and their environment whereas disclosure control is simply an operation applied to data.

and patterns. However, the atomic activities are contextualised by the specific data situation. As a result, interpreting particular patterns for effective anonymisation decision-making support will be a non-trivial task. This feature may include the following information for each activity:

- ◇ Parts of the data for which the activity takes place.
- ◇ The purpose of the activity.

For instance, feature Ac3 is stated in DGI_1 as *"The University will process your personal information, including where applicable your image, for a range of contractual, statutory or public interest purposes, including the following:*

To deliver and administer your education, record the details of your studies (including any placements with external organisations), and determine/confirm your academic achievements (e.g. results, prizes)

To administer student related policies and procedures including appeals, complaints, grievances, disciplinary matters ...".

Feature Ac4 (Retention and Deletion): Information about the time period for retention of the data until it is permanently deleted of the data from the receiving party's systems.

For instance, feature Ac4 is stated in DGI_2 as *"The minimum archive duration for research data and records are subject to the University's Records Management Policy and Records Retention Schedule.".*

Feature Ac5 (New Activity): Information about whether activities that are not listed under feature 5 or new purposes for the data that may arise in the future are allowed or not. DGIs may:

- ◇ Refuse permission for any new activity or new purpose.
- ◇ Refuse permission for specific new activities or purposes.
- ◇ Require the receiving party to notify or gain permission from the disclosing party for the new activity or purpose.

The DGI may describe specific other activities but it may not. For instance, feature Ac5 is stated in DGI_4 as *"This Agreement does not, and shall not be construed to, constitute the grant to the Receiving Party of (a) any right or license to use any Confidential Information for any purpose other than for the Purpose in accordance with this Agreement."*

Feature Category Sharing and Disclosure (S)

Feature S1 (Sharing with third Parties): Information related to sharing of the data with third parties. This feature may include the following information:

- ◇ A specification of third party with whom the data is (to be) shared.
- ◇ The data items to be shared.
- ◇ The purpose of the sharing.
- ◇ The anonymisation methods including disclosure control techniques (that should be) employed by the receiving party before sharing the data with third parties.
- ◇ The activities that will/can be carried out on the data by the third party.
- ◇ The deadline for the deletion of the data permanently from the third party's systems.
- ◇ The IT infrastructure that the third party shall take to safeguard the confidentiality of the data.

For instance, feature S1 is stated in DGI_4 as *"The Receiving Party agrees that it shall only use the Confidential Information for the Purpose and shall not disclose such Confidential Information to third parties except, and in each case solely in connection with the Purpose, to (i) its and its Affiliates' directors, officers, employees, agents, consultants, advisors and service providers (Representatives) who are informed of the confidential nature of the Confidential Information and are bound by obligations*

of confidentiality and non-use no less restrictive than those contained herein and (ii) its potential co-investors in the Company already bound by confidentiality obligations with the Company."

It is worth noting that the substance of this feature refers to many of the other features. This is because sharing with third parties would almost certainly itself be governed by a (yet to be created) DGI. This feature therefore specifies in advance some of the contents of such a DGI.

Feature S2 (Public Release): The information related to the public release of data that may be done by the receiving party. This feature may include the following information:

- The data items to be published.
- The purpose of the publishing.
- The disclosure control techniques that must be applied by the receiving party before data are published.

For example, feature S2 is stated in DGI_{12} as *"Work from the project may be written up in a journal article to be agreed in consultation with PharComp. Any output to be published will be checked to ensure that they are not disclosive."*

Feature S3 (Disclosure to Court): Information related to the situation where the receiving party receives a request to or becomes legally obliged to disclose the data (either fully or partially) to a court, administrative agency or other agency with relevant statutory authority.

For instance, feature S3 is stated in DGI_4 as *"The Receiving Party may disclose the Confidential Information if compelled to do so by a court, administrative agency or other tribunal of competent jurisdiction; provided, however, that in such case the Receiving Party shall provide prompt written notice to Company in advance of the disclosure, to the extent reasonably possible, so that Company may seek a protective order or other remedy from said court or tribunal and the Receiving Party shall only disclose that portion of such Confidential Information that, in the opinion of its legal counsel is required to be disclosed. In the event that Confidential Information is required to be disclosed pursuant to this paragraph, the Receiving Party shall exercise all reasonable efforts to obtain reliable assurance that confidential treatment will be accorded the Confidential Information."*

Feature S4 (Inadvertent disclosure): The information related to the situation where the receiving party inadvertently identifies any individuals from the data.

