DARE UK

Participatory Design and Evaluation of a Data Provenance Dashboard: SARA project

Katherine O’Sullivan, Milan Markovic, Chinasa Odo, Ana Ciocarlan

31/10/2023

‘This work was funded by UK Research & Innovation [Grant Number 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)’.

UK Research and Innovation

HDRUK
Health Data Research UK

ADRUK
Data-driven change

Participatory Design and Evaluation of a Data Provenance Dashboard: SARA project

1. Background

In healthcare research, working with clinical datasets often involves navigating strict governance permissions to access routinely collected to allow for the secure processing and linking of health data for research purposes when it is not practicable to obtain consent from individual patients [1]. These datasets are not readily suited for analysis due to the inclusion of patient-identifiable information and therefore must be processed in accordance with strict legal and governance frameworks. As part of this process data are pseudo-anonymised so that individuals cannot be identified whilst also allowing for reproducibility of research outputs. In Scotland, the processing of this data takes place in Trusted Research Environments (TREs), also known as Safe Havens [2], where data analysts working within the TRE extract patient identifiable information, pseudo-anonymise and link the data according to the specific ethical permissions granted for the research project. Due to the sensitive nature of the data, however, researchers cannot see the data during processing, and have described this process as a ‘black box’ since the methods used are rarely shared, cannot be evaluated and may suffer from undetected errors [3]. The solution to this problem – maintaining patient confidentiality whilst ensuring transparency, openness and quality assurance – is the tracking and recording of data processing within a TRE, capturing and reporting on information about data, users and activities, known as data provenance [4].

This work package builds on results from our previous exploratory study funded by Wellcome (project ref. 219700/Z/19/Z), which focused on exploring opportunities for provenance-based metadata reporting within the Grampian Data Safe Haven (DaSH) environment based on standard provenance models [5]. However, to operationalise this approach within a real TRE two outstanding tasks must be addressed: (1) Co-design user interfaces for provenance data collection and visualisation with end users (e.g. data analysts, researchers, data 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 semi-automated manner.

This report documents the approach used to design and validate the introduction of data provenance tracking in a TRE. It reports on the participatory design sessions, low-fidelity design validations and, upon delivery of the prototype dashboard, the final useability evaluation.

2. Participatory design sessions

In this project, a participatory design approach was adopted to promote democratic collaboration with major stakeholders during the system’s development cycle. 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. This approach empowers users to have a say in shaping the system they will ultimately use, ensuring it better meets their requirements and preferences. For this project, we undertook participatory designs using three approaches: Contextual inquiries, user requirements interviews, and co-design workshops.

We followed a human-centred approach to investigate how to design transparent, accessible, and reliable interfaces. Focusing on user experience and employing appropriate methods to deliver and visualise information is the key to enabling user empowerment. Humans are susceptible to a variety of biases when they evaluate how information is presented to them [6], and they expect artificial systems to provide explanations akin to other humans [7]. This may influence the trusting beliefs of the users. Therefore, developing a good understanding of human aspects is essential in the design process.

2.1. Contextual inquiries

A contextual inquiry was conducted to gain deeper insights into the users' real-world work environment and their interactions with data provenance. This method involved observing users in their natural settings in order to understand user's activities and processes. The information gathered during these interviews played a pivotal role in shaping the design of the system. It informed key design choices, including user requirements, the creation of user personas, the selection of features, architectural decisions, and the overall content strategy. This user-centred approach ensured that the system would be well-tailored to the users' actual needs and workflows. The main requirements from the contextual inquiries were as follows:

Contextual Inquiries - Requirements for a Data Provenance Tracking and Reporting System

- The system should allow researcher to fill in a Data Specification that captures the required datasets and variables.
- Data capture from spec sheet should be designed in such a way that one can use it to compare extracted data.
- Put a mechanism to check that the imported data has not been corrupted.
- Make a list of what the sensitive data variables are and ensure they can be tracked in the data file to check throughout the data processing workflow.
- Duplicate values are highlighted.
- Ensure that all the sensitive data are excluded from the release table.

