Institution: Newcastle University

Unit of Assessment: 11

Title of case study: Assuring data integrity and trust across organisational boundaries: The PROV Model for provenance

Period when the underpinning research was undertaken: 2011-2017

Details of staff conducting the underpinning research from the submitting unit:

Name(s):	**Role(s) (e.g. job title):**	**Period(s) employed by submitting HEI:**
Professor Paolo Missier	Professor of Big Data Analytics	2011 – Present
Professor Paul Watson	Professor of Computer Science	1995 - Present

Period when the claimed impact occurred: 2014-2020

Is this case study continued from a case study submitted in 2014? N

1. Summary of the impact (indicative maximum 100 words)

The ease with which electronic data are reproduced, modified, and shared is bringing vast opportunities accompanied by the major societal challenge of assuring data integrity. Data provenance, the record of the history of information products, can be used to engender trust in data and to facilitate its reproducibility. Yet, its widespread implementation has until recently been elusive, due in part to a lack of common representation and formal model. Newcastle University research has underpinned the design of an extensible community-based standard data model and formal ontology, denoted “PROV”, that has been adopted across geographies, sectors, disciplines and types of organisation.

PROV has become the most pervasive standard worldwide for seamlessly sharing data provenance within and across organisations, promoting positive changes in their working practices for data governance. This case study highlights its impact at NASA (USA), at the National Archives and the Gazette (UK), and at Astra Zeneca (global). It also mentions its role at NHS Digital, the Allotrope Foundation, and other organisations.

2. Underpinning research (indicative maximum 500 words)

The W3C Working Group (WG) on provenance was formed in 2011 following an earlier community effort, the Open Provenance Model (OPM). Missier was an active contributor to the OPM, and between 2011 and 2013 his research has enabled the development of a data model for provenance that met the challenges of an interoperable standard. Newcastle’s contribution is twofold. Missier’s research between 2011 and 2013 provided a solid foundation for the design and later adoption of a formal data model and ontology, and was instrumental to shaping PROV into a viable model that could be accepted, adopted, and extended by multiple communities. Then, after PROV was released, Missier’s further research demonstrated how a number of technical hurdles, including comparing two provenance documents [P5, P6] and safely providing abstractions over PROV in order to protect sensitive provenance information [P7], could be overcome to implement PROV into practical systems.

The design process itself was 2 years long and complex, requiring leadership and coordination across about 40 diverse organisations. Working along the Working Group Chairs, Missier took a leadership role in shaping the model, as evidenced by the key normative PROV documents where Missier is one of the key editors. These documents are:

[PROV-DM] The PROV Data Model: http://www.w3.org/TR/prov-dm/

[PROV-N] The Provenance Notation: http://www.w3.org/TR/prov-n/

[PROV-CONSTRAINTS] Constraints of the PROV Data Model: http://www.w3.org/TR/prov-constraints/, in addition to non-normative documents where Missier has been the principal designer (PROV-DICTIONARY: https://www.w3.org/TR/2013/NOTE-prov-dictionary-20130430/) and dissemination notes (PROV-PRiMER: https://www.w3.org/TR/2013/NOTE-prov-primer-20130430/).

These formal PROV documents incorporate design elements that are grounded in Missier's work.

- PROV is represented using both relational modelling and semantic modelling principles. Using the latter to express provenance was first proposed in [P1], which can therefore be considered a precursor to PROV-O, the Ontology Web Language (OWL) specification of PROV [http://www.w3.org/TR/2013/REC-prov-o-20130430/].
- A PROV plan is a generic modelling element that can be used to describe how a piece of data is produced. A PROV modelling pattern prescribes how this can be specialised to represent concrete processes, such as scientific workflows. Missier contributed first by clarifying formally how provenance should be structured to accommodate workflow plans with certain desirable characteristics [P2], and later by co-authoring an extension to PROV, called D-PROV (later "ProvONE") [P3], that lets scientists describe the process structure itself, as part of the history of the data produced by the process. Developed independently of the PROV WG and within the scope of the DataONE project (https://www.dataone.org/ -- a large US-based repository of Climate and Ecology datasets where Missier has been co-chair of the Provenance Working Group since 2010), this extension contributed to PROV's adoption within scientific communities including DataONE, where PROV-supported provenance is now actively promoted.
- In [P3], Missier also demonstrated that provenance could be expressed using Prolog, with advantages in terms of query capabilities over other implementation models. This representation has been used for instance in the ReComp research prototype (https://blogs.ncl.ac.uk/recomp/, EPSRC funding). The PROV-N document which he co-edited (https://www.w3.org/TR/prov-n/) embraces this approach by providing a Prolog-like syntax and model that is both human-readable and machine-processable.