For example, feature S4 is stated in DGI_5 as *"Disclosure of Information, if in written form, shall be stamped CONFIDENTIAL or bear markings of like import. If verbally or visually disclosed, such as through a facility tour, Information shall be summarised in writing, to the extent practicable, within 60 days of disclosure, stamped in the manner indicated above, and forwarded to the recipient. Information will be considered confidential by both Parties and will be received and handled as set forth herein."*

Feature S5 (Intentional Disclosure): Information related to the situation that the receiving party intentionally attempts to identify any individuals from the data.

For instance, feature S5 is stated in DGI_4 as *"The Parties understand and agree that any disclosure or misappropriation of any of the Confidential Information in violation of this Agreement may cause the Company irreparable harm, the amount of which may be difficult to ascertain, and therefore agree that the Company shall have the right to apply to a court of competent jurisdiction for specific performance and/or an order restraining and enjoining any such further disclosure or breach and for such other relief as the Company shall deem appropriate. Such right of the Company is to be in addition to the remedies otherwise available to the Company at law or in equity."*

#	Feature Name	DGI 1	DGI 2	DGI 3	DGI 4	DGI 5	DGI 6	DGI 7	DGI 8	DGI 9	DGI 10
1	Parties	+	+	+		+	+	+	+	+	+
2	Data	+	+	+	+	+	+	+	+	-	+
3	Deletion	+	-	+	-	+	+	+	+	+	+
4	Security Technique	+	+	+	-	+	+	-	+	+	+
5	Activities	+	+	+	+	+	+	+	+	+	+
6	Capture	-	-	+	+	+	+	-	+	-	+
7	Collect	-	-	+	-	+	+	-	+	+	+
8	New Activity	-	+	-	-	-	-	+	-	-	+
9	Third Party	+	+	+	-	+	+	+	+	-	+
10	Public Release	+	+	+	+	-	+	+	-	-	+
11	Disclose to Court	+	-	-	+	+	-	+	-	-	+
12	Disclose Inadvertently	+	+	-	-	+	-	+	+	-	-
13	Disclose Intentionally	+	+	-	+	-	-	+	-	-	-
14	Recovery (Lost)	-	-	-	-	+	-	-	+	-	-
15	Recovery (Damage)	-	-	-	-	+	-	-	+	-	-
16	Recovery (Destroy)	-	-	-	-	+	-	-	-	-	-
17	Obligation (Sending)	+	-	-	+	-	-	-	-	-	-
18	Obligation (Receiving)	+	+	-	+	-	-	-	-	-	-
19	Right (Disclosing)	+	+	-	-	-	+	+	+	-	+
20	Right (Receiving)	+	-	-	-	-	+	-	+	-	-
21	Right (DS)	-	-	-	+	+	-	-	+	-	+
22	Regulation	+	+	+	+	+	+	+	+	-	+
23	Effective date	-	-	+	+	+	+	+	+	+	+
24	Terminate Date	+	-	-	-	-	-	-	+	-	-
25	Miscellaneous	+	+	+	+	-	-	+	+	-	+
26	Version History	-	-	+	+	+	+	+	+	+	+

Fig. 6. The comparison of the selected DGIs according to the extracted features, assuming support of the feature by + and failing to support the feature by -

Feature category Rights and Obligations (RO)

Feature RO1 (Obligations of Disclosing Party): Other obligations of the disclosing party that are not specified in other features.

For example, feature RO1 is stated in DGI_5 as *"Each Party acknowledges and agrees that the other Party would not have an adequate remedy at law and would be irreparably harmed in the event that any provision of this Agreement is not performed in accordance with its terms or is otherwise breached. Accordingly, each Party agrees that the other Party shall be entitled to seek equitable relief, including an injunction and/or specific performance, in the event of any breach of the provisions of this Agreement, in addition to all other remedies at law or in equity."*

Feature RO2 (Obligations of Receiving Party): Other obligations related to the receiving party that are not specified in other features. For instance, the information related to the the recovery actions to be taken by the receiving party in the event of data loss, destruction or damage which commonly describes the back-up and recovery process in the advent of this event.

For example, feature 14 is stated in DGI_{12} as *"UKAN is looking at the data at the metadata level; if the data is damaged, UKAN will obtain replacement copies from PharmComp".*

7.3.1 **Feature RO3 (Rights of Disclosing Party):** Other rights related to the disclosing party that are not specified in other features.

For instance, feature 16 is stated in DGI_2 as *"The University of Manchester is responsible for: Empowerment of organisational units, providing appropriate means and resources for research support operations, the upkeep of services, organisational units, infrastructures, and researcher education..."*

Feature RO4 (Rights of Receiving Party): Other rights related to the receiving party that are not specified in other features.

