International Journal of Population Data Science

Journal Website: www.ijpds.org

Semi-automated data provenance tracking for transparent data production and linkage to enhance auditing and quality assurance in Trusted Research Environments

Katherine O'Sullivan[1,*], Milan Markovic[2], Jaroslaw Dymiter[3], Bernhard Scheliga[3], Chinasa Odo[4], and Katie Wilde[3]

Submission History

Submitted:	02/06/2024
Accepted:	14/11/2024
Published:	06/02/2025

[1]University of Sheffield, Research and Innovation IT, 10-12 Brunswick St, Sheffield, S10 2FN
[2]University of Aberdeen, Department of Natural and Computing Sciences, 32 Elphinstone Rd, Aberdeen AB24 3EU
[4]University of Aberdeen, Grampian Data Safe Haven, Health Sciences Building, Foresterhill, Aberdeen, AB25 2ZD
[3]University of Bradford, Faculty of Health Studies, Bradford, BD7 1DP

Abstract

Introduction

We present a prototype solution for improving transparency and quality assurance of the data linkage process through data provenance tracking designed to assist Data Analysts, researchers and information governance teams in authenticating and auditing data workflows within a Trusted Research Environment (TRE).

Methods

Using a participatory design process with Data Analysts, researchers and information governance teams, we undertook a contextual inquiry, user requirements interviews, co-design workshops, low-fidelity prototype evaluations. Public Engagement and Involvement activities underpinned the methods to ensure the project and approaches met the public's trust for semi-automating data processing. These helped inform methods for technical implementation, applying the PROV-O ontology to create a derived ontology following the four-step Linked Open Terms methodology and development of automated scripts to collect provenance information for the data processing workflow.

Results

The resulting Provenance Explorer for Trusted Research Environments (PE-TRE) interactive tool displays the data linkage information extracted from a knowledge graph described using the derived SHP ontology and results of rule-based validation checks. User evaluations confirmed PE-TRE would contribute to better quality data linkage and reduce data processing errors.

Conclusion

This project demonstrates the next stage in advancing transparency and quality assurance within TREs by semi-automating and systematising data tracking in a single tool throughout the data processing lifecycle, improving transparency, openness and quality assurance.

Keywords

data provenance; semi-automation; trusted research environments; secure data environments; safe haven; data linkage; transparency; information governance; audit

*Corresponding Author:
Email Address: k.k.osullivan@sheffield.ac.uk (Katherine O'Sullivan)

https://doi.org/10.23889/ijpds.v10i2.2464

Introduction

Large-scale analysis of health and social care data has the potential to transform health and population outcomes, yet to achieve these outcomes researchers must rely on complex project-specific data extractions and linkage across multiple datasets (either via unique identifier(s) or probability matching) across multiple data sources both within and across organisations. A major challenge for the specialist staff providing data, particularly providers of sensitive, personal data, is ensuring that the extraction and linkage of this complex data to create bespoke, research-ready datasets is performed in line with the research project's ethical permissions but without direct involvement from researchers since their ethical permissions do not permit them to see patient-identifiable information.

For example, a research team applies to undertake research in the public good to identify patients that have particularly characteristics. They may as for patients of a certain age range (e.g.18-50) with a specific disease (e.g. Chronic obstructive pulmonary disease (COPD)) who present at Emergency department attendance, inpatient admissions, and who receive certain medicines prescribed via their GP. They may also ask for any deaths data and deprivation data (based on postcode) to understand outcomes of particular drug interventions. Within a Trusted Research Environment (TRE), also known as a Secure Data Environment (SDE), specialist staff (Data Analysts) must translate researchers' data requirements into programming code to create the cohort, and then find all records of these individuals meeting the study's data requirements across the datasets and associated variables. In many TREs or SDEs, Data Analysts remove personal identifiers of individuals (name, address, etc.) as well as minimise certain critical pieces of information about individuals so as to provide useful information to researchers whilst also ensuring patient confidentiality and privacy (e.g. replace date of birth with age, replace postcode with deprivation quintile or decile and replace patient ID with a study-specific ID). Finally, the Data Analysts need to link all datasets together so that researchers can analyse individuals across each of the distinct datasets. Researchers must rely on Data Analysts to interpret their cohort specifications, prepare bespoke extractions (which may be across dozens of datasets, with hundreds of variables per dataset), minimise all potentially identifiable variables and correctly link individuals across the datasets. Researchers receive this prepared data ready for analysis, but do not have the ability to validate the data due to the identifiable nature of the data until it is fully de-identified and prepared. For Data Analysts, each step in this extraction and linkage process requires interpretation and decision-making, and whilst it may be documented within the stored procedures of the programming queries, these are not able to be shared with the researchers as the code itself may also contain patient identifiable information (e.g. specific patient identifier exclusions, etc.).

At the same time that researchers must ensure their analysis is open, transparent, and reproducible [1], data extraction and linkage to produce research-ready datasets within TREs is restricted and opaque, and lacking clarity around processing decisions made during the data production workflow [2]. Indeed, it has been noted that errors related to data processing before, during and after linkage can bias results [3]. Errors introduced during data processing can be benign (e.g. minor variations without statistical significance) or can be catastrophic (e.g. incorrectly including or excluding individuals from the cohort), where the entire analysis could be flawed, leading to policies, regulations or practices implemented that could harm rather than help individuals and/or populations.

The GUILD guidance [1] aimed to address errors in data linkage and analysis by providing data processors a series of principles that would provide data linkers, analysts and researchers to help assess linkage errors. For Data Analysts extracting and linking data there are a number of key recommendations: data providers should share information to explain how the data set was created and maintained, including how data was collected, cleaned and standardised; when linking data, descriptions and justifications of the linkage characteristics should be provided (e.g. which fields have been used to link across datasets); information such as methods and algorithms for linkage, and accuracy at the aggregate level (such as comparison of aggregate counts, uniqueness, etc.) should be shared; and any methods for disclosure control (e.g. removing patients with protected characteristics or sensitive disease categories) applied to the linked data should be published before providing to researchers. These principles aim to reduce the risk of errors being introduced in data processing prior, whilst also providing assurance to researchers about the quality and provenance of the data made available to them. However, an unknown is to what extent data processors are following the GUILD guidance, if at all, and how this information is being tracked and recorded through the data production workflow in TREs. Indeed, one of the challenges is systematising data tracking from ingress to egress in a single tool for multiple purposes: capturing metadata, inspecting the data as it is processed and having an auditable record for information governance or audit purposes. If GUILD guidance is being implemented by TREs, it is likely limited to metadata creation (e.g. datasets, variables and row counts) in a static form at the final stage of the data production workflow, since tracking all activities, decisions and people involved throughout the data production is hugely resource intensive across multiple systems, since it is likely to be done manually. To overcome this gap, we aim to provide a user-driven approach to designing and implementing a (semi-)automated provenance-based solution that fulfils the GUILD guidance for tracking, documenting and reporting a full data production lifecycle from data ingress to release of project-specific datasets to researchers for analysis. Moreover, we propose an end-to-end prototype software solution that offers Data Analysts the ability to inspect data records during data processing to improve data quality and reliability.

In this article, we describe the creation and implementation of a prototype data provenance tracking tool, or Provenance Explorer for Trusted Research Environments (hereafter, PE-TRE), in the Grampian Data Safe Haven (DaSH) in northeast Scotland, designed to improve transparency and trustworthiness of data processing of research data within a secure data setting. The prototype development involved five aspects: (1) co-design of user interfaces with end users for provenance data collection and visualisation, resulting in a low-fidelity design; (2) Public Information and Engagement

(PIE) activities allowing members of the public to evaluate the low-fidelity design to ensure it met public trust; (3) extension and refinement of the Safe Haven Provenance (SHP) ontology [4, 5]; (4) design and implementation of mechanisms for semi-automated collection of data linkage provenance described using the SHP ontology; (5) implementation and user evaluation of the provenance dashboard. PE-TRE is offered as an open-source solution for organisations to customise to their unique environments and data production workflows, whilst providing a systematic framework for automating data provenance tracking across TREs, in order to improve transparency and trust in sensitive data processing.

Background

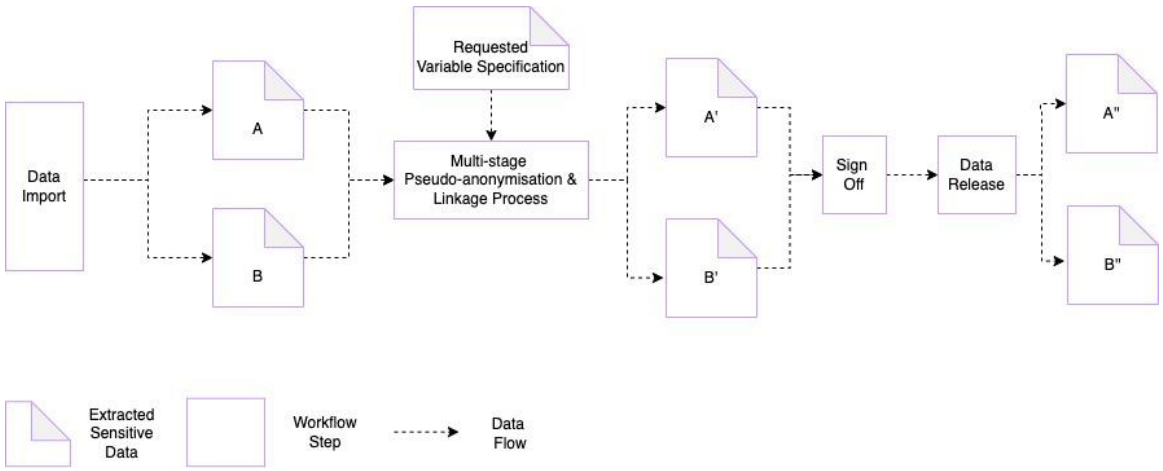
In Scotland, routinely collected but sensitive health and social care data is made available for research undertaken for public benefit [6, 7]. In ethics applications to access this sensitive data, researchers must identify a cohort of individuals, select the specific datasets and variables they require (hereafter, researcher specification, or researcher specification file) to answer their research questions, applying the principle that data requested should be the minimum necessary information required [8]. If a project is approved, data is then extracted, pseudonymised and linked by trained Data Analysts working in Scottish Government accredited 'Safe Havens' to ensure this sensitive, personal data is protected and treated with the highest levels of confidentiality and security [9, 10]. Researchers then analyse the pseudonymised data within these TREs. Finally, researchers request analysis release from the TRE to publish their research, of which a critical component is a review of the research outputs by specialist TRE staff to verify that the analysis has aggregated results to ensure no small numbers (<5) are released, thereby safeguarding individuals' privacy [11].

However, health data providers face challenges in extracting and linking this complex data and safeguarding its safe release for research. They must operate within strict information governance regulations to provide data as specified by researchers, but they are not experts in how the data was collected in the clinical setting or the systems used to capture this data. Moreover, since the data itself must be pseudonymised before releasing to the researchers to protect patient confidentiality, data providers cannot share the patient-level data during extraction or linkage with researchers to check for correctness. In addition to information governance and data familiarity challenges, processing data from ingestion to release for researchers to analyse can occur across multiple environments and is often manual, ad hoc and time consuming, with data checks and data tracking across multiple systems and captured and stored in separate locations without a uniform audit trail. These challenges pose a number of significant data quality assurance risks. First, Data Analysts may process data incorrectly when working from researcher-provided documentation (e.g. a research protocol or dataset selection) since they simply interpret the researchers' request differently than the researcher intended. However, this discrepancy could ultimately bias results since cohorts could be created erroneously. Second, they may unintentionally release patient-identifiable data when undertaking manual, time-consuming data checks since the data sizes can be significant (hundreds of millions of rows, with thousands of variables) and may simply miss a potentially identifiable variable, particularly if the data is messy at point of entry. Third, without a uniform audit trail that accurately tracks data flow across infrastructures and/or systems, corruption errors could be introduced in data processing, linkage or transfers. Finally, without a uniform audit trail, evidencing the full lifecycle of data processing for information governance or internal/external audits can become a laborious task. It requires the collection and compilation of any relevant information required to demonstrate the compliance of data processing according to applicable agreements and standards for both individual projects and wider data processing requirements set by data controllers and organisations. When that information sits in multiple systems and in various formats, it requires a significant amount of manual effort to identify, collate and synthesise this information and relevant information may be missed. Figure 1 shows a typical data production workflow of two datasets from import to release to researcher.