2.2. User Requirements Interviews

User requirement interviews were conducted to enable end users to articulate their expectations and needs from the proposed Data Provenance tool. These interviews were instrumental in understanding the desired functionality and performance of the tool by the proposed users: Data Analysts, Researchers and Information Governance specialists. Moreover, they served as the foundation for subsequent phases of specification, design, and the creation of user personas, which were used in collaborative design workshops. Headline requirements from these interviews by user group were as follows:

User Requirement Interviews – Researcher Requirements for a Data Provenance Tracking and Reporting Tool

- Make a list of what the sensitive data variables are and ensure they can be tracked in the data file to check throughout the data processing workflow.
- Dataset Selection Justification: Researchers are provided with a list of available patient records datasets. They choose from this list and provide justifications for their selection. These justifications include explaining why specific datasets are needed and where linkages should be established between them.
- System to automatically provide a summary of what has been extracted using the inclusion and exclusion criteria selected.
- Provide the range of the data extracted to show transparency.

User Requirement Interviews – Data Analyst Requirements for a Data Provenance Tracking and Reporting Tool

- Ability to track different versions of researcher requests (e.g. if additional data variables are requested after the first version release).
- Provide automated checks for basic data processing requirements (e.g. date ranges, min/max values, etc., number of rows, unique CHIs, to remove manual checks).

User Requirement Interviews – Information Governance Requirements for a Data Provenance Tracking and Reporting Tool

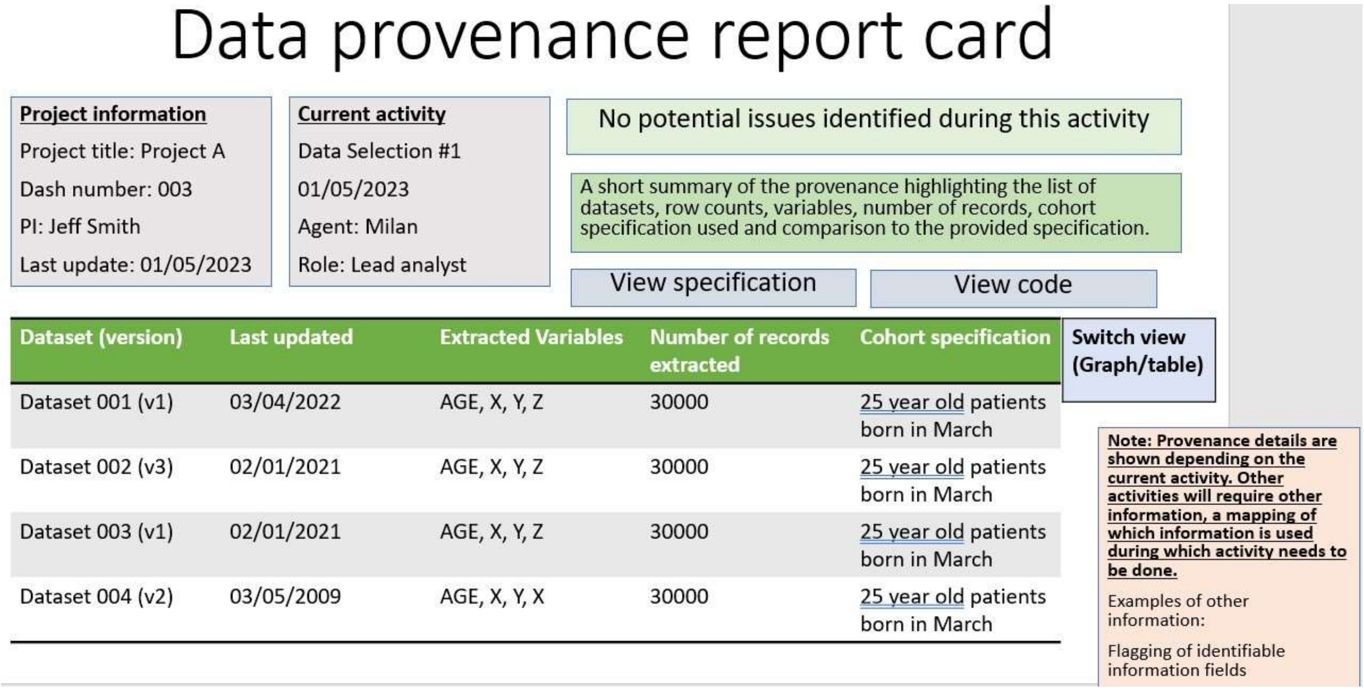
- Make a list of what the sensitive data variables are and ensure they can be tracked in the data file to check throughout the data processing workflow.
- Dataset Selection Justification: Researchers are provided with a list of available patient records datasets. They choose from this list and provide justifications for their selection. These justifications include explaining why specific datasets are needed and where linkages should be established between them.