For example, feature 17 is stated in DGI_5 as *"Nothing in this Agreement shall constitute a waiver of any patent or other rights either Party may have in its Information, nor shall it constitute the grant of any license or any right under any patent by either Party to the other."*

Feature RO5 (Rights of Data Subject): The rights related to the data subjects. This feature may include the following parts:

- ◇ The right to access to their personal information.
- ◇ The right to correct their personal information.
- ◇ The right to have their personal information deleted.
- ◇ The right to restrict or suspend the processing of their personal information.
- ◇ The right to seize transferring of their personal information to the receiving party.

For example, in the DGI_1 that the data subjects are the disclosing party, feature 18 is stated as *"Under certain circumstances, by law you have the right to: 13.1. Request access to your personal information (commonly known as a "data subject access request"). This enables you to receive a copy of the personal information we hold about you and to check that we are lawfully processing it. 13.2. Request correction of the personal information that we hold about you. This enables you to have any incomplete or inaccurate information we hold about you corrected ..."*

Feature Category Metadata (MD)

Feature M1 (Regulation): Information related to the laws and regulations applied to the operation of the DGI.

For instance, feature M1 is stated in DGI_5 as *"This Agreement shall be interpreted and construed according to the laws of the Commonwealth of Massachusetts USA, without regard to the choice of law provisions thereof."*

Feature M2 (Period): The information related to the start date (effective date) when the DGI comes into force and the duration of the time that the DGI will be effective.

For instance, feature M2 is stated in DGI_4 as *"This non-disclosure agreement (the Agreement) is entered into as of 19/03/2021 (the Effective Date) by ...*

The confidentiality and non-use obligations under this Agreement shall continue for five (5) years from the Effective Date."

Feature M3 (Termination): Information related to the Termination of the DGI including the situation where either party wishes to terminate the agreed DGI at any time before the specified termination.

For example, feature M3 is stated in DGI 5 as *"Either Party may terminate this Agreement at any time by providing written notice of termination to the other Party."*

Feature M4 (Execution): The information related to the miscellaneous statement. This feature may describe how the DGI is agreed and signed while emphasising the complete understanding of DGI by both parties.

For example, feature M4 is stated in DGI_4 as *"This Agreement may be executed in writing or in electronic form (such as Skribble, DocuSign or AdobeSign, or which contains an electronic scan of the signature) and be delivered by post, courier or email; the counterpart so executed and delivered shall be deemed to have been duly executed and validly delivered and be valid and effective for all purposes. This Agreement expresses the full and complete understanding of the Parties with respect to the subject matter hereof and supersehdes all prior or contemporaneous proposals, agreements, representations and understandings, whether written or oral, with respect to the subject matter."*

Feature M5 (Version History): Information about the version history of the DGI; if the DGI may be modified in future. This feature may include the updated parts and the date of the amendment.

For example, feature M5 is stated in DGI_2 as *"Version amendment history:*
Version Date: 1.2 March 2021
Previous Review Dates: March 2018, February 2019
Next Review Date: March 2022"

7.4 Evaluation Step

In Section 7.3.1 sets of keywords were identified that characterised each feature described in Section 7.3.2. Five features were associated with a single set of keywords. The remainder were characterised by two sets of keywords, where one of the keywords in the first set must be followed (not necessarily immediately) by a keyword from the second set. In addition, Section 7.3.1 allowed for synonyms or derived terms (hyponyms) as well as the specified keywords. For instance, "intent" would be a synonym of "purpose" and might be used to convey the same meaning. Also, a "shipper" is a type of "company" and might be used in DGIs relating to the shipping industry.

We refer to the above characterisations as rules. If text contains words from each of the corresponding sets of keywords (or their synonyms / hyponyms) and in the correct order, then we say that the text complies with the rule for the feature. Thus we can attempt to identify sections of documents that possess the features detailed in Section 7.3.2 by checking for compliance.

The analysis of a document comprises 3 steps:

(1) Extract text from document
(2) Process text to generate suitable data for analysis
(3) Perform the analysis

7.4.1 Text Extraction. Documents might come in a variety of file formats. Many are PDF, but user agreements published on-line will often be HTML. Plain text formats (such as HTML) contain document data in an easily accessible format, whereas PDFs contain information regarding the positions of characters on a page. Thus arranging the collection of characters into meaningful words, sentences, paragraphs, sections and so on requires much more work, and is difficult to do reliably. The Python library Textract (https://github.com/deanmalmgren/textract) can be used to extract text from a wide variety of file formats. It was used to extract text as a single string from

the files that we considered. As the considered files were PDFs some additional processing of the returned strings was necessary. The details are not presented here.