The patient-identifiable databases are A and B; the data is then extracted according to the cohort description, variable selection, and undergoes a multi-stage de-identification and minimisation and linkage, resulting in Dataset A' and B'. These then undergo a quality assurance signoff and are released to the researchers as files A'' and B''. Errors can be introduced at each stage of this process, and due to the number of steps involved, tracking the full data provenance is challenging given the substantial number of processing activities.

To design, build, test and evaluate the effectiveness of (semi-)automating data provenance of sensitive data processing within a TRE, we worked with the Grampian Data Safe Haven (DaSH), one of the four regional Scottish-government accredited Safe Havens in North East Scotland. DaSH holds longitudinal health, social care and population data for the Aberdeenshire, Aberdeen City and Moray population with some datasets going back as far as 1950s [12], and it serves as a case study for introducing data provenance into TREs. DaSH's trusted research environment architecture operates across two organisational environments. Patient-identifiable data is extracted from databases on the NHS Grampian Health Board infrastructure. Once a research study's cohort is identified and accompanying datasets extracted by DaSH analysts, this data is pseudonymised within the NHS network and then checked twice for integrity and to ensure all patient identifiable data has either been removed or appropriately pseudonymised. The data then undergoes a final signoff for files to be transferred by SFTP to the University of Aberdeen network, where the files receive a second round of pseudonymisation on the DaSH analyst platform. The twice-pseudonymised files are then linked so that researchers can undertake analysis across the relevant project data from multiple datasets. These files are checked and then signed off for release to researchers within the TRE for analysis. Currently in the UK, most TREs only exist in a single environment (e.g. the second stage processing within the DaSH analyst platform described above). For that reason, the development of the prototype provenance trace and accompanying interactive dashboard, focussed on implementation on the University of

Figure 1: Data production workflow in a TRE



Aberdeen platform only rather than across two platforms. Implementing in a single platform most closely resembles the existing architecture of many TREs in the UK, with one area for data analyst processing and a second area for researchers to securely access and analyse the data. Again, like most TREs, DaSH does not currently implement any approach that would record a unified provenance trace of the data linkage process and hence only partial and fragmented information is recorded in a digital format (e.g. in multiple manually created spreadsheets).

The benefit of a single provenance trace in TREs would allow the full history of the data from its origin through its various transformations in the linkage workflow to improve quality and promote transparency. Likewise, categorising the capture of this information in a standardised, formal structure means that the information is captured within a standard provenance model. The W3C PROV standard defines data provenance as 'information about *entities*, *activities*, and *people* involved in producing a piece of data or thing, which can be used to form assessments about its quality, reliability or trustworthiness' [13]. The PROV standard provides a common data model (CDM) for recording provenance information in a way that is understandable to machines by formalising a vocabulary of concepts such as *entities*, *activities and agents* and relationships between them (e.g. linking agents to activities for which they were responsible). Such provenance traces can then be used for producing automated audit reports and error detection (data discrepancies) in data extraction and linkage. A further benefit of employing a CDM is that it promotes interoperability and exchange of information across organisations, allowing the exchange of information across heterogeneous environments.

Our initial research established a first version of the Safe Haven Provenance (SHP) ontology by extending the PROV standard for a domain specific application. Furthermore, two early prototype provenance reporting templates were developed, one for Data Analysts within the TRE to assist with data checking during extraction and linkage and the other for researchers that would help assist researchers in detecting errors during the processing as it did not display potentially identifiable information [3]. In the research presented in this paper, we aimed to address two outstanding challenges to operationalise the prototype provenance model in a real-world environment: (1) co-design user interfaces for provenance data collection and visualisation; (2) description and formalisation of a TRE ontology based on the requirements of the real-world application deployment trial; and (3) design and implement mechanisms for collecting provenance information about the individual activities. The outcome of this research provides an open-source solution for TREs to design and implement data provenance capture and reporting to improve transparency, accuracy, and quality assurance in secure data settings.

Related work

The use of W3C PROV has been previously demonstrated as means to increase reproducibility and trust of computer-generated outputs, for example, by Curcin and colleagues in the context of diagnostic clinical decision support systems [14]. The authors employ the use of provenance templates which together with domain-specific ontologies represent domain abstractions of meaningful provenance information that can be mapped to the actions of the software. Such an approach stipulates the provenance tracking functionality to be embedded within the decision support software and provenance is stored in a graph database on a server to provide question answering capabilities over the provenance graphs. In our approach we follow a similar process, however, this is adapted for an offline environment where automated scripts unobtrusively extract provenance information from a file system, describe it using domain-specific ontology and store it in a local file. Furthermore, we exploit the semantic rules to help users to automatically check compliance of the data linkage process by comparing and retrieving important parts of the provenance trace.

Xu and colleagues [15] highlights the importance of challenges related to data visualisations of fine-grained provenance graphs which might be difficult to navigate and overwhelming to the user. They specifically note the issue of data literacy of the user and security (e.g. sensitive data can be hidden or obfuscated). For this reason, we employ the user-centric approach where the provenance visualisation interfaces were co-designed with the end users. We also abstracted the

provenance information (e.g. only aggregated statistics about a dataset are recorded) to reduce the amount of potentially sensitive information captured in a provenance record.

Johnson and colleagues [16] demonstrated that system flow mapping of electronic health records captured in clinical settings required interpretation and transformation before it could be used for research, but that data provenance was crucial especially for researchers who were not familiar with the clinical workflow and might miss or misunderstand data. Black and colleagues [17] showed that data provenance was crucial in identifying changes in electronic medical records software that introduced data quality issues and data provenance could facilitate remediation of these issues. Can and Yilmazer [18] used an ontology-based model to validate and verify the access to personal data related to history and timings of vaccines. Importantly, they demonstrated that this auditing of tracing versions of data, changes made, people who accessed and permissions for access improved patient outcomes when administering vaccines. Finally, Sun and colleagues [19] focused on the origins of data using the open provenance model to create a fully automated provenance prototype to undertake data quality assessments for healthcare management. Their focus, however, was on 'dirty data' (e.g. real data in hospital systems that may have inaccuracies, incompleteness, etc., and has not been cleansed or standardised) and how trustworthy the source data was when combining different healthcare records.

To summarise, while the role of provenance has been extensively discussed in a variety of medical and research data contexts and its benefits are widely agreed, it seems that development of practical provenance applications in this context remains a challenge.

The remainder of this article is structured as follows: We present the context and infrastructure in which the prototype PE-TRE was developed and deployed. We review existing approaches to implementing a provenance-based solution for data tracking. Next, we explain the methods used to develop and validate a low-fidelity prototype. After, we describe the approach to implementing a prototype solution within a real-world environment and its evaluation by the end users. Finally, we discuss outcomes (including limitations) and opportunities for future work.

Methods

There were two main obstacles to overcome to operationalise data provenance tracking within a real-world environment: (1) Co-design user interfaces for provenance data collection and visualisation with end users (e.g. Data Analysts, researchers, information governance and auditors) to maximise the utility of provenance information in the context of error prevention and trustworthiness of produced data assets; and (2) Design and implement mechanisms for collecting provenance information about the individual activities (e.g. data extraction, anonymisation) and their inputs (e.g. characteristics of the produced datasets) within the data linkage workflow in an unobtrusive, reliable, and, critically, a semi-automated manner to improve consistency and timeliness.

In this project, we adopted a participatory design approach to promote democratic collaboration and improved innovation with major stakeholders during the system's development cycle [20, 21]. The primary goal was to create a user-centric system that aligns with the needs and expectations of its intended users by actively involving them in the design process [22]. This approach empowers users to have a say in shaping the system they will ultimately use, ensuring it better meets their requirements and preferences. In this project, we undertook a three-step participatory design process including contextual inquiries, user requirements interviews, and co-design workshops, complimented by a low-fidelity prototype evaluation to ensure the data collected met the user's needs. Finally, we conducted a usability evaluation of the working prototype to confirm that the system was user friendly. Participant numbers throughout the methods and findings are small due to the specialist nature of the work and size of the teams involved in DaSH. The participatory design and low-fidelity prototype evaluation are described below.

Stage 1. Contextual inquiry

The first step to designing an effective data provenance tracking tool was to undertake a contextual inquiry [23] with Safe Haven Data Analysts (*invited n=6; attended n=3*) within the Grampian Data Safe Haven to gain deeper insights into the real-world work environment and their interactions as data is ingested, extracted, pseudonymised, linked and checked prior to releasing data to researchers in the TRE. This method involves observing users in their natural environment to understand users' activities and processes and is essential in ensuring technology can accommodate and enable their real work [24]. Participants were asked to describe their typical process for data extraction and linkage, as well as the challenges they encounter when carrying out these tasks.

Stage 2: User requirements interviews

User requirement interviews were conducted to enable end users to articulate their expectations and needs from the proposed provenance dashboard. These interviews were instrumental in understanding the desired functionality and performance of the tool by the proposed user groups: Data Analysts (*invited n=4; attended n=2*), Researchers (*invited n=4; attended n=2*) and Information Governance specialists (*invited n=3; attended n=2*). Moreover, they served as the foundation for subsequent phases of specification, design, and the creation of user personas, which were incorporated into collaborative design workshops.

Stage 3: Co-design workshop

We undertook a co-design workshop to co-create the low-fidelity prototype of the PE-TRE with the prospective users who participated in the contextual inquiries and user requirements interviews, applying the 'user-as-wizard' methodology [25] to explore design challenges and solutions. This co-design session was run as a collaborative workshop that brought together the various stakeholders to generate ideas, solve problems, and create solutions collectively to the requirements previously identified and how best to implement them in a tool including visually and functionally.

Stage 4: Low-fidelity prototype evaluation

Results of the contextual inquiries, interviews and co-design workshop were compiled, and researchers developed low-fidelity designs based on the synthesis of the collaborative ideas generated by the participating stakeholders. A final validation session was conducted with users who participated in the earlier co-design process to confirm that the proposed design accurately represented the ideas generated during the workshop. This session provided users with an opportunity to review the low-fidelity prototype of the PE-TRE and identify any potential changes or adjustments that might be needed. It ensured that the system aligns closely with user expectations and requirements, enhancing its effectiveness and user satisfaction.

Stage 5: Public involvement and engagement

As we noted in a separate report about our design and implementation of public involvement and engagement, trustworthiness of data-access processes by listening and responding to public inputs is crucial to maintaining a social licence to process confidential patient data for research purposes [26]. Therefore, involving the public in evaluating the PE-TRE was essential for enabling the research team to understand the areas of data provenance that are of most concern to (or are less understood by) members of the public and to obtain public trust in the processing of data for research undertaken in the public good.

Stage 6: Technical implementation

Results of Stages 1-5 were used to derive technical system requirements for a prototype provenance infrastructure used to evaluate the feasibility of our approach. The system comprised three main components, namely the Safe Haven Provenance Ontology (SHP), semi-automated provenance collection components, and a visual provenance explorer tool for trusted research environments (PE-TRE). Following the identified system requirements we have revised the previous version of the SHP and created an updated version with additional classes and properties following the four-step Linked Open Terms methodology [27]. The ontology was used to manually produce examples of provenance traces that then informed the development of the PE-TRE tool. Finally, automated scripts and web-based user interfaces were developed to automate the collection of provenance information to support the test deployment of our system in the DaSH environment.