In the post-release phase of PROV's life, Missier continued to produce research results that helped establish PROV's role as an important community data model, encouraging adoption. Firstly, a provenance data model is only useful if it supports a provenance database that can be queried effectively. Missier and his colleagues at UC Davis, USA, showed how this can be accomplished in practice [P4]. Secondly, Missier and Watson showed how two provenance traces obtained from the execution of two independent processes could be compared with each other [P5 and, more recently, P6]. This is a key requirement to enable the reproducibility of data produced using scientific workflows. Finally, Missier and colleagues proved a suite of formal properties of PROV models, showing that it is possible to abstract out elements of a provenance document without compromising its integrity. This is a key property when provenance is to be exchanged between organisations with limited mutual trust [P7].

3. References to the research (indicative maximum of six references)

[P1] Zhao, J., Sahoo, S.S., Missier, P., Sheth, A., & Goble, C. Extending Semantic Provenance into the Web of Data. IEEE Internet Comput. 2011;15(1):40–8. DOI: 10.1109/MIC.2011.7
[P2] Missier, P. & Goble, C. Workflows to Open Provenance Graphs, round-trip. Future Generation Computer Systems (FGCS). 2011; 27(6): 812--819. DOI: 10.1016/j.future.2010.10.012
[P3] Missier, P., Dey, S., Belhajjame, K., Cuevas, V., & Ludaescher, B., D-PROV: extending the PROV provenance model with workflow structure. In Procs. TAPP'13, Lombard, IL, 2013. DOI: 10.1.1.370.5403
[P4] Missier, P., Ludascher, B., Bowers, S., Altintas, I., Dey, S., & Agun, M. Golden Trail: Retrieving the Data History that Matters from a Comprehensive Provenance Repository. International Journal of Digital Curation. 2011; 7(1). DOI: 10.2218/ijdc.v7i1.221
[P5] Missier, P., Woodman, S., Hiden, H., & Watson, P. Provenance and data differencing for workflow reproducibility analysis. Concurrency and Computation: Practice and Experience. 2013; 28(4): 995–1015. DOI: 10.1002/cpe.3035
[P6] Thavasimani, P., Cala, J., & Missier, P. Why-Diff: Exploiting Provenance to Understand Outcome Differences from non-identical Reproduced Workflows. IEEE Access, 2019. DOI: 10.1109/ACCESS.2019.2903727

[P7] Missier, P., Bryans, J., Gamble, C., & Curcin, V., Abstracting PROV provenance graphs: A validity-preserving approach, Future Generation Computer Systems. 2020; 111:352 - 367. DOI: 10.1016/j.future.2020.05.015

Grants:
[G1] 2012-2013, Trusted Dynamic Coalitions, EPSRC / DSTL EP/J020494/1 (£98,000). Awarded to: Newcastle University. PI: P. Missier.
[G2] 2016-2019, ReComp: sustained value extraction from analytics by recurring, selective re-computation. EPSRC Making sense from data initiative, £585,000. Awarded to: Newcastle University. PI: P. Missier.
[G3] 2017-2020, CEM-DIT: Communication and Trust in Emergencies, funding: Office of Naval Research Global, £110,000. Awarded to: Heriot-Watt, Coventry, Newcastle University. PI: P. Missier.

4. Details of the impact (indicative maximum 750 words)

Newcastle University's research has contributed to the unique extensible design and widespread adoption of the PROV standard for data provenance. The model was endorsed by the World Wide Web Consortium (W3C) in 2013, and since then it has become the de facto standard for capturing and exchanging provenance. Data-intensive organisations have also adopted PROV for internal use to add value to their datasets.