7.4.2 Text Processing. The resulting single string was lower cased and processed into substrings, each substring representing a relevant chunk of text (sentence, paragraph or section). These were further processed using the Natural Language Toolkit (https://github.com/nltk/nltk). This was used to split each substring into a list of words (and punctuation characters that have no impact on subsequent analysis). The Natural Language Toolkit (NLTK) bindings to WordNet (https://wordnet.princeton.edu/) were used to map any synonyms / hyponyms of our keywords to the keywords themselves. (This is a one-time operation that does not need repeating unless the keywords are changed.) The synonym / hyponym mappings were used to find synonyms / hyponyms and replace them with the relevant keyword.

7.4.3 Analysis. Each chunk of text (list of words and punctuation characters) was tested for each rule. Initially each chunk was scanned and the indices of any keywords found were added to lists, so two sets of keywords would result in two lists of indices. If either list was empty, then the chunk did not comply with the rule. If neither list was empty, then we simply needed to check that there existed some index in the first list that was less than some index in the second. If, and only if, this was the case, then the chunk complied with the rule. (For features with only a single set of keywords compliance was established by simply finding a single keyword.)

A second approach was developed which generated a compliance weight. The weight is simply the number of ways in which compliance can be established. That is, the number of distinct ways that pairs of indices (i,j) can be chosen from our lists of indices, i from the first list and j from the second, such that $i<j$. The idea behind this is that compliance is less likely to be due to chance if it can be established in multiple ways.

Table 9 shows the results of analysing an end user licence agreement using the features and rules described in Sections 7.3.1 and 7.4.1. The values shown are compliance weights, and features (columns) and chunks (rows) that have all zero weights have been suppressed. The largest weight is for feature Ac1 and the paragraph indexed 31. The paragraph is reproduced below.

"To preserve at all times the confidentiality of information pertaining to individuals and/or households in the data collections where the information is not in the public domain. Not to use the data to attempt to obtain or derive information relating specifically to an identifiable individual or household, nor to claim to have obtained or derived such information. In addition, to preserve the confidentiality of information about, or supplied by, organisations recorded in the data collections. This includes the use or attempt to use the data collections to compromise or otherwise infringe the confidentiality of individuals, households or organisations."

This clearly relates to data collection/capture. Preliminary experimentation with files that are not DGIs (for example, CVs) shows that they tend to generate few "false positives". So a similar approach might be used for identifying DGIs, or classifying them into sub-types.

7.4.4 Issues. The whole process of extracting text, processing text and performing meaningful analysis is far from error-free. Extracting text from binary files such as PDFs is more difficult than extracting text from files that are already in a plain text format. Although libraries such as Textract can be a great help they tend to extract all text from PDFs, including headers and footers. They also tend to process files page by page, not reliably distinguishing between ends of paragraphs and ends of pages. Sections (that might contain a number of paragraphs) are harder to identify than paragraphs. Section headings can be difficult to distinguish from short paragraphs. Fully automated and accurate text extraction will probably require a dedicated, task-specific PDF parser. Such a parser is currently under development.

	Text	Ac1	Ac2	Ac3	RO3	RO4	RO5
19	To use and to make personal copies of any part...	2	0	0	0	0	0
27	To give access to the data collections, in who...	0	4	4	0	0	0
29	To ensure that the means of access to the data...	0	0	0	1	1	0
31	To preserve at all times the confidentiality o...	10	4	0	0	0	0
37	That the members of the Data Team may hold and...	0	3	1	0	0	0
41	That any personal data submitted by me is accu...	0	0	1	0	0	0
45	At the conclusion of my research (or if earlie...	0	8	2	0	0	0
51	I understand that non-compliance with any of t...	0	0	0	1	1	2
57	The members of the Data Team accept no liabili...	0	2	1	2	2	0
59	Whilst steps have been taken to ensure all lic...	0	3	0	0	0	0
61	I agree to indemnify and shall keep indemnifie...	0	2	0	0	0	0
62	If the whole or any part of a provision of thi...	0	0	2	0	0	0
80	to give access to the data collections only to...	0	2	0	0	0	0
88	to send the UK Data Service bibliographic deta...	0	2	0	0	0	0

Table 9. Example analysis of the paragraphs of the UK Data Service End User Licence