Results

Stage 1. Contextual inquiry

The information gathered during the contextual inquiries played a pivotal role in shaping the production of materials used throughout the participatory design process, such as the creation of user personas, an initial key requirement list, a content strategy for the dashboard and an approach to the prioritisation of design features. A number of key requirements for the prototype PE-TRE were identified at this stage, including:

- A mechanism to check that imported data has not been corrupted;
- Compile a list of sensitive data variables and ensure they can be tracked in the data file to check for pseudonymisation or minimisation, as data progresses through the workflow;
- Highlight any duplicate values; and
- For data to be released to researchers for analysis within the TRE, implement a check that all the sensitive data are excluded from the release table.

Stage 2: User requirements interviews

Interviews resulted in a number of features requested for the PE-TRE by user groups with several overlapping requirements. We then reviewed the requirements and categorised them as 'in scope' (e.g. relevant to the provenance trace capture and visualisation) and 'out of scope' (e.g. either outside the project scope or for future enhancements once the prototype was operational). The in-scope requirements include:

- Dataset Selection Justification: Researchers are provided with a list of available patient records datasets. They choose from this list and provide justifications for the variables (dataset fields) required within each dataset, aligned with the project protocol and to accompany the ethics application for the research study. This should also capture where linkages should be established between them. (*Researchers, Analysts, Information Governance*)
- System to automatically provide a summary of what has been extracted using the inclusion and exclusion criteria selected by researchers. (*Researchers, Analysts*)
- Provide the range of the data extracted to show transparency. (*Researchers, Analysts Information Governance*
- Provide automated checks for basic data processing requirements (e.g. date ranges, min/max values, number of rows, unique identification numbers) to remove time-consuming manual checks. (*Analysts*)

Stage 3: Co-design workshop

The co-design workshop offered the participants the opportunity to sketch out options for the visual design of the visual provenance explorer and accommodate multiple user views about visualisation preferences of the key pieces of information that would be helpful to interrogate in the data production workflow.

Stage 4: Low-fidelity prototype evaluation

Results of the contextual inquiries, interviews and co-design workshop were compiled, and researchers developed low-fidelity designs based on the synthesis of the collaborative ideas generated by the participating stakeholders (Figure 2).

A final validation session was conducted with users who participated in the earlier co-design process to confirm that the proposed design accurately represented the ideas generated during the workshop. This session provided users with an opportunity to review the low-fidelity prototype of the PE-TRE and identify any potential changes or adjustments that might be needed. It ensured that the system aligns closely with user expectations and requirements, enhancing its effectiveness and user satisfaction.

The feedback provided by the users included four main suggestions on improving the low-fidelity design for implementation in the PE-TRE:

- Provide a unified log of information if any changes occur during data processing; this will keep stakeholders informed about changes or any modifications throughout the processing lifecycle.
- Include a feature that displays a range of variables, highlighting if any variables are missing from what researchers have requested. For instance, if a researcher has requested data for a range of years (e.g. 2000–2010), the system should indicate if any specific year, such as 2003, is missing from that range. This can be a valuable tool for Data Analysts and Researchers to quickly identify any gaps or discrepancies in the data they have requested, ensuring they are working with the complete set of variables they need for their research.
- Show which Data Analyst is working on which project and during each stage of the data processing.
- Include information about any changes in the cohort during the data extraction process by providing an explanation of what that change entails.

These were prioritised as key requirements in the implementation of the data provenance dashboard since they were captured as part of the evaluation of the prototype.

Stage 5: Public involvement and engagement

A detailed report on the public's views was produced by our project collaborators Ipsos Scotland [28]. However, the main findings from workshops with the public to meet their threshold for transparency in processes were twofold: first, recording a decision log as part of the dashboards, including decisions made about data included or excluded (and why); and second, documenting a quality standard or statement about quality assurance procedures to show staff had followed a consistent procedure.

The public views around maintaining data privacy and confidentiality also specified several core requirements. First, a dashboard or displays for researchers should have a mechanism to flag 'small numbers' (e.g. <5 or <10 patients) to minimise the risk of patient identifiability. Second, they suggested replacing any text that a researcher might see in the dashboard (such as drug names or a specific medical condition) with less specific descriptions (or code the descriptions to prevent identification via the specific description). Third, they suggested streamlining the dashboard so that they only contain the necessary information for those involved in the research to do their jobs effectively. Finally, they specified that there should be clear communication between the researchers and Data Analysts to help make decisions that balance patient privacy with relevance to the research.

Participants were receptive towards semi-automation and its potential use for both record-keeping and managing free-text patient data, but highlighted some conditions they felt should be in place for its use to be considered acceptable:

- The automation would need to take account of differences in the data (for example, different languages) and inconsistencies (e.g. use of abbreviations).
- There should be regular random spot checks by TRE staff to ensure the semi-automation is working properly over time.
- There should be ongoing maintenance of the algorithms (code) underpinning the semi-automation to ensure that it does not repeat mistakes or embed biases.

Fundamentally, the public acknowledged the benefits of scale and speed in the use of semi-automation in creating a provenance trace and tracking system as long as there is human involvement to interpret the nuances and make decisions around potential anomalies and errors.

We weighted all of the feedback from the public as 'essential' in the development and operationalisation of the provenance dashboard so that the tool met the public's requirements for enhancing openness and transparency and improving data protection and data privacy.

Stage 6. Technical Implementation

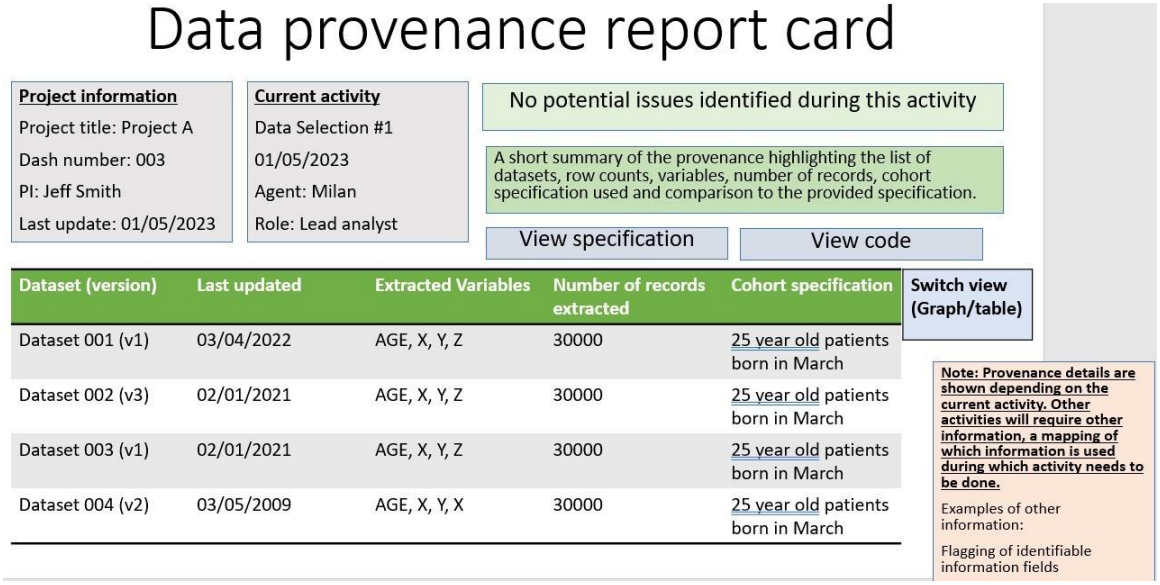
The Safe Haven Provenance Ontology (SHP)

In this section we describe our approach for modelling the provenance information of the data linkage process in the form of provenance knowledge graphs [29] using the SHP ontology [5]. Such knowledge graphs then provide a 'smart' data layer underpinning the audit mechanisms required by the PE-TRE. A knowledge graph represents information in a graph-based format, where entities (i.e. individuals, places, data items) are represented as nodes, and the relationships between these entities are represented as edges connecting the nodes. This structure enables the graph to capture how different pieces of information are interconnected, providing a comprehensive framework for understanding relationships and dependencies among various data points. By structuring data in this format, knowledge graphs facilitate more intuitive and powerful ways to navigate and analyse complex datasets, uncovering insights that might not be readily apparent in more traditional data storage and representation models. Moreover, knowledge graphs support logical rules and queries that can be used to create intelligent application layers for validating and reporting data to the end users.

We utilise the W3C PROV-O ontology to provide a high-level base vocabulary for describing causal provenance traces, that is, where actions and outputs are linked to some preceding actions and outputs that had some influence on them.

In the TRE context, each data linkage process (i.e. data extraction request handled by TRE Data Analysts) is described as a provenance trace consisting of activities (e.g. data extraction, data validation, etc.), entities (e.g. existing

Figure 2: Data provenance tool low-fidelity design



data and resources used, and new data produced by the activities), and agents (e.g. humans responsible for activities taking place and existence of individual entities). To enable queries about specific details of the data linkage provenance trace (e.g. specific activity types and their results) the generic provenance vocabulary defined by PROV-O was extended with domain-specific vocabulary formalised in the SHP ontology.

Figure 3a) illustrates three core components (i.e. Activity, Entity, Agent) of PROV-O that are used to describe causal provenance graphs of activities, their inputs and outputs, as well as descriptions of human or software agents bearing some responsibilities for the activities taking place. Figure 3b) illustrates a portion of the SHP ontology that extends PROV-O with new sub-classes of those three main concepts. In this example, a new sub-class of Activity, namely *Data Extraction*, is defined to describe the process of extracting data related to the cohort from databases according to the specific variables (dataset fields) selected by researchers. Another sub-class *Dataset* is defined to describe the extracted cohort-based subset of the database(s). Finally, the Safe Haven *Analyst* describes the person responsible for the validation checks during extraction, pseudonymisation and/or linkage.

Table 1 lists the concepts defined by the SHP ontology. The sub-classes of *shp:TREActivity* describe a number of common activities occurring during the data linkage process. In terms of agents, the SHP ontology currently defines sub-classes of *prov:Person* class that correspond to the roles of the Data Analysts involved in the data linkage process. For future use cases, it would be possible to also describe organisation and software agents (e.g. multiple organisations working on a federated data linkage) by extending corresponding PROV-O classes *prov:Software* and *prov:Organisation* (not shown in the table). The entities that are used and generated by different activities represent: the source of the extracted data (*shp:Database*); the extracted subset of those data sources (*shp:Dataset*) and their characteristics such as statistical summaries (*shp:EntityCharacteristic*); variables defining the individual items in the extracted data (*shp:Variable*) and their associated constraints (*shp:VariableConstraint*); reports summarising results of data checks (*shp:TREReport*); and various other supporting documents used by the Analysts during the data linkage process (*shp:TREProjectDocument*).

Provenance Trace Examples Described Using the SHP Ontology

In this section, we discuss how the SHP Ontology can be used to capture provenance information about specific TRE activities.