2.3. Co-design Workshop and validation session

The purpose of the Co-design workshop was to co-create the Data Provenance tool system with the prospective users, applying the the-user-as-wizard methodology [8] to explore design challenges and solutions. This co-design session was run as a collaborative workshop that brought together stakeholders to generate ideas, solve problems, and create solutions collectively; invited participants were individuals connected to the Grampian Data Safe Haven (DaSH) as the prototype being developed was designed around the data processing workflow of DaSH. The validation session was conducted with users who participated in the previous codesign process and aimed to confirm that the proposed system design accurately represents the ideas generated during the workshop. This session provided users with an opportunity to review the design of the tool 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.

Low-fidelity Visualisation of User Requirements

The low-fidelity prototype represents a culmination of collaborative ideas from Data Analysts, Researchers and Information Governance specialists. Participants from all users wanted to monitor data movements, ensuring that the extracted data is free from sensitive information and to provide data processing summaries, aiding in the assessment and management of data provenance throughout the research process. Participants wanted a practical way to monitor project progress and ensure data quality and privacy.



Data provenance report card

Project information
Project title: Project A
Dash number: 003
PI: Jeff Smith
Last update: 01/05/2023

Current activity
Data Selection #1
01/05/2023
Agent: Milan
Role: Lead analyst

No potential issues identified during this activity

A short summary of the provenance highlighting the list of datasets, row counts, variables, number of records, cohort specification used and comparison to the provided specification.

View specification

View code

Dataset (version)	Last updated	Extracted Variables	Number of records extracted	Cohort specification
Dataset 001 (v1)	03/04/2022	AGE, X, Y, Z	30000	25 year old patients born in March
Dataset 002 (v3)	02/01/2021	AGE, X, Y, Z	30000	25 year old patients born in March
Dataset 003 (v1)	02/01/2021	AGE, X, Y, Z	30000	25 year old patients born in March
Dataset 004 (v2)	03/05/2009	AGE, X, Y, X	30000	25 year old patients born in March

Switch view (Graph/table)

Note: Provenance details are shown depending on the current activity. Other activities will require other information, a mapping of which information is used during which activity needs to be done.

Examples of other information:

Flagging of identifiable information fields

Figure 2: Data Provenance Tool Low-fidelity design

The main outcomes of the validation were:

- Provide updates / documentation of information if any changes occur during the data provenance process; this will keep stakeholders informed about changes or any modifications.
- 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.

3. Public Involvement – Low-fidelity prototype evaluation

Involving the public in evaluating our low-fidelity prototype 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, as well as include ideas within decision-making in the design and development of the prototype dashboard [9]. As noted in accompanying reports, enhancing the trustworthiness of data-access processes by listening and responding to public inputs was Enhancing the trustworthiness of data-access processes by listening and responding to public inputs is crucial to maintaining a social licence to processes confidential patient data for research purposes. A separate, detailed report on public engagement has been produced for this project [10], however, the main findings from workshops with the public are included here.