The generation of synonyms and hyponyms is not fully automatic. Words can have several meanings / definitions, and we are generally only interested in one. This requires some checking and possible manual intervention when generating mappings of synonyms / hyponyms to keywords. For example, the word ”process“ is present in the keywords for Ac3. But WordNet contains 13 definitions for process and only one is the intended definition. If synonyms and hyponyms for a (particular) wrong definition were chosen, then words such as ”march“ and ”file“ would be replaced by the word ”process“. This might generate false positives. Of course, constructing these mappings is a one-time process for a given collection of keywords; so a little checking / manual intervention is not unreasonable. A more difficult issue occurs when a keyword (or a synonym or hyponym of the keyword) is present in the text, but not with the relevant meaning. For instance ”treat“ is a synonym for the relevant definition of ”process“, and ”propagate“ is a hyponym. Ideally we would like to infer the sense of any of these words if they appeared in text, and only allow them to influence analysis if used in the appropriate sense. Doing this manually would be time consuming, and would need to be performed for each document analysed. Of course, parsing sentences to extract meaning is a fundamental aspect of natural language processing, but we have as yet not tried to incorporate this in our analysis.

The analysis approach itself is a little *ad hoc*. There might well be other aspects of a text chunk that could be exploited. If we constructed a corpus of annotated DGIs, then there is a range of machine learning approaches that could be applied to the problem. The current approach is also a little sensitive to the size of the chunk. For example, large paragraphs will tend to generate a larger number of positive results (whether true or false) simply because they contain more words. Nevertheless, preliminary results using imperfect text extraction, imperfect processing, and a relatively *ad hoc* analysis approach are encouraging.

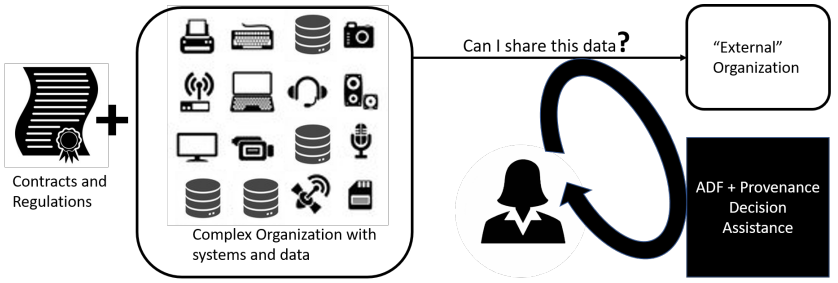

Fig. 7. Support for the decision maker around information exchange.

8 CREATION OF A PROTOTYPE THAT PROVIDES ANALYSIS FOR SHARING CONCERNS TO DECISION MAKERS USING *RP*[4] PROVENANCE-ENCODED DATA AND ADF RULES

Currently, this decision making is performed manually by a human actor within an organisation (see Figure 7 white boxes) who must understand the regulations and contracts that apply to the data, the complex set of data and system relationships within an organisation and their impact on future sharing decisions. The goal of this work is to provide a support tool (see Figure 7) to improve decision making about information sharing and anonymisation.

Our approach to this takes multiple steps. First, we utilise the concepts from the *Anonymisation Decision Framework (ADF)* which operationalises the processes of functional anonymisation [1, 3], as described further in Section 2.2. We propose the use of data provenance to represent these concepts in a machine-interpretable way that allows automated reasoning over the factors that influence data sharing decisions. Data provenance has already been applied in the modelling of similar problems such as quantitatively modelling the pros and cons of data sharing between two organisations [4], situation awareness and decision making [41], controlling of direct and indirect data flows [42], big data security and privacy [5].

In order to provide the capacity for better reasoning, we first map the concepts in provenance to those required by the ADF to ensure the essential concepts can be modelled. Based on that evaluation, we identify a concept required for good information sharing reasoning, **data environments**, that is currently not representable in the W3C PROV, the provenance interperability standard [13, 14]. We briefly analyse possible mechanisms to extend the W3C PROV in order to support this essential element.

Using these building blocks, it is then possible to create a semi-automated tool to assist decision makers in better information sharing. The tool takes in contracts, regulations and other Data Governance Instruments (DGIs). It also takes in known historical data movement, processes, actors and data environments. Using the DGIs, it creates a set of valid information sharing instances that a decision maker can review and choose among.

9 CONCLUDING REMARKS

In the knowledge economy, large amounts of personal and corporate-sensitive data are collected to support decision-making, policy analytics, service delivery etc. There are inherent confidentiality and privacy risks in sharing such data, and yet organisations often need or want to do so for contractual, or business processing reasons. The problem of identifying possible disclosure risks[18],

[18] these are risks that information may leaked about identifiable individuals

particularly before they occur, is a core concern, and methods to identify and mitigate such risk run from statistical techniques to manipulate the data [43] through attempts to provide provable guarantees [44] to holistic frameworks for protecting data [3, 12].