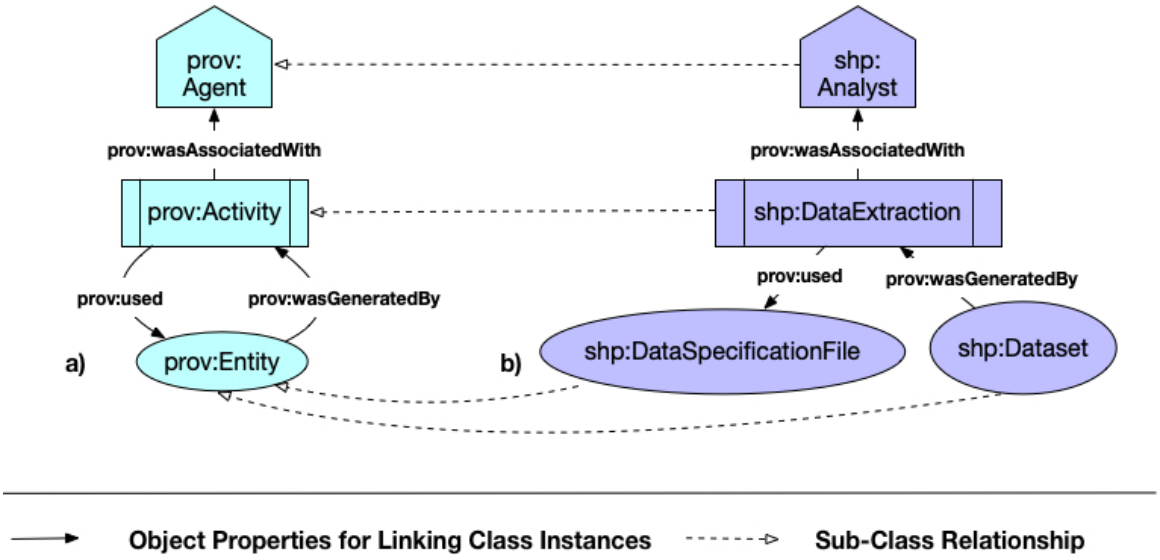
To expand on the example shown in Figure3b), in Figure 4 (below) we illustrate the description of a specific data extraction activity and its inputs and outputs. The graph illustrates instances of SHP classes forming a provenance trace where each instance is identified via unique International Resource Identifier (IRI) (e.g. ex:P1/DSF), and the class the instance belongs to is shown in parentheses.

For example, the entity instance describing the *shp:Data SpecificationFile* used during the data extraction is identified with IRI ex:P1/DSF where "ex:" is a prefix abbreviating a common IRI base. The figure captures the part of provenance trace informing about Alice (*shp:Analyst*) performing a data extraction activity (*shp:DataExtraction*) from the SMR01 database (*shp:Database*) using the variables specified by researchers (*shp:DataSpecificationFile)* and generating a dataset file ex:P1/imp/SMR01_Rel_v1.csv (*shp:Dataset*) and ID mapping file ex:P1/imp/Link_Rel_v1.csv (*shp:Dataset*).

Figure 5 (below) illustrates how to model aggregated statistics and other high-level information about a dataset without including the raw data into a provenance record. This approach, rather than working with raw data, increases the security of the provenance records since it will not include any potentially sensitive information about individual dataset entries.

The *shp:EntityCharacteristic* concept describes a specific aspect of an entity, in this case the SMR01_Rel_v1.csv dataset which was extracted from the SMR01 database (see Figure 5). The raw values in the dataset are grouped in the table where each column header represents a variable.

Figure 3: Part a) of the figure illustrates a simplified overview of the PROV-O ontology consisting of three main components prov:Agent, prov:Activity, and prov:Entity. Part b) illustrates an example of sub-classes of the corresponding PROV-O components defined in the SHP ontology used to describe a data extraction process that uses data source specification file to produce a dataset and is performed by a Safe Haven data analyst



The group of variables contained in the dataset is described using the *shp:SelectedVariables*, which is a sub-class of *prov:Collection*. Each of the variables is then modelled as a separate member of this collection; for example, *ex:SMR01/var/DISCHARGE_DATE* represents the date in the SMR01 database informing when a patient has been discharged from a hospital. The IRIs of the variables should remain consistent across the different provenance traces. This way we can later query, for example, for the most common variables that are being extracted within the TRE across different projects. The *shp:EntityCharacteristic* further captures additional details about the raw data associated with a specific variable in a dataset. For example, Figure 6 illustrates how data properties *shp:notNull*, *shp:minValue* and *shp:maxValue* describe a number of entries without null values, oldest discharge date and the latest discharge date found in the *ex:P1/imp/Link_Rel_v1.csv* dataset.Similarly, the provenance trace may include a description of the variables requested by the researchers, which guide the data extraction process. This information is captured in a *Data Specification File*. Each file contains one or more *shp:DataSourceSpecifications* that specify the source database

Table 1: An overview of the SHP ontology concepts and their relationship to PROV-O. Different levels imply sub-class relationships between concepts (left to right)

Level 1 class/type	Level 2 class/type	Level 3 class/type
prov:Agent	prov:Person	Analyst; Lead Analyst; Research Coordinator; Researcher; Technical Lead
prov:Entity	prov:Collection	Requested Variables, Selected Variables
	TRE Report	Disclosure Check Report; Validation Check Report
	TRE Project Document	Cohort; Data Linkage Plan; Data Specification File; Dataset Linkage Id Mapping; Statistical Summary; UID Mapping
	Data Source Specification	n/a
	Database	n/a
	Dataset	n/a
	Variable	Sensitive Variable
prov:Activity	TRE Activity	Cohort Creation; Data Check; Data Extraction; Data Transfer; Dataset Linkage; Dataset Release; Flat File Creation; Id Linkage; Original Id Removal; Project Id Assignment; Pseudonymisation; Study Id Assignment; Update
owl:Thing	Entity Characteristic; Variable Constraint	n/a

For example, all Data Extraction activities are a type of TRE Activity, and all TRE Activities are a type of PROV-O Activity but not all TRE Activities are Data Extraction activities and not all prov:Activities are PROV-O Activities. Concepts that are only subclasses of owl:Thing are not linked to any external ontologies.

Figure 4: An example provenance graph describing Data Extraction activity, activity inputs, activity outputs, and the person performing the activity

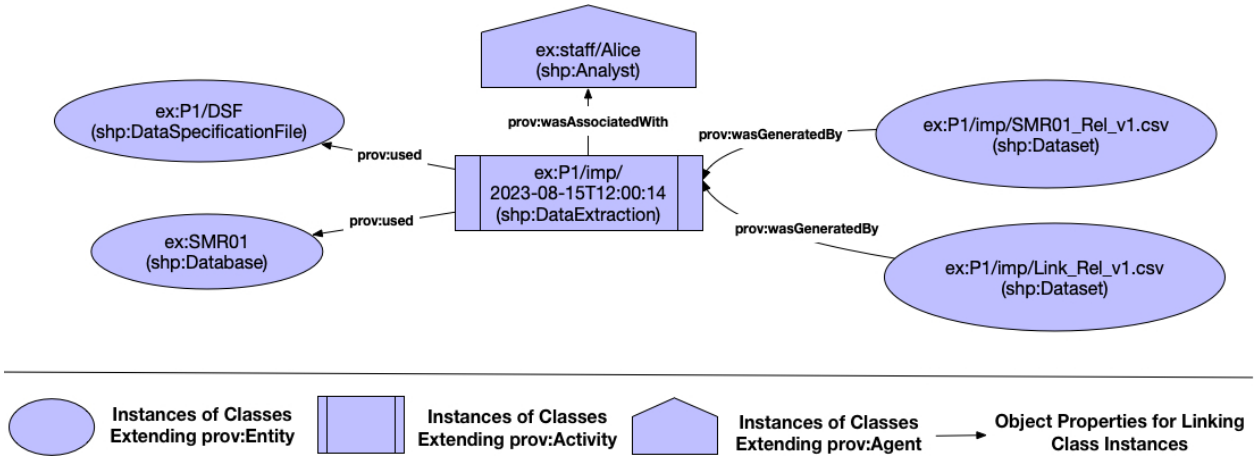
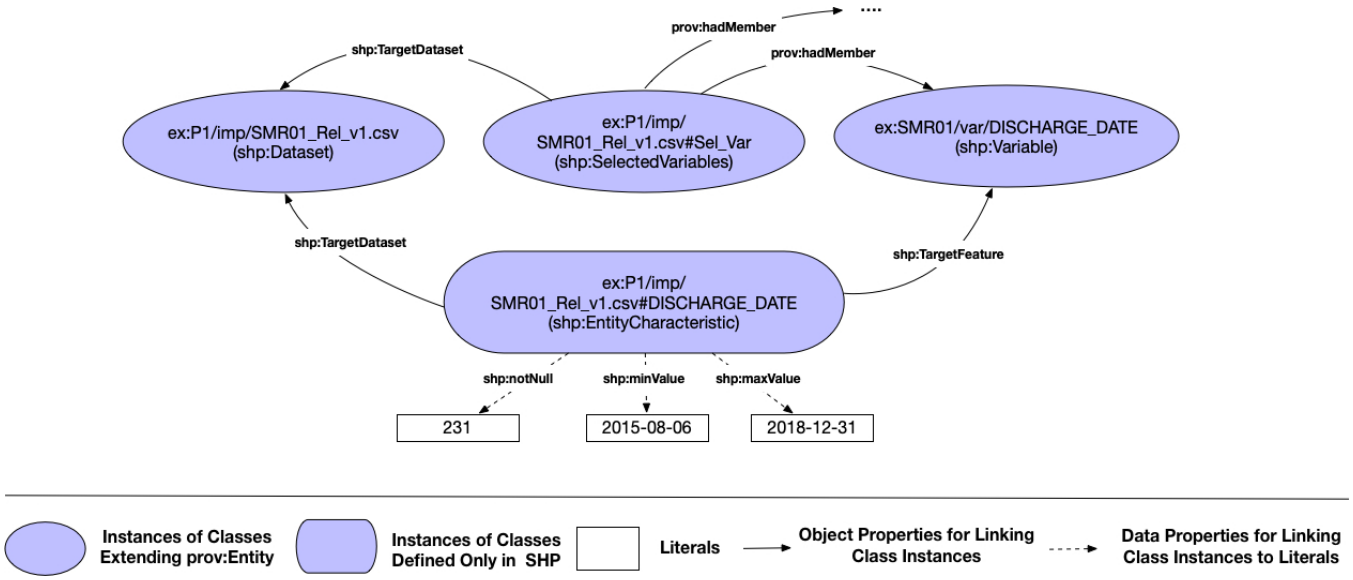


Figure 5: An example provenance graph describing aggregated data statistics related to Discharge Date variable contained within a dataset



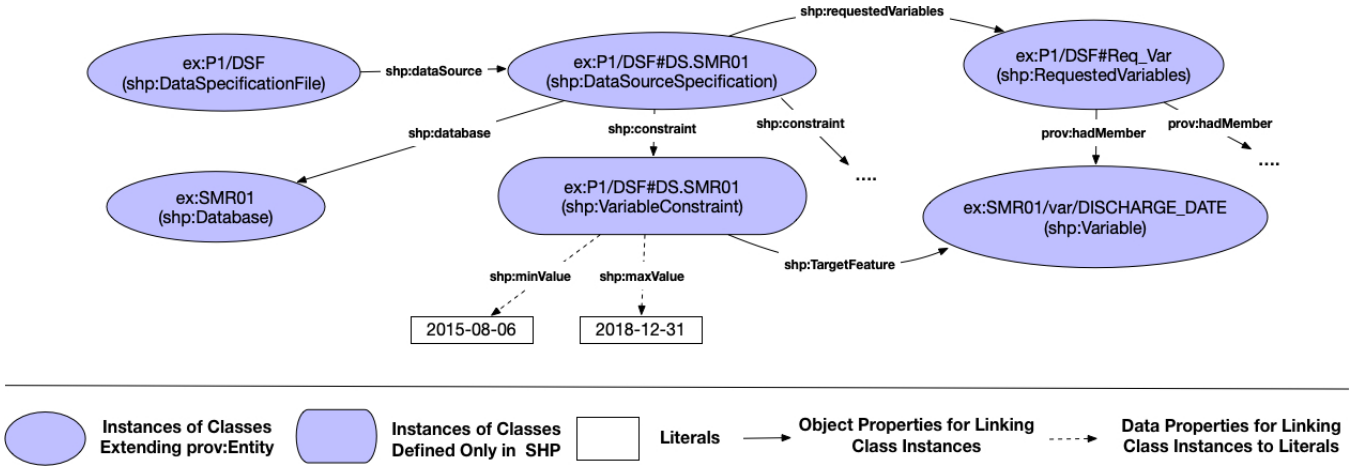
from which the data should be extracted, a collection of variables to be extracted (*shp:RequestedVariables*) and any additional constraints (*shp:VariableConstraint*) that should be applied to individual variables. For example, Figure 6 (below) illustrates a constraint *ex:P1/DSF#VC.SMR01* imposed on the *ex:SMR01/var/DISCHARGE_DATE* variable to state that only records between 06 Aug 2015 and 31 Dec 2018 should be extracted.