To ensure a transparent process, participants recommended:

- Having a decision log as part of the dashboards, including decisions made about data included or excluded (and why).
- Having a quality standard or statement about quality procedures to show staff had followed a consistent procedure.

To maintain confidentiality, participants recommended:

- Having low numbers banded to minimise the risk of patient identifiability.
- Replacing some details – like the type of drug or the specific condition – with less specific descriptions.
- Streamlining the dashboard so that they only contain the necessary information for those involved in the research to do their jobs effectively.
- Ensuring that there is 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 use of semi-automation was deemed acceptable (acknowledging the benefits of scale and speed) as long as there is human involvement to interpret the nuances and make decisions.

4. Usability 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. Since the design of the Data Provenance tool involved the user during the early stage of the system development, carrying out a usability evaluation in this project is to help identify usability issues and areas for improvement, resulting in a more user-friendly and successful product; the cognitive walkthrough' approach was chosen to evaluate the ease of new users to accomplish tasks within a newly-designed system [11]. Finally, participants of the useability evaluation were also provided with an evaluation questionnaire, guided by the trust in technology constructs of functionality, helpfulness, and reliability [12] and following the Technology Acceptance Model 2 [13]. The main findings from the Useability Evaluation are as follows:

Usability Evaluation of Data Provenance Tracking and Reporting System

- Allows researchers to fill in a Data Specification that captures the required datasets and variables (NB: this is a separate output (available from: https://github.com/TRE-Provenance) but integrates directly with the provenance tracking and reporting tool).
- Data capture from Data Specification is designed in such a way that it can be used throughout the data workflow to compare extracted data, automatically providing a summary of what has been selected and using inclusion/exclusion criteria provided.
- The automation is able to check that the imported data has not been corrupted, and there is functionality to undertake a manual inspection.
- Sensitive variables have an automated check, and there is functionality to undertake a manual inspection throughout the workflow, including release files.
- Each dataset/file can be inspected against other datasets/files to compare data inputs/outputs.
- Lists of sensitive data variables can be included as comments and can be tracked throughout the data processing workflow.
- Comment feature allows for duplicate values to be highlighted by manually including comments.
- Decisions can be captured by manually including comments.
- Activities timeline' recorded by date/time displays the relevant data processing stage and reflects if different versions of the data are produced (and undergo full processing in the timeline); 'Activities' timeline acts as a quality check to ensure all processing steps are undertaken in correct order.
- Automated checks for basic data processing requirements (date ranges, min/max values where constraints are set, number of rows in data, number of unique CHIs and male/female %).
- Dashboard made visually less cluttered than low-fidelity version to only show necessary/limited information unless specific selections are made.
- Dashboard allows for the Data Analyst to show aggregate data to Researchers during the data production phase for improved communication and transparency.
- TRE staff can undertake detailed (manual) inspection of all automated checks during processing to ensure that the semi-automation is working properly over time.
- Code can be inspected to ensure ongoing maintenance – no AI/machine learning included in tool to prevent biases or 'learned' mistakes.
- Tool designed for humans to inspect the data provenance to improve quality assurance, transparency and openness in data processing, whilst ensuring patient confidentiality and privacy.

Overall, participants in the usability evaluation felt that there were three main benefits to the new data provenance tool:

- Improve speed in checking data, as the tool brings together the data workflow in a single location that automates some standard checks whilst enabling more detailed manual checking, particularly the ability for side-by-side comparisons of inputs/outputs. Data Analysts felt that the tool would help them most effectively in validation checks of the data extraction, pseudonymisation and linkage. However, they also saw a benefit whilst building data extractions since the automation allowed for easy comparison against the Data Specification provided by researchers.

- Improve transparency for researchers in how the data was extracted and linked and allow Researchers to 'sense check' the data as it is produced by viewing the unique CHIs, row counts (which are often linked to particular health episodes, such as hospital admissions or medicines prescribed), and male/female percentage. Participants felt that the dashboard allowed them to inspect the records manually prior to sharing the dashboard views with Researchers to ensure patient confidentiality and privacy.