It is now widely accepted that disclosure risk resides not just in the data themselves but in the relationship between the data and their environment [1, 45, 46]. 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" [47] and refer to the relationship between data and their environment as the data situation. The complexity with data sharing is that one is moving data from one environment to another and therefore changing risk.

The current project has carried out a thorough exploration of the relationship between anonymisation and data provenance and in particular between the anonymisation decision making framework and the provenance standard PROV.

A full list of outputs and activities is described in the appendices. However, our overarching finding is that the two constructs/formalisms have much in common and the underlying mechanisms are mutually supportive. Further work which examines how machine learning tools might unify the two as part a singular data systems management approach is likely to be productive.

ACKNOWLEDGEMENTS

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. doi: 10.1016/j.clsr.2018.02.001.

[2] Paul Ohm. Broken promises of privacy: Responding to the surprising failure of anonymization.

[3] Mark Elliot, Elaine Mackey, Kieron O'Hara, and Caroline Tudor. *The anonymisation decision-making framework*. UKAN publications, 2016.

[4] Taeho Jung, Seokki Lee, and Wenyi Tang. Using provenance to evaluate risk and benefit of data sharing. In *13th International Workshop on Theory and Practice of Provenance (TaPP 2021)*, 2021.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] Nils Ulltveit-Moe and Vladimir Oleshchuk. A novel policy-driven reversible anonymisation scheme for xml-based services. *Information Systems*, 48:164–178, 2015.

[10] John Rumbold and Barbara Pierscionek. Contextual anonymization for secondary use of big data in biomedical research: proposal for an anonymization matrix. *JMIR medical informatics*, 6(4):e7096, 2018.

[11] Kamyar Hasanzadeh, Anna Kajosaari, Dan Häggman, and Marketta Kyttä. A context sensitive approach to anonymizing public participation gis data: From development to the assessment of anonymization effects on data quality. *Computers, Environment and Urban Systems*, 83:101513, 2020.

[12] Mark Elliot, Elaine Mackey, and Kieron O'Hara. *The anonymisation decision-making framework 2nd Edition: European practitioners' guide*. UKAN publications, 2020.

[13] PROV Data Model. https://www.w3.org/TR/prov-dm/, 2013. accessed on 2020-04-20.

[14] 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.

[15] Barbara Blaustein, Adriane Chapman, Len Seligman, M. David Allen, and Arnon Rosenthal. Surrogate parenthood: Protected and informative graphs. *Proc. VLDB Endow.*, 4(8):518–525, may 2011.

[16] James Cheney and Roly Perera. An analytical survey of provenance sanitization. In *International Provenance and Annotation Workshop*, pages 113–126. Springer, 2014.

[17] Wai Kit Sze and R. Sekar. Provenance-based integrity protection for windows. In *Proceedings of the 31st Annual Computer Security Applications Conference*, ACSAC 2015, page 211–220, New York, NY, USA, 2015. Association for Computing Machinery.

[18] 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, 2018.

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

[20] 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.

[21] 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.

[22] Víctor Cuevas-Vicenttín, Saumen Dey, Sven Köhler, Sean Riddle, and Bertram Ludäscher. Scientific workflows and provenance: Introduction and research opportunities. *Datenbank-Spektrum*, 12(3):193–203, October 2012. doi: 10.1007/s13222-012-0100-z.

[23] Tope Omitola, André Freitas, Edward Curry, Séan O'Riain, Nicholas Gibbins, and Nigel Shadbolt. Capturing interactive data transformation operations using provenance workflows. In *Lecture Notes in Computer Science*, pages 29–42. Springer Berlin Heidelberg, 2015. doi: 10.1007/978-3-662-46641-4_3.

[24] Leonardo Murta, Vanessa Braganholo, Fernando Chirigati, David Koop, and Juliana Freire. noWorkflow: Capturing and analyzing provenance of scripts. In *Lecture Notes in Computer Science*, pages 71–83. Springer International Publishing, 2015. doi: 10.1007/978-3-319-16462-5_6.