PROV has gained extensive reach geographically (UK, EU, USA, Australia), across disciplines (Geoscience, Climate studies, Medicine, public information services) and sectors (Government, Business, Science). Links to web-published information by beneficiaries of PROV are collated in [E1, E5]. The investment required to incorporate PROV into existing data stores, creating extensions, and changing working practices is indicative of its value to beneficiaries.

The use cases below illustrate the level of impact of PROV on three high profile organisations: **NASA/ USGCRP (US Global Change Research Program**), global Pharma company **Astra Zeneca**, and the **National Archives** in the UK. An introduction to how they use PROV and why, is provided in [E1]. Below we highlight key points. In the words of their programme managers, the benefit has been in *making their information more authoritative and trustworthy in the eyes of their users*, and to engender positive changes of internal working practices concerning data governance.

1. NASA / USGCRP

NASA JPL manage the US Global Change Information System (GCIS) https://data.globalchange.gov/, on which the National Climate Assessment (NCA) reports in the USA are based. These publicly available reports, commissioned by the USGCRP (Global Change Research Program), inform and influence policy debate on climate change and the environment, within the USA and internationally. Impacts include:

Change in working practice & policy. The use of provenance in the GCIS was recommended by the Federal Advisory Committee on climate assessment and mandated by the US Administration (President Obama at the time) [E3 Appendix 3]. PROV, along with the GCIS ontology and other metadata vocabularies, is used systematically in the GCIS to *enforce the traceability of all of the about 50,000 individual resources held in the database* [E3 section 5]. According to Dr. Sherman of USGCRP [E6], *all contributors to the GCIS* are now required to curate their contributed content, including any climate science data. PROV metadata *must* now be supplied alongside any resource contributed.
Since 2012 NASA first, and now USGCRP, have demonstrated long-term commitment to PROV and to data curation more broadly, by providing sustained funding for 3 FTE staff [E6].

Effect on policy debate provided by transparency and assurance of the data held by the GCIS. PROV and the GCIS ontology have effectively promoted a culture of data curation and data trust within the climate science community. This percolates down to the reports themselves,

which are public, and where PROV elements are exposed both for human reading and in machine-processable formats (XML, RDF), providing trust into climate science as part of public discourse. According to Curt Tilmes, Data Scientist at NASA detailed to GCRP at the time, this was a significant benefit of PROV and enabled them to have "*incontrovertible consensus*" to support climate debate [E1 page 13].

NASA has extended the PROV model [E2] for use in the Planetary Data System (PDS4) which contains scientific data from the solar system planetary missions. The data, which includes historical datasets from the Voyager and Cassini missions is now continually updated and verified through the PROV extension. This enables valuable analysis: for example, the provenance schema for the Voyager ISS geometric calibrated images allows tracing information to be used by exoplanet scientists for analysis [E1 pg. 13].

PROV has benefitted satellite construction. Satellites are made up of a large number of constituent parts procured via a long chain of intermediary suppliers. Counterfeit parts pose a particular problem for NASA. Dr Tilmes states that PROV data provides the audit trail that allows identification of the source of counterfeit or faulty parts [E1 pg. 7].

2. UK National Archives and the Gazette

The National Archives in the UK (NA) maintains 11 million historical government and public records in addition to current documents. Their data underpin the UK Government Gazette and legislation database. To assure data after digitisation, the NA have mandated the systematic inclusion of provenance [E7] as part of all the documents published by the Gazette. Recorded web traffic to the UK Government Web Archive suggest over 1.7 billion redirected hits to the website (NA website).

Change of working practice as a result of the requirement by the National Archives (NA) to include provenance. All of Gazette data must now be supported by provenance statements. PROV has made this possible and cost-effective, as all the hard design work has already been done. The Gazette has committed dedicated staff to maintain and support provenance curation (Dr. Cresswell interview [E7]).