- Improved transparency for Information Governance/Research Coordinators to ensure the Datasets and variables processed in the data workflow match those that were approved by Ethics permissions.

Participants commented that the tool would provide an easy validation mechanism for quality assurance during internal/external audits that project data was produced in the correct order and would demonstrate to Information Governance teams that all data processing protocols were applied consistently.

5. Challenges and Limitations

There were some challenges and limitations in this process. The first was engaging the identified user groups consistently and throughout the process. Participation in the project was voluntary, and it was difficulty to involve Researchers and Information Governance experts not embedded in the DaSH team due to limited availability. The second was determining which 'requirements' by the user groups were considered to be in scope and out of scope for the project. Although data provenance involves the activities, entities and people associated with data production, this can be interpreted widely by participants, and therefore a number of other 'requirements' were introduced in interviews and workshops that did not sit fully within the scope of this project. For example, Researchers and Information Governance teams wanted better signposting and information about applying for ethical permissions. This is related to the Data Linkage Plan and Spec File production (since these documents are required for ethical permission applications) but are not specifically related to the capture of data as it proceeds through the workflow. As such, all requirements were captured, and the project team had to evaluate whether they fell under the scope of this project. Finally, some of the requirements that were considered in scope were not able to be actioned fully in the Data Provenance tool because the workflow cannot be fully documented within the environment – for example, details related to validation checks and signoffs occur outside of the environment (using Microsoft PowerAutomate Approvals functionality) and could not be easily integrated within the secure Safe Haven environment. These procedures for recording outcomes of validation checks and signoffs will need to be changed to take place within the environment but could not be properly scoped or delivered within the limited timeframe of this project.

References

[1] O'Sullivan, K. and Wilde, K. (2023) "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*, 4(2). doi: 10.23889/ijpds.v4i2.1817.

[2] Scottish Government. (2015) Charter for Safe Havens in Scotland: Handling Unconsented Data from National Health Service Patient Records to Support Research and Statistics [Internet]. Available from: https://www.gov.scot/publications/charter-safe-havens-scotland-handling-unconsented-data-national-health-service-patient-records-support-research-statistics/.

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

[4] W3C, PROV-O: The PROV Ontology. W3C Recommendation 30 April 2013. [Internet]. Available from: https://www.w3.org/TR/prov-o/#:~:text=The%20PROV%20Ontology%20(PROV%2DO,systems%20and%20under%20different%20contexts.

[5] Outputs produced by this project included the SHP ontology (https://www.w3id.org/shp), a light-weight vocabulary for documenting provenance of a data linkage workflow, and an initial version of a restricted software prototype demonstrating an interactive interface for visualising provenance information (https://github.com/SafeHavenProvenance/SH2PROV).

[6] Kahneman, D. (2011)*Thinking, fast and slow.* Farrar, Straus and Giroux.

[7] De Graaf, M., Malle, B. (2017). “How people explain action (and autonomous intelligent systems should too)”. *2017 AAAI Fall Symposium Series*.

[8] Masthoff, J. (2006) “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, Dublin, Ireland, 20/06/06*. <http://www.easy-hub.org/workshops/ah2006/doc/UCDEAS06_Masthoff.pdf>

[9] Dunbar, S., Tilbrook, A., O'Sullivan, K., Casey, A., (2023) *Final Public Involvement and Engagement Report: SARA project* [forthcoming: October 2023].

[10]. Mulholland, C., Simpson, E., Abernethy, S., *Risk assessment and mitigation in health data research: findings from deliberative workshops and an online survey*. Ipsos Scotland, September 2023 [forthcoming, November 2023].

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

[12] McKnight, D., Carter, M., Thatcher, J., and Clay, P. (2011) “Trust in a specific technology: An investigation of its components and measures”. *ACM Trans. Manag. Inform. Syst.* 2(2). DOI = 10.1145/1985347.1985353.

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