[25] Timothy McPhillips, Tianhong Song, Tyler Kolisnik, Steve Aulenbach, Khalid Belhajjame, R. Kyle Bocinsky, Yang Cao, James Cheney, Fernando Chirigati, Saumen Dey, Juliana Freire, Christopher Jones, James Hanken, Keith W. Kintigh, Timothy A. Kohler, David Koop, James A. Macklin, Paolo Missier, Mark Schildhauer, Christopher Schwalm, Yaxing Wei, Mark Bieda, and Bertram Ludäscher. YesWorkflow: A user-oriented, language-independent tool for recovering workflow information from scripts. *International Journal of Digital Curation*, 10(1):298–313, May 2015. doi: 10.2218/ijdc.v10i1.370.

[26] Saumen Dey, Khalid Belhajjame, David Koop, Meghan Raul, and Bertram Ludäscher. Linking prospective and retrospective provenance in scripts. In *7th {USENIX} Workshop on the Theory and Practice of Provenance (TaPP 15)*, 2015.

[27] Qian Zhang, Yang Cao, Qiwen Wang, Duc Vu, Priyaa Thavasimani, Timothy McPhillips, Paolo Missier, Peter Slaughter, Christopher Jones, Mathew B. Jones, and Bertram Ludäscher. Revealing the detailed lineage of script outputs using hybrid provenance. *International Journal of Digital Curation*, 12(2):390–408, August 2018. doi: 10.2218/ijdc.v12i2.585.

[28] James Cheney and Roly Perera. An analytical survey of provenance sanitization. In *Lecture Notes in Computer Science*, pages 113–126. Springer International Publishing, 2015. doi: 10.1007/978-3-319-16462-5_9.

[29] Barbara Blaustein, Adriane Chapman, Len Seligman, M. David Allen, and Arnon Rosenthal. Surrogate parenthood. *Proceedings of the VLDB Endowment*, 4(8):518–525, May 2011. doi: 10.14778/2002974.2002979.

[30] Michael Stonebraker, Paul Brown, Alex Poliakov, and Suchi Raman. The architecture of SciDB. In *Lecture Notes in Computer Science*, pages 1–16. Springer Berlin Heidelberg, 2011. doi: 10.1007/978-3-642-22351-8_1.

[31] James Cheney. A formal framework for provenance security. In *2011 IEEE 24th Computer Security Foundations Symposium*. IEEE, June 2011. doi: 10.1109/csf.2011.26.

[32] Qun Ni, Shouhuai Xu, Elisa Bertino, Ravi Sandhu, and Weili Han. An access control language for a general provenance model. In *Lecture Notes in Computer Science*, pages 68–88. Springer Berlin Heidelberg, 2009. doi: 10.1007/978-3-642-04219-5_5.

[33] 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, Ltd, August 2012. doi: 10.1002/9781118348239.

[34] George T. Duncan, Mark Elliot, and Juan-José Salazar-González. *Statistical Confidentiality*. Springer New York, 2011. doi: 10.1007/978-1-4419-7802-8.

[35] Thomas Pasquier, Xueyuan Han, Mark Goldstein, Thomas Moyer, David Eyers, Margo Seltzer, and Jean Bacon. Practical whole-system provenance capture. In *Proceedings of the 2017 Symposium on Cloud Computing*. ACM, September 2017. doi: 10.1145/3127479.3129249.

[36] Chenyun Dai, Dan Lin, Murat Kantarcioglu, Elisa Bertino, Ebru Celikel, and Bhavani Thuraisingham. Query processing techniques for compliance with data confidence policies. In *Lecture Notes in Computer Science*, pages 49–67. Springer Berlin Heidelberg, 2009. doi: 10.1007/978-3-642-04219-5_4.

[37] Amani Abu Jabal, Maryam Davari, Elisa Bertino, Christian Makaya, Seraphin Calo, Dinesh Verma, and Christopher Williams. ProFact: A provenance-based analytics framework for access control policies. *IEEE Transactions on Services Computing*, pages 1–1, 2019. doi: 10.1109/tsc.2019.2900641.

[38] Harshvardhan Jitendra Pandit, Declan O'Sullivan, and Dave Lewis. Extracting provenance metadata from privacy policies. In *Lecture Notes in Computer Science*, pages 262–265. Springer International Publishing, 2018. doi: 10.1007/978-3-319-98379-0_32.

[39] Line Pouchard, Kevin Huck, Gyorgy Matyasfalvi, Dingwen Tao, Li Tang, Huub Van Dam, and Shinaje Yoo. Prescriptive provenance for streaming analysis of workflows at scale. In *2018 New York Scientific Data Summit (NYSDS)*. IEEE, August 2018. doi: 10.1109/nysds.2018.853895.