Representing SHP Provenance Traces as RO-Crates

The RO-Crate vocabulary has been developed as a common representation for research data and their associated metadata [30]. Recently, RO-Crate has gained visibility in the UK TRE landscape as one of the potential emerging standards to create a federated, interoperable TRE infrastructure [31] for exchanges of information in a common format. As such, the RO-Crate vocabulary can be seamlessly aligned with the SHP vocabulary to capture additional metadata descriptions and further enhance the information described by the RO-Crate. In our context, RO-Crates describe a knowledge graph consisting of *Data Entities*, *Contextual Entities* and *Actions* using the JSON-LD [32] format. Below, we illustrate how the RO-Crate and the SHP ontologies were combined to provide the means for capturing semantic descriptions of the data linkage process.

Actions describe any processes or methodologies applied to Data Entities. In our approach, we capture the actions that lead to a creation of new dataset files (for example, when a signed off file is released to a researcher and the released version of the dataset file is created in the researcher directory). Figure 7 (below) illustrates an example description of such data release activity. To make the files more readable for humans, JSON-LD supports the use of lookup keys which refer to the full IRI of the referenced concept defined separately in the context element (not shown in figure). For example, in our example "agent" refers to *Class: prov:Agent* defined in the PROV-O ontology [33]. The activity is associated with a unique ID (i.e. *ex:P1/release/2023-08-21T14:56:14+00:00*)

Figure 6: An example provenance graph describing a structure of the Data Specification File specifying the source, variables, and variable constraints for the data extraction activity



denoted by the "@id" element. The *"@type"* element stores information about the types of the action which in this case corresponds to "*Create Action*" (i.e. an RO-Crate term used for activities that create new resources) and a class name from the SHP ontology identifying this action as *Data Release Activity*. Furthermore, the label, human-readable description and the time at which the activity ended is also described. The description of the action also points to the "object" which is the identifier of the file used as an input and the "result" of the action (i.e. the released file). The files are then further described as Data Entities.

These are essential components of the RO-Crate representing the digital artefacts produced during the linkage process such as datasets, mapping files, data specification files, etc. Figure 8 (below) illustrates an example description of a dataset produced by the signoff activity. Each entity is again associated with a unique id (i.e. *ex:P1/signoff/SMR01_Rel_v1.csv*). The type of the entity in our example corresponds to "*File*" (i.e. an RO-Crate term used for data files) and "*shp_DataSet*" (i.e. the *Data Set* concept defined in the SHP ontology). Among other metadata such as label and description, the data entity also includes a description of the file hash, and information about the creator (i.e. the *wasAttributedTo* element points to identifier *ex:staff/Alice*). Finally, the "*exifData*" element points to identifiers of additional Contextual Entities associated with this data file (i.e. in this case, information about the data contained in this file that corresponds to the Discharge Date variable).

Contextual Entities in RO-Crate provide additional metadata that supports the understanding and reuse of Data Entities. These entities can describe people (like researchers or contributors), organisations, instruments used in the research, or even funding information. Contextual entities describe the additional context of the data outputs. Figure 9 (below) illustrates a high-level description of the data contained in the signed off file (see Figure 8) relating to the *Discharge Date* variable.

The example snippet includes the min and max values contained in the data file, which are one of the data quality indicators of each variable considered during the data validation of the data extract. The "target file" and "target feature" are utility concepts and capture links to the relevant *Data Entities* to which this *Contextual Entity* refers.

Semi-Automated Provenance Trace Creation

To create the provenance knowledge graph described using SHP and RO-Crate vocabularies, we aimed to automate as many data collection steps as possible to prevent introducing additional, manual tasks, whilst considering the need to manually inspect the provenance trace to provide robust quality assurance of the data processing as well as the automation function itself.

The semantic descriptions of the data specification file (i.e. the description of the datasets and variables associated constraints requested by researchers) were created automatically by augmenting the internal DaSH project request process. The previous process requiring researchers to complete Excel and Word document forms was replaced by a prototype web-based system [34]. Among other details (e.g. project description, project aims, etc.) the system allows researchers to select datasets, variables and associated constraints (e.g. min and max ranges, data sources, etc.). In the web-based system researchers use selectable drop-down menus to make their variable selections, replacing the existing manual process that relies on the manual completion of Excel and Word documents. The web tool then automatically generates the semantic descriptions of corresponding researcher choices using SHP and RO-Crate vocabularies and stores them in a JSON-LD file that forms the first part of the provenance knowledge graph. The remaining parts of the provenance knowledge graph were created by automated PowerShell scripts deployed in the Microsoft Server Environment [35]. The scripts observed the changes in the predefined folder structure, which was designed to support record-keeping during the data linkage process and follows the five main stages of data processing: import extracted data, export processed data, check processed data, sign off processed data and release processed data. Results of Data Analyst actions (e.g. creation of pseudo-anonymised dataset files) were recorded in timestamped subfolders corresponding to one of the five stages. Actions such as creation of new files in specific folders then triggered automated descriptions of

Figure 7: A snippet of a JSON-LD representation of the RO-Crate describing the data release activity


```
{
"@id":"ex:P1/release/2023-08-21T14:56:14+00:00",
"@type":["CreateAction", "shp_DatasetRelease"],
"agent":[{"@id":"dash:staff/Alice"}],
"description":"Data made available to researchers.",
"label":"Data Release",
"endTime": "2023-08-21T14:56:14+00:00",
"object":[ {"@id":" ex:P1/signoff/SMR01_Rel_v1.csv"}],
"result":[ {"@id":" ex:P1/release/SMR01_Rel_v1.csv "} ]
}
```


Figure 8: A snippet of a JSON-LD representation of the RO-Crate describing a signed off dataset file


```
{
"@id":"ex:P1/signoff/SMR01_Rel_v1.csv",
"@type":["File", "shp_DataSet"],
"description":" In-patient hospital data capturing information throughout
patient's hospital stay, diagnosis and operations performed.",
"label":"P1/signoff/SMR01_Rel_v1.csv",
"path":"file:///c:/P1/signoff/SMR01_Rel_v1.csv",
"shp_hash":"e0d675e5f316bef78bfdf5a008837600",
"wasAttributedTo":{"@id":"ex:staff/Alice"},
"exifData":[
   {"@id":" ex:P1/signoff/SMR01_Rel_v1.csv#DISCHARGE_DATE" }
]
}
```


new *Data Entities* and *Actions* in the provenance knowledge graph.

Additional provenance information such as links between the datasets and their database of origin (i.e. *prov:wasDerived From*) were automatically created by processing the names of the created files which followed predefined naming convention (i.e. <project id>_<*source database*>_<*file workflow stage*>_<*version*>.*csv)*. Furthermore, a separate script processed each of the newly created files to generate a set of corresponding *Context Entities* that described summary statistics of the data file contents (e.g. min and max data values corresponding to individual variables contained within the dataset).

In our test deployment, the provenance knowledge graph required only minimal manual data input, such as human readable text summaries of the dataset contents (e.g. inpatient hospital data capturing information about patient's stay) and the results of the sign off activities (which could not be collected automatically at the time of the experiment, as the sign off system existed outside the TRE infrastructure).

The data collection scripts in our test deployment are a TRE-specific data collection layer that was designed specifically for the DaSH operational environment to demonstrate the feasibility of semi-automated provenance collection in a real TRE environment. The scripts are generic and will extract provenance for all standard DaSH projects producing .csv files (support for other file formats is planned as a future enhancement). Other TRE environments may implement their own data collection approaches. For example, a cloud-based TRE implementing a workflow management system could extend its logging functionality to automatically generate relevant SHP and RO-Crate provenance descriptions

Figure 9: A snippet of a JSON-LD representation of the RO-Crate describing additional metadata about the Discharge Date variable contained in the signed off dataset file


```
{
"@id":" ex:P1/signoff/SMR01_Rel_v1.csv#DISCHARGE_DATE",
"@type":["shp_EntityCharacteristic"],
"shp_dataType":"date",
"shp_minValue":{
  "@value": "06-08-2015",
  "@type": "http://www.w3.org/2001/XMLSchema#date" },
"shp_maxValue":{
  "@value": "31-12-2018",
  "@type": "http://www.w3.org/2001/XMLSchema#date"},
"shp_targetFile":{"@id":" ex:P1/signoff/SMR01_Rel_v1.csv"},
"shp_targetFeature":{"@id":"ex:SMR01/var/DISCHARGE_DATE"}
}
```


at individual workflow stages. Provided that the resulting RO-Crate descriptions are aligned with the terminology defined in the SHP ontology the provenance records can be viewed and analysed using our provenance explorer tool described in the next section.

Provenance Explorer for Trusted Research Environments (PE-TRE)

Building on the low fidelity prototype, we developed a Java-based prototype desktop application for visualising provenance data. The PE-TRE visualises provenance traces stored as files in a JSON-LD format that comply with the RO-Crate specification as well as the associated comments described in a separate file using the standard W3C Web Annotation Vocabulary [36].

The main screen of PE-TRE (Figure 10a) lists the project details at the top with a button linking to a separate, detailed view of the data specification file (see Figure 11). Below this is an interactive list of files that were released to the researcher (Figure 10b). The list of individual activities that were performed during the data linkage process is displayed further below on the left side of the interface (Figure 10c). When a specific activity is selected, more details are shown on the panel to the right of the activity list. This includes the description of the activity, staff who completed the activity, list of activity inputs and outputs, results of any associated automated validation checks, and additional comments associated with this activity (Figure 10d).

Figure 11 (below) illustrates a detailed view of the variable specification file. The view contains A scrollable list with information about the requested variables, data sources, and additional variable constraints specified by the researcher (Figure 11a). In addition, the view also contains results of a number of validations, for example, to determine whether the files imported into the DaSH environment contain data linked to all Data Sources specified in the Data Specification File (Figure 11b). Comments added either by the Data Analysts or researchers are also displayed (Figure 11c).

Finally, Figure 12 (below) illustrates the Dataset details view that is used to inspect the individual dataset files in more detail. The view contains a brief description of the file, its location in the file space and some high-level statistics such as row count (Figure 12a). The results of the validation rules which focus on validating dataset files against constraints defined in the Data Specification File and internal list of sensitive variable names are also included (Figure 12b). Additionally, Figure 12b displays a failed validation check caused by a variable introduced during the data linkage process (*DaSH 123 StudyNumber SMR01*) as the variable is present in the dataset file but not in the Data Specification File, indicating that an additional variable has been extracted that should not be present in the data released to researchers. Aggregated data statistics calculated for each variable present in the dataset file are displayed in a scrollable list with small numbers highlighted (Figure 12c). Finally, the tool also enables adding comments by analysts to enhance understanding of the presented data (Figure 12d).