Maintenance of authority & correctness of data after digitisation. The Gazette is the authoritative source for public documents in the UK, as notices published by them are afforded legal standing - traceability and trust in the information are therefore paramount. Originally only available as physical verified documents, the data has maintained integrity through the use of PROV during the digitisation effort. Data retrieved from the Gazette must be traceable: It is "*the official public record, it has credibility, it has that kind of grand strength […] now there is a provenance trail for every single notice […] it will tell you what happened to the notice since it came to us for publishing, and every step that happens within that notice journey*." Janine Eves, Business and Operations Director, The Gazette [E4].

Traceability of legislation data. Legislation.co.uk, underpinned by the National Archives, is the official web-accessible database of the statute law of the United Kingdom. PROV provides traceability to legislation data. This has been especially useful to maintain the trace of legislation originating from EU law to support exit arrangements (Dr. Cresswell interview [E7]).

3. AstraZeneca

AstraZeneca is a global pharmaceutical company with a portfolio of speciality care and primary care medicines. According to Dr Tom Plasterer, Director of Bioinformatics, Data Science & AI [E8], a provenance model was needed to support internal processes and PROV provided a community-accepted solution for that. The process of adopting PROV along with other ontologies started in 2013 as part of a million-dollar project, where PROV is estimated to

account for about 5-10%, with continued maintenance to date. This effort resulted in a **change of working practices,** where the use of shared vocabularies now informs data governance and promotes transparency: "*its vital importance covers … processes from drug discovery, target identification, target validation, through to trial design, evaluation, clinical trials, and how data is managed*" [E1]. PROV has also enabled a new internal business intelligence system called CI360 to be developed, which bring **competitive advantage** to the company. It alerts scientists of possibly actionable news about competitors' products. The technology is based on the popular concept of "nanopublications" (http://nanopub.org/), where scientific statements are systematically annotated with provenance assertions. *PROV ensures that the statements are properly corroborated and thus safely actionable*.

Selected further cases of PROV in use.

Other notable cases of PROV adoption, are documented in [E5]. Amongst these:

- The Allotrope Foundation, an international consortium of pharmaceutical, biopharmaceutical, and other scientific research-intensive industries, develops and adopts specifications to standardize the acquisition, exchange, storage and access of analytical data captured in laboratory workflows. PROV is part of this suite.
- PROV has been adopted by Health Level Seven International (HL7) Fast Healthcare Interoperability Resources (FHIR), part of NHS Digital UK. Beneficiaries are clinicians, researchers, and regulators who are better able to trace, reproduce, and analyse scientific data.
- CSIRO, Australia's largest government research organisation, has extended PROV for earth sciences data.
- The Netherlands Government's database of national registration of land and buildings incorporate PROV extension, "BAG". Ministries, water boards, police forces and security regions, are obliged to use the data from the registrations.

5. Sources to corroborate the impact (indicative maximum of 10 references)

[E1] *Commissioned report*: Impact Evaluation of PROV, Cactus Impact Science, July 2020. https://www.impact.science/wp-content/uploads/2020/08/Evaluation-of-Impact-of-PROV.pdf. Also available are direct transcripts of interviews with all the corroborators mentioned in the report and in this ICS.

[E2] *NASA's PROV extension for PDS4:* https://113qx216in8z1kdeyi404hgf-wpengine.netdna-ssl.com/wp-content/uploads/2019/05/130_crichton.pdf.

[E3] *NCA report section:* U.S. Global Change Research Program (USGCRP) (2018). 4th National Climate Assessment: Data Tools and Scenario Products. https://nca2018.globalchange.gov/chapter/appendix-3/

[E4] *Transcript of Interview* with Janine Eves, Business and Operations Director, The Gazette, 30/1/2020

[E5] *A collection of references* to selected and notable documented implementations and extensions to PROV: https://blogs.ncl.ac.uk/paolomissier/2021/02/07/w3c-prov-some-interesting-extensions-to-the-core-standard/

[E6] *Email exchange* with Dr. Reid Sherman, USGCRP (following additional interview)

[E7] *Additional Interview notes & contact for corroboration*: Dr. Stephen Cresswell, The Gazette, UK.

[E8] *Additional Interview notes & contact for corroboration*: Dr. Tom Plasterer, Director of Bioinformatics, Data Science & AI, AstraZeneca.