[40] Luc Moreau, Belfrit Victor Batlajery, Trung Dong Huynh, Danius Michaelides, and Heather Packer. A templating system to generate provenance. *IEEE Transactions on Software Engineering*, 44(2):103–121, February 2018. doi: 10.1109/tse.2017.2659745.

[41] 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.

[42] 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.

[43] 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.

[44] Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*, pages 94–103. IEEE, 2007.

[45] What is personal data? | ICO. accessed on 2021-10-25.

[46] Luk Arbuckle and Felix Ritchie. The five safes of risk-based anonymization. *IEEE Security & Privacy*, 17(5):84–89, 2019.

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

Appendices

A ACTIVITIES

The following is the set of activities supported by this project:

- Partnership between two Turing Partner Universities: Southampton and Manchester, drawing together both Computer Science and Social Science for an interdisciplinary approach to this real world problem.
- Alan Turing Institute Workshop **Provenance, security & machine learning**. Monday 11 Nov 2019 Time: 09:45 - 17:30. Organizers: Dr James Cheney, Dr Adriane Chapman, Dr Paolo Missier, Kate Wicks.
- Conversations with data holders and other stakeholders, in order to disseminate and critique early results in a critical-friendly environment, and to ensure the practicality and real-world focus of the work.
- A presentation delivered to the cross-sector UNECE work session on statistical data confidentiality - which generated significant interest. "Modelling data environments within PROV to assist anonymisation decision-making"
- Cross project meeting with members the ESOC-life EU framework project this led to the investigators joining the authorship team for a co-authored paper on provenance standards.

B OUTPUTS

The following is the set of outputs supported by this project:

- M.A. Jarwar, A. Chapman, M. Elliot and F. Raji, (2021) "Modelling data environments within PROV to assist anonymisation decision-making", Proceedings of UNECE/Eurostat Expert Meeting on Statistical Data Confidentiality, 2021. https://unece.org/statistics/events/SDC2021

- Set of compiled and released Use Cases that highlight requirements for Function Anonymisation with associated provenance.
- Jarwar, M. A., Chapman, A., Elliot, M., Raji, F. (2021). Provenance, anonymisation and data environments: a unifying construction. arXiv preprint arXiv:2107.09966.
- *In preparation* Muhammad Aslam Jarwar, Adriane Chapman, Mark Elliot, Fatemeh Raji, and Tom Blount. 2022. Modelling Data Environments Within PROV to Assist Decision making for Anonymisation. Target journal: Computing Law and Security review. Intended submission October 2022.
- *In preparation* Fatemeh Raji, Mark Elliot, Adriane Chapman, Muhammad Aslam Jarwar, Tom Blount. 2022. Retrospective, Prospective, Permitted, Prescriptive and Proscriptive (RP4) Provenance for Anonymisation Decision-Making. Target journal: Law, Innovation and Technology. Intended submission December 2022.
- *under review* Rudolf Wittner, Petr Holub, Cecilia Mascia, Francesca Frexia, Heimo Muller, Markus Plass, Clare Allocca, Fay Betsou, Tony Burde, Ibon Cancio, Adriane Chapman, Martin Chapman, Melanie Courtot, Vasa Curcin, Johann Eder, Mark Elliot, Katrina Exter, Elliot Fairweather, Carole Goble, Martin Golebiewski, Bron Kisler, Andreas Kremer, Sheng Lin-Gibson, Anna Marsano, Marco Maccavelli, Josh Moore, Hiroki Nakae, Isabelle Perseil, Ayat Salman, James Sluka, Caterina Strambio-De-Castillia, Michael Sussman, Jason R. Swedlow, Kurt, Zatloukal, and Jorg Geiger, (2022) Towards a Common Standard for Data and Specimen Provenance in Life Sciences. Submitted to Journal of Biomedical Informatics. Preprint available here: https://tinyurl.com/Provenance-standard
- Software/code/tools/methods developed/released:
  - PROVExtension (https://github.com/aslamjarwar/ProvExtension). Extends the PROV model and the supporting visualization code to allow for Environments to be represented in PROV.
  - A semi-automated tool to help the decision makers in understanding the complex relationships and constraints of multiple Data Governance Instruments (DGIs) and parameters existed in the ADF environments. The Reasoning Example can be found here: https://github.com/aslamjarwar/ReasoningExample and the code for the Simulator itself can be found here: https://github.com/aslamjarwar/Data-Environment-Simulation
  - RP4-NL: Code for processing DGIs to create RP4 provenance: https://github.com/DuncanSmith147/RP4
- A proposal drafted for future work, and assessed by EPSRC as within their remit.