Using semantic annotations for automated validation

Modelling provenance traces as knowledge graphs opens opportunities for intelligent rule-based data validation that can increase the data quality and speed up manual efforts during the data validation process. For example, rules can be expressed as SPARQL [37] queries that evaluate logical patterns over the provenance graphs and can highlight the dataset files that violate some predetermined conditions. Figure 13 (below) illustrates an example of a generic SPARQL

Figure 10: Main screen of the Provenance Explorer: a) Project details and navigation button to the Data Specification File view; b) A list of released dataset files; c) A list of activities performed during the data linkage process; d) Activity details including descriptions of inputs, outputs, results of validation rules, and comments

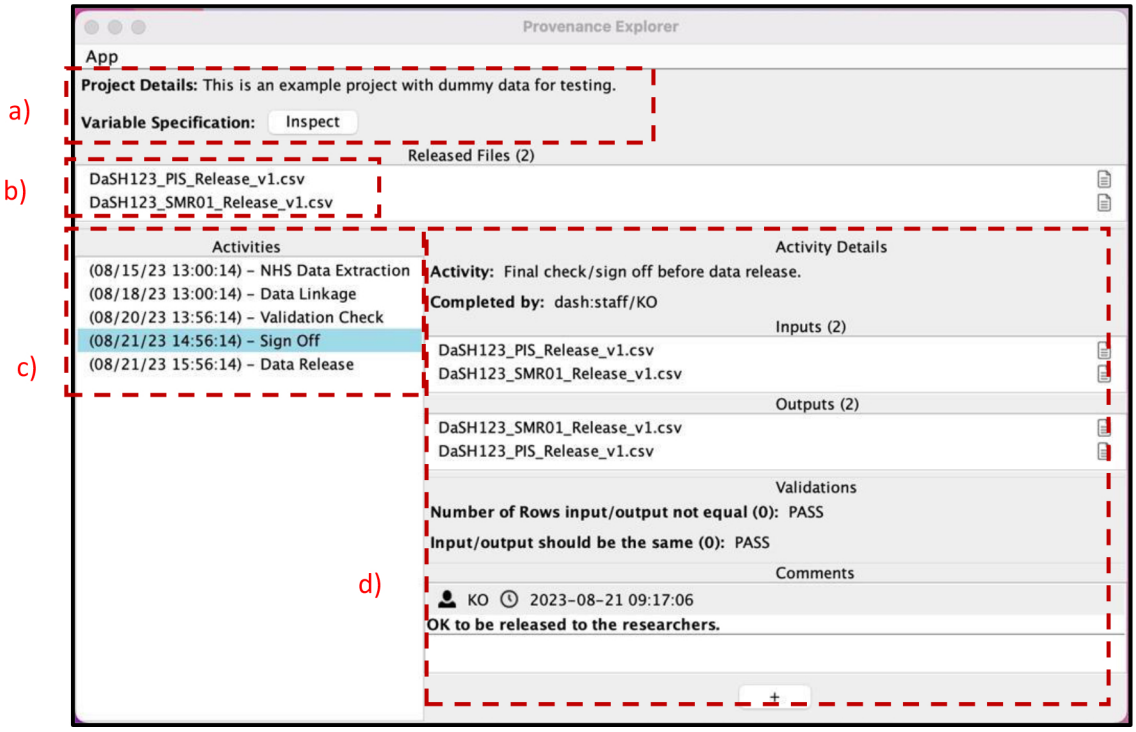
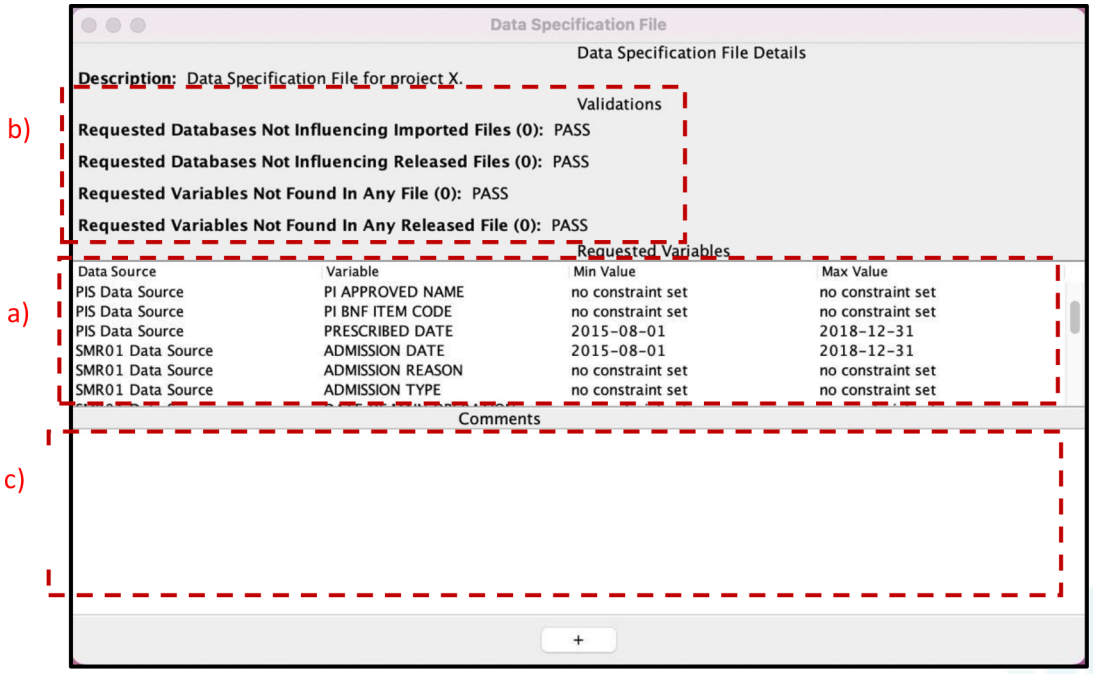


Figure 11: The Data Specification File view: a) A list of requested variables including the source databases and any additional constraints; b) Results of the validation rules; c) Comments added during the data linkage process



query for evaluating inputs and outputs of a specific activity where the activity is specified by an activity IRI at the time the rule is executed. In this example, the information about the file hash (i.e. a unique fixed length string representation of the file content) recorded for the activity inputs is compared with the hash recorded for the activity outputs. For some activities such as a data release, the files should not change (i.e. hash values should be the same) since they have already passed all the required modifications and are only needed to be released to the researcher for analysis within the TRE. Thus, if the file is changed for any reason during the release process this would be an unexpected/unauthorised change discovered by the validation rule and a potential deviation or breach within the TRE would be prevented.

Figure 12: The Dataset Details view: a) Details of the inspected file together with high level aggregate statistics; b) Results of the validation rules; c) Aggregated data statistics per each variable present in the dataset file; d) Comments added during the data linkage process

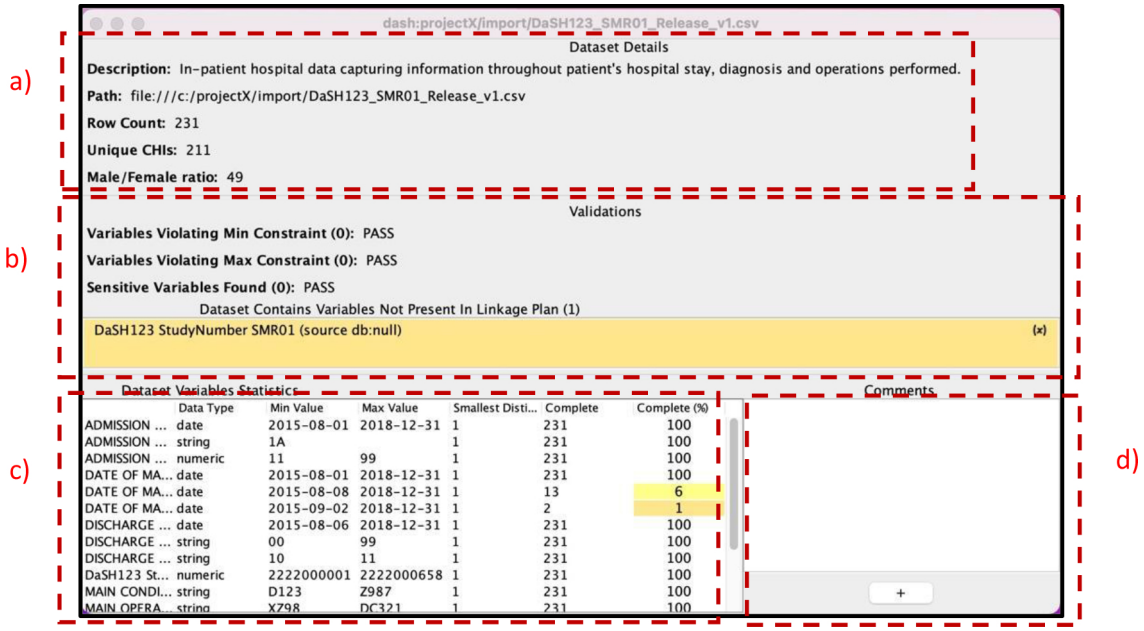


Figure 13: An example SPARQL query to test whether the data stored in the CSV files processed by an activity (e.g. a data release) have changed based on the comparison of the HASH associated with the files

```
SELECT DISTINCT ?input ?output ?datasetL

WHERE {<activityIRI> a <http://schema.org/CreateAction>;
                     <http://schema.org/object> ?input;
                     <http://schema.org/result> ?output.
       ?input a shp:DataSet; rdfs:label ?datasetL.
       ?output a shp:DataSet; rdfs:label ?datasetL.
       ?input shp:hash ?inputHash.
       ?output shp:hash ?outputHash.
       FILTER (?inputHash != ?outputHash)
      }
```

The prototype application includes a number of different validation rules; however, these could be easily extended in the future. Examples of the implemented validation rules include:

1. Checking number of rows contained in the dataset file before and after a specific activity. For some activities such as Sign Off, it is not expected to remove or add any data to the file (Figure 10d).

2. Checking the file hash of the file before and after a specific activity (Figure 10d). Similarly, to the previous validation check some activities should not be altering the content of the file however others such as Data Linkage are expected to. A hash function produces a unique string representation of the file content that can be used to evaluate whether any content of the file has been modified.

3. Checking the provenance trace against the Data Specification File to validate:
   a. whether all requested databases were queried, and results were imported in the environment (Figure 11b);
   b. derivation history of the released files to check if each of the requested databases yielded at least one dataset file released to the researcher (Figure 11b);
   c. the presence of requested variables, sensitive variables and unspecified variables in any of the dataset files (Figure 12b); and
   d. violations of constraints associated with requested variables caused by dataset file content (e.g. values out of specified date range) (Figure 12b).

User evaluation

Usability evaluation involves assessing how well a product or system meets the needs and expectations of its users in terms of effectiveness, efficiency and user satisfaction [38]. Since the design of the Data Provenance tool involved (potential) users

from the initial stages of the system development, carrying out a usability evaluation of the prototype identifies usability issues and areas for improvement that can be integrated into a full production version. The usability evaluation should therefore result in a more user-friendly and successful product once deployed as a live system and integrated into business processes.

Cognitive walkthrough of the prototype tool

We chose the 'cognitive walkthrough' approach for our evaluation of the prototype since many users prefer to learn software by exploration [39]. Participants in the cognitive walkthrough ($n = 7$) were asked to walk through a series of tasks to assess the PE-TRE interface as a new user of the application. The evaluation began with the purpose of the study and the intended use of the provenance tool. Participants were then asked to complete a series of tasks using the PE-TRE, and whilst interacting with the system, they were also asked to describe their experiences aloud. There were two observational analysis criteria used to assess the tool's effectiveness as well as to uncover potential issues and areas for improvement. The first criterion was users' ability to achieve the correct result and their understanding of the action(s) required to successfully perform an overall task. The second criterion was whether users observed several key aspects of the tool: (1) identified the correct item/action; (2) whether the interactive elements required to perform an action were visible and easy to identify; if labels were easy to understand; and (4) the user was able to successfully move through each stage of the overall task.

Overall, the usability evaluation participants were able to successfully perform tasks related to the requirements set out in the contextual enquiries and co-design workshops, confirming the effectiveness of the tool and its ease of use. All participants commented on the ease of use in finding the Data Specification File (e.g. the file completed by researchers that includes the required datasets and variables for the project, see Figure 10a, above) and the ability to inspect inputs and outputs in a side-by-side view (see Figure 10d, above). This allowed participants to carefully review the data files to ensure the data was as expected throughout the workflow. Additionally, they noted the usefulness of multiple comment boxes (see Figures 10d, 11c and 12d, above) that allowed them to log and reference important information during each stage of the data workflow, in addition to higher level comments that could be captured against the cohort or the datasets required for the project. All participants noted the ability to inspect the data regardless of the automated validations to ensure that the validation rules were working as expected to monitor the tool's performance. This evaluation observation is particularly important given the public's views on semi-automation as well as validating the public's system monitoring requirement as an important quality assurance and auditing mechanism.

However, some of the elements had lower usability effectiveness than others. The first was challenges in identifying some of the automated validations. Three participants struggled to find the automated validations on the main screen (see Figure 10d, above) although they are contained underneath the 'Outputs' box. All three said that the validation check location on the main screen meant that it got lost between the 'Inputs' and 'Outputs' sections and the 'Comments' section (which notably have a heading and a white box underneath, whereas the Validation section, although in bold text, is contained within the standard grey colour of the tool). This could be in part due to this being their first interaction with the dashboard and therefore they had to take in the system as a whole in a very short period of time. Interestingly, though, the same three individuals had no problems finding the validations on the 'Dataset details' screens (see Figure 12b, above), which has a different design layout. The second element that had lower usability effectiveness than others was related to labelling and terminology. Three participants noted that the terminology of validations was not easy to understand (see Figure 12b, above); in particular, they all misunderstood what the 'Dataset details' validations were intending to capture. The first two validations automate a check on whether any datasets have been erroneously included in the extraction and linkage in the data workflow, and the second two validations automate a check on whether any variables not selected by researchers are included in the data files. Three participants said the wording between the first two validations and second two validations were too similar and were not clear on what they were meant to check. All three participants also said that a 'zero' value could be misinterpreted, even if the validation states 'PASS'. Finally, the third element that had a lower usability effectiveness was again related to terminology: two participants noted that there was only a single value next to the 'male/female ratio' calculation (Figure 12a) and that the value did not seem to represent a ratio, but rather a percentage. They also noted that the percentage was unclear since it was a single value (49%), which they assumed was the percentage of males in the cohort (and then inferred that 51% of the cohort was female, but couldn't be sure).

Following the effectiveness tasks, participants were asked questions related to their perception of the system, including what results or consequences of using the provenance tool could be, and any improvements that could be made.

All participants deemed the PE-TRE to be useful in checking data in the workflow, helping to automate basic checks whilst also allowing for detailed inspection of each activity for detailed data validation checks. Participants who worked closely with researchers and/or governance teams also noted that it would facilitate better involvement for both groups in the quality assurance process. In particular, they explained that the PE-TRE would allow researchers to see data as it is being processed in aggregate, facilitating clinical sense-checking of the data extraction. For governance teams, they commented that they would be able to show this as evidence of a clear audit trail of all entities, activities and agents involved in producing the sensitive data for research and would meet any governance requirements for audits.

Improvements suggested included a change in the visual design of the dashboard by 'sectioning' the interface so that automated validations were easier to find, either by different colours, different sized fonts or other visual cues. Participants also agreed that the validation terminology should be improved before deployment so that it was more clearly worded so that all potential users of the system (e.g. Data Analysts, Governance staff and Researchers) would clearly understand what the automated validations referred to. Finally, the term

'male/female ratio' needed to be amended to 'percentage of males/females', with a % against each gender. They also suggested displaying the full numerical value of each, but noted that this could potentially contain small numbers depending on the research study, so acknowledge that percentage may be a better proxy in the current iteration.

Evaluation questionnaire

Participants of the usability evaluation were also provided with a separate evaluation questionnaire to complete anonymously after the cognitive walkthrough. The questionnaire was designed using the trust in technology constructs of functionality, helpfulness, and reliability [40] and following the Technology Acceptance Model 2 [41]. Survey questions and responses are available via Github [42]. The separate evaluation ensures that any responses given by the participants during the cognitive walkthrough, particularly those around the effectiveness or ease of use of the tool, do not differ from the participants' anonymously-provided responses and acts as a validation of the cognitive walkthrough.

In a number of key areas, participants exclusively agreed or strongly agreed that the tool would be useful in their work, including their intention to use the tool (agree or strongly agree = 100%) and perceived usefulness in performing their work (somewhat agree or strongly agree = 100%). The output quality was rated as effective (somewhat agree or agree = 100%), with all stating there were no significant issues with the system (somewhat agree or agree = 100%). Although some categories in the area of perceived usefulness were not rated as highly as others ('using the system improves my performance in my job'; 'using the system in my job increases my productivity'; 'using the system enhances my effectiveness in my job'), none of the participants disagreed, and the majority at least somewhat agreed, agreed or strongly agreed it would increase productivity and enhance effectiveness.

One area where responses were more contradictory to the verbal responses in the cognitive walkthrough was in the area of Job Relevance. Although all participants were either neutral or agreed that the usage of the system is important ($n = 7$, 100%), this is lower than the responses given during the cognitive walkthrough where all participants stated that the usage of the system would be important. This is potentially a discrepancy between perceived usefulness if the prototype were to be implemented, vs. the prototype evaluation where the tool was not yet in use. Similarly, 'In my job, usage of the system is relevant', there was one outlier who chose 'strongly disagree' – again, this may be due to the fact that the tool had not been fully deployed for use at the time of evaluation and therefore the responder saw the system as not being relevant to current ways of working or current working processes; all other participants responded neutral, somewhat or strongly agree. This suggests that potentially the question was misinterpreted, rather than the provenance tool being irrelevant to the participant's job.

Discussion

To our knowledge, this is the first design and implementation of interactive provenance tracking within a TRE) that offers the functionality to (semi-)automate provenance capture and reporting. Our approach presents an innovative solution by demonstrating that the PROV-O common data model can be implemented within the context of a secure data environment undertaking sensitive data extraction, pseudonymisation and linkage, and that it is possible to automate almost all components integral to the data processing workflow. The accompanying human-readable dashboard, PE-TRE, prioritises interactivity and assures that agents still manually inspect the data to reduce linkage errors. It offers transparency to researchers to guarantee the data for their research is error-free, contributing to increased reliability of the data underpinning research. Finally, it offers information governance teams an audit trail for quality assurance purposes across both individual projects and wider data processing regulations. Not only does our solution fulfil the key principles of the GUILD guidance, but it also demonstrates a major step in implementing the PROV common data model that, if adopted across the TRE landscape, can facilitate an interoperable, federated approach to provenance capture across heterogeneous environments. This is a fundamental requirement of the DARE UK Federated Architecture Blueprint [43], namely, that organisations operating as part of a federation adopt a set of standards and protocols that enable the exchange of information to improve trust and trustworthiness of sensitive data processing.

There were a number of challenges faced in the provenance collection and dashboard creation, in part due to the real-world constraints of the environment in which it was developed. First, a substantial portion of the provenance trace is inferred from the creation and movement of files between folders and sub-folders within the TRE (e.g. 'Import' folder, 'Export' folder, etc.); however, not all activities performed resulted in a file creation, which meant that some information could not be tracked or logged automatically. When this was identified during the course of the research, the TRE was able to change its folder and subfolder structure and naming conventions to improve the automated recording and classification of certain events. Many TREs use file creation and movement as a standard operating procedure, but they will also need to examine their folder and subfolder structure and naming conventions of both folders and files to help improve the trace. This will have the added benefit of more standardised folder and file naming systems to improve consistency between organisations looking to implement a provenance-based solution for improving transparency in the processing of sensitive data. Each organisation will have their own specific SOPs for file creation and movement such that they will need to customise the provenance trace to their specific requirements. However, we have provided the base code for the trace open source so that organisations can amend it to their environments and workflows whilst maintaining overall utility and adherence to the framework we have designed. Any adjustments to the code maintaining the overall SHP Ontology will be able to utilise the PE-TRE dashboard without any adjustments.

Similarly, the provenance creation will need to be customised if file creation and movement are not standard SOPs; for example, a TRE could use direct SQL-to-SQL transfer methods. The provenance monitor would need customising to capture SQL-to-SQL log data. As previously mentioned, cloud-based TREs could implement a workflow

management system to extend logging functionality to generate relevant SHP and RO-Crate provenance descriptions at specific workflow stages. Finally, while PowerShell offers a low-entry programming language for implementing provenance capture mechanisms, more powerful programming solutions might be more suitable for particular production systems due to the need for processing very large data files which are common in TREs. However, by offering our code open-source, organisations using SQL-to-SQL or other data creation and transfer mechanisms or architectures can use our code as the basis for categorising and amending code according to their specific processes. What we have offered is a proof of concept that automated provenance tracking is possible in secure data environments on sensitive data, and we have provided a clear methodology and all necessary components that can be customised to an individual TRE's specifications.

One limitation in our existing prototype is that certain steps in our test environment, such as approvals/rejections of validation checks and data transfer/release signoffs, sit in software managed outside of the TRE infrastructure; therefore enhanced metadata about these activities could not be automatically captured within the TRE environment in a straightforward way and in the timeframes of this research project (nine months). Future improvements would be to either bring the validation check and signoff software into the TRE and capture the information automatically, or to change the mechanism for the approvals/rejections of validation checks and data transfer/release signoffs if the software was not able to be brought into the TRE environment.

Finally, the framework prototype is currently optimised for use by Data Analysts within the TRE and does not offer segmented or audience-specific 'views' for information governance teams or researchers. Although this was an original goal to produce audience-specific views of PE-TRE, the length of the research project (nine months) meant that this was unachievable during such a short period of time. However, the tool in its current form can certainly be used to demonstrate quality assurance to IG teams by offering a single, accessible track of the data as it progresses through the workflow; likewise, Data Analysts are easily able to view any potentially disclosive fields to assess whether they are able to share project data during its production with researchers.

The ease of scalability across the UK TRE landscape is one of resourcing, but the methods and outputs we provide are reproducible and customisable. To implement this at scale, each TRE would need to map its workflow to the SHP Ontology and would need to update descriptions and integrate with the RO-Crate vocabulary, amending the open-source code we have published to their own environment and processing protocols. As previously noted, however, some processes may need to be altered within the TRE to better capture reportable information in an automated way.

Future work will focus on developing audience-specific 'views' to display essential information more relevant to user groups. We are also interested in capturing the analysed project data and the data egressed from TREs as part of the output disclosure control process to provide a full, end-to-end provenance trace from ingress into the TRE to egress from the TRE, further enhancing quality assurance and auditing. We will also explore provenance tracking scalability in an interoperable, federated network of TREs, including demonstrating utility within cloud-based environments, to ensure quality assurance across organisations when preparing data for both pooled data provisioning as well as data in situ for federated querying across a variety of architectures.

Conclusion

In this paper we have presented the Provenance Explorer for Trusted Research Environments (PE-TRE), a prototype tool for semi-automating data provenance tracking in TREs. Our approach presents an open-source solution for TREs to improve openness, transparency, and trustworthiness in the data workflow in producing sensitive data for research, which is customisable to an individual TREs requirements. Whilst there are still some areas of the extraction process that are not tracked by the tool and require further development, it was acknowledged by user groups and the public there is significant benefit in providing quality assurance to sensitive data processing for research in the public benefit. Further implementation of our PE-TRE in TREs will improve consistency and equivalency of service delivery within and across organisations. Moreover, by using coding vocabulary that facilitates federated solutions between organisations, our PE-TRE has the potential to contribute to a truly interoperable UK TRE landscape.

Acknowledgements

We would like to acknowledge the DaSH Team in delivering and evaluating the PE-TRE (Adrian Martin, Helen Rowlands, Joanne Lumsden, Vicky Munro, Amal Sebastian and Michael Gent) and the support and contributions from researchers at the University of Aberdeen Centre for Health Data Science and NHS Grampian (Jessica Butler, Simon Sawheny, Dimitra Blana) and University of Aberdeen and NHS Grampian Information Governance teams (Jody McKenzie, Fiona Stuart and Alan Bell), and University of Aberdeen's Department of Natural and Computing Science (Ana Ciocarlan). Further acknowledgement to contributors to the DARE UK Semi-Automated Risk Assessment project (PI: Arlene Casey), with contributions from DataLoch (Stuart Dunbar, Amy Tilbrook), West of Scotland Safe Haven (Charlie Mayor), eDRIS (Scottish National Safe Haven) (Jackie Caldwell) and University of Sussex (Liz Ford).

Conflict of Interest statement

None Declared.

Ethics statement

We received ethical approval from the University of Aberdeen Physical Sciences & Engineering Research Ethics Review Board for this project (Project ID: 652841; 14 April 2023).

Data availability statement

All data outputs from this research are made available open source via: https://github.com/TRE-Provenance/TRE-Provenance-Monitor.

Funding statement

This work was funded by UK Research & Innovation [MC_PC_23005] as part of Phase 1 of the DARE UK (Data and Analytics Research Environments UK) programme, delivered in partnership with Health Data Research UK (HDR UK) and Administrative Data Research UK (ADR UK). Previous work was funded by the Wellcome Trust (project ref. 219700/Z/19/Z).

References

1. Gilbert, R, Lafferty, R, Hagger-Johnson, G, Harron, K, Zhang, L, Smith, P, Dibben, C. and Goldstein, H. GUILD: Guidance for Information about Linking Data sets. Journal of Public Health 2018 40(1):191–198. https://doi.org/10.1093/pubmed/fdx037

2. Munafò, M, Nosek, B, Bishop, D. et al. (2017) A manifesto for reproducible science. Nature Human Behaviour 2018 1:0021. https://doi.org/10.1038/s41562-016-0021

3. Scheliga, B, Markovic, M, Rowlands, H, Wozniak, A, Wilde, K, Butler, J. Data provenance tracking and reporting in a high-security digital research environment. International Journal of Population Data Science 2022 7(3). https://doi.org/10.23889/ijpds.v7i3.1909

4. Butler, J, Scheliga, B, Markovic, M., Rowlands, H. Safe Haven Provenance Ontology [Internet]. 2022 [cited 1 June 2024]. Available from: https://safehavenprovenance.github.io/.

5. Markovic, M, O'Sullivan, K, Dymiter, J, Martin, A, Rowlands, H, Ciocarlan, A, Odo, C, Robb, C. Trusted Research Environment Ontology. 2023 [cited 1 June 2024]. Available from: https://tre-provenance.github.io/SHP-ontology/releases/v0.2/index-en.html

6. Scottish Government. Charter for Safe Havens in Scotland. 2015 [cited 1 June 2024]. Available from: https://www.gov.scot/publications/charter-safe-havens-scotland-handling-unconsented-data-national-health-service-patient-records-support-research-statistics/pages/4/.

7. Information Commissioner's Office. Principles and grounds for processing. 2023 [cited 1 June 2024]. Available from: https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/the-research-provisions/principles-and-grounds-for-processing/.

8. The UK Caldicott Guardian Council. *The Caldicott Guardian Principles*. [cited 1 June 2024]. Available from: https://www.ukcgc.uk/the-caldicott-principles.

9. Gao, C, McGilchrist, M, Mumtaz, S, Hall, C, Anderson, LA, Zurowski, J, Gordon, S, Lumsden, J, Munro, V, Wozniak, A, Sibley, M, Banks, C, Duncan, C, Linksted, P, Hume, A, Stables, CL, Mayor, C, Caldwell, J, Wilde, K, Cole, C, Jefferson, E. A National Network of Safe Havens: Scottish Perspective. Journal of Medical Internet Research 2022 '4(3):e31684. https://doi.org/10.2196/31684

10. Gao, C, Mayor, C, McCall, S, Wilde, K, Cole, C, Jefferson, E. Towards a Scottish Safe Haven Federation. International Journal of Population Data Science 2022 7(3). https://ijpds.org/article/view/1837

11. Desai, T, Ritchie, F, Welpton, R. (2016). Five Safes: Designing data access for research. 2016 [cited 1 June 2024]. Available from: https://www2.uwe.ac.uk/faculties/BBS/Documents/1601.pdf

12. O'Sullivan, KK, Wilde, K. A profile of the Grampian Data Safe Haven, a regional Scottish safe haven for health and population data research. International Journal of Population Data Science 2023 4(2). https://doi.org/10.23889/ijpds.v4i2.1817

13. W3C. PROV-Overview. PROV-O: The PROV Ontology. 2013 [cited 1 June 2024]. Available from: https://www.w3.org/TR/2013/NOTE-prov-overview-20130430.

14. Curcin, V, Fairweather, E, Danger, R, Corrigan, D. Templates as a method for implementing data provenance in decision support systems. Journal of Biomedical Informatics 2017 65: 1-21. https://doi.org/10.1016/j.jbi.2016.10.022

15. Xu, S, Rogers, T, Fairweather, E, Glenn, A, Curran, J, Curcin, V. Application of Data Provenance in Healthcare Analytics Software: Information Visualisation of User Activities'. AMIA Joint Summits on Translational Science Proceedings, 2017: 263–272.

16. Johnson, K, Kamineni, A, Fuller, S, Olmstead, D, Wernli, K. 'How the provenance of electronic health record data matters for research: a case example using system mapping. EGEMS 2014 2(1): 1058. https://doi.org/10.13063/2327-9214.1058

17. Black, JE, Terry, A, Cejic, S, Freeman, T., Lizotte, D, McKay, S, Speechley, M, Ryan, B. Understanding data provenance when using electronic medical records for research: Lessons learned from the Deliver Primary Healthcare Information (DELPHI) database. International Journal of Population Data Science 2023 8(5). https://doi.org/ 10.23889/ijpds.v8i5.2177

18. Can O, Yilmazer D. Improving privacy in health care with an ontology-based provenance management system. Expert Systems 2020 37:e12427. https://doi.org/10.1111/exsy.12427

19. Sun, Y, Lu, T, Gu, N. A method of electronic health data quality assessment: Enabling data provenance. IEEE 21st International Conference on Computer Supported Cooperative Work in Design (CSCWD) 2017: 233-238. https://doi.org/10.1109/CSCWD.2017.8066700

20. Trischler, J, Pervan, SJ, Kelly, SJ, Scott, DR. The Value of Codesign. Journal of Service Research 2018 21: 75–100. https://doi.org/10.1177/1094670517714060

21. Bannon, LJ, Ehn, P. Design matters in participatory design. Routledge handbook of participatory design 2012, 37–63.

22. Beyer, H, Holtzblatt, K. *Contextual Design: Defining Customer-Centered Systems*. 1988. San Francisco: Morgan Kaufmann.

23. Mirel, B. Contextual inquiry and the representation of tasks. ACM SIGDOC Asterisk Journal of Computer Documentation 1996 20(1), 14-21.

24. Holtzblatt, K, Beyer, HR Contextual Design. Interaction Design Foundation - IxDF. 2014 [cited 1 June 2024]. Available from: https://www.interaction-design.org/literature/book/the-encyclopedia-of-human-computer-interaction-2nd-ed/contextual-design

25. Masthoff, J. The user as wizard: A method for early involvement in the design and evaluation of adaptive systems. In Weibelzahl, S, Paramythis, A, & Masthoff. J (eds). Fifth Workshop on User-Centred Design and Evaluation of Adaptive Systems: held in conjunction with the 4th International Conference on Adaptive Hypermedia & Adaptive Web-based Systems. 2006 [cited 1 June 2024]. Available from: http://www.easyhub.org/workshops/ah2006/doc/UCDEAS06_Masthoff.pdf.

26. Dunbar, S, Casey, A, O'Sullivan, K, Tilbrook, A, Ford, E, Linksted, P, Mayor, C, Caldwell, J, Markovic, M, Ciocarlan, A, Harrison, K, Mills, N, Wilde, K. SARA Public Involvement and Engagement Final Report. 2023 [cited 1 June 2024]. Available from: https://zenodo.org/records/10084410

27. Poveda-Villalon, M, Fernandez-Izquierdo, A, Fernandez-Lopez, M, Garcia-Castro, R. LOT: An industrial oriented ontology engineering framework. Engineering Applications of Artificial Intelligence 2022 111: 104755. https://doi.org/10.1016/j.engappai.2022.104755

28. Mulholland, C, Simpson, E, Abernethy, S. Risk assessment and mitigation in health data research. Findings from deliberative workshops and an online survey. 2023 [cited 1 June 2024]. Available from: https://doi.org/10.5281/zenodo.10229070

29. Hogan, A, Blomqvist, E, Cochez, M, d'Amato, C, Melo, GD, Gutierrez, C, Kirrane, S, Gayo, JEL, Navigli, R, Neumaier, S, Ngomo, ACN. Knowledge graphs. ACM Computing Surveys 2021 54(4): 1–37.

30. Soiland-Reyes, S, Sefton, P, Crosas, M, Castro, LJ, Coppens, F, Fernández, JM, Garijo, D, Grüning, B, La Rosa, M, Leo, S, Ó Carragáin, E. Packaging research artefacts with RO-Crate. Data Science 2022 5(2): 97–138.

31. Giles, T, Soiland-Reyes, S, Couldridge, J, Wheater, S, Thomson, B, Beggs, J, Gallier, S, Cox, S, Lea, D, Biddle, J, Doal, R, Tammuz, N, Wilson, B, Cole, C, Sapey, S, Thompson, S, Jefferson, E, Quinlan, P, Goble, C. *TRE-FX: Delivering a federated network of trusted research environments to enable safe data analytics* 2023 [cited 1 June 2024]. Available from: https://doi.org/10.5281/zenodo.10529669

32. JSON-LD Community Group. JSON for Linking Data. Cited 1 June 2024. Available from: https://json-ld.org/.

33. W3C. 41. Starting Point Terms: Agent. PROV-O: The PROV-Ontology. 2013 [cited 1 October 2024]. Available from: https://www.w3.org/TR/prov-o/#Agent.

34. Dymiter, J, Markovic, M. TRE Provenance Monitor 2023 [cited 1 October 2024]. Available from: https://github.com/TRE-Provenance/TRE-Provenance-Monitor.

35. W3C. *Web Annotation Vocabulary*. 2017 [cited 1 June 2024]. Available from: https://www.w3.org/TR/annotation-vocab/.

36. W3C. *SPARQL 1.1 Query Language*. 2013 [cited 1 June 2024]. Available from: https://www.w3.org/TR/sparql11-query/.

37. Markovic, M. TRE Provenance Variable Specification Tool. 2023 [cited 1 October 2024]. https://github.com/TRE-Provenance/Variable-Specification-Selection-Tool.

38. International Standards Organization. Part 11: Usability: Definitions and concepts. ISO9241-11:2018, Ergonomics of human-system interaction. 2018 [cited 1 June 2024]. Available from: https://www.iso.org/obp/ui/#iso:std:iso:9241:-11:ed-2:v1:en.

39. Wharton, C, Riemann, J, Lewis, C, Poison, P. The cognitive walkthrough method: a practitioner's guide. In Nielsen, J, Mack, R. (eds.). Usability inspection methods. 1994. 105–140.

40. McKnight, D, Carter, M, Thatcher, J, Clay, P. Trust in a specific technology: An investigation of its components and measures. ACM Trans Manag Inform Syst 2011 2(2). https://doi.org/10.1145/1985347.1985353

41. Venkatesh, V, Davis, F. A Theoretical Extension of the Technology Acceptance Model: Four Longitudinal Field Studies. Management Science 2000 46(2):186-204. https://doi.org/10.1287/mnsc.46.2.186.11926

42. O'Sullivan, K. TRE Provenance Reports: Evaluation Questionnaire. 2023 [cited 1 June 2024]. Available from: https://github.com/TRE-Provenance/Reports/blob/main/PE-TRE%20User%20Evaluation%20Questionnaire.pdf.

43. Baxter, R. DARE UK Federated Architecture Blueprint. 2023 [cited 16 October 2024]. Available from: https://dareuk.org.uk/wp-content/uploads/2023/04/DARE-UK-Federated-Architecture-1-Initial.pdf.