Institution: King's College London

Unit of Assessment: Computer Science and Informatics (UoA 11)

Title of case study: Provenance - A fundamental data governance technique

Period when the underpinning research was undertaken: 2010 - 2020

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:
Miles, Simon	Reader	03/2007 - present
Curcin, Vasa	Senior Lecturer	07/2014 - present
Moreau, Luc	Professor	08/2017 - 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)

Provenance is a record of the processes by which data was produced, by whom, how, and from what other data. Research on provenance at King's over a decade, and later significant and sustained world-wide adoption, often without King's direct involvement, have led to the global recognition that provenance is a critical facet of good data governance for businesses, governments and organisations in general. The impact of King's pioneering work has manifested itself in i) commercial, governmental and research organisations launching new products incorporating provenance functionality, ii) multiple standardisation bodies providing guidance to software engineers, iii) scientific communities coalescing around provenance to ensure trusted information exchange, and iv) regulators asserting that provenance is a technique to address regulatory requirements. Overall, due to King's research, provenance is now widely regarded as an essential function of an IT system, to provide a trusted account of what the system performed and the data it manipulated.

2. Underpinning research (indicative maximum 500 words)

The World Wide Web Consortium (W3C), the standardisation body for the web, defined provenance as a record that describes the people, institutions, entities, and activities involved in producing, influencing, or delivering a piece of data or a thing. Back in 2013, W3C published PROV – a standard for expressing such a record to make it possible to store, exchange, query and manipulate it in interoperable ways.

King's 2014 impact case study described Miles's contribution to the requirements [1] underpinning provenance and the methodology required to put it in practice. Since 2014, new research has been conducted and novel significant impact has emerged. King's researchers have conducted world-leading research that builds on PROV to deepen the understanding of provenance, to develop software engineering methodologies and techniques to deploy provenance, and to conceive provenance-based techniques for systems to produce explanations about their decisions. This research has been led by Miles (2007-present), Curcin (2014-present) and Moreau (2017-present). Their teams have worked with a range of applications that share a critical data governance imperative of demonstrating the quality of data, including health management systems (UKRI grants EP/N027426/1 and KTP-509790), automated decision systems in finance (UKRI grants EP/R511559/1 and EP/S027238/1), and command and control systems (US Navy grant N629091812079). Through this body of research, and specifically the following three strands, provenance emerges as a fundamental technique for data governance.

[1. Understanding] In collaboration with leading researchers who specified PROV, Miles conducted a retrospective analysis that explicitly characterised the scope, requirements, guiding principles, and design decisions that resulted in PROV [2]. This post-standardisation analysis demonstrated the broad consensus about the design of PROV, helped increase the

understanding of PROV and its interoperability, and was a scientific output that has contributed to its widespread adoption.

[2. Creating and managing] A key practical challenge that hampers adoption of new technologies is the effort required to deploy them in practice; specifically, for provenance, a challenge was the effort involved in automatically creating provenance that accurately describes the actions performed by an application. With this concern in mind, research has taken place in two complementary approaches, both supporting the novel paradigm of declarative construction of provenance.

First, Curcin defined provenance templates [3] as *abstract provenance fragments* representing meaningful domain actions (UKRI grant EP/N027426/1). Templates were conceived to generate a model-driven service interface for domain software tools to routinely capture the provenance of their data and tasks. By exposing a domain-focused interface for provenance, Curcin demonstrated that provenance templates can capture the audit trail of a task and its resulting data. The use of provenance thereby enables users to place their trust in systems; in particular, it facilitates reproducible research, which was demonstrated by a range of provenance-based queries, in the context of Learning Health Systems. Curcin's provenance templates were subsequently used in an Innovate-UK project (KTP-509790), where the provenance templates technology was demonstrated to be deployable within legacy applications in order to enrich a product with novel data governance functionality.

Second, Moreau instigated UML2PROV [5], a technique capable of producing provenance automatically for programs specified according to the industry-adopted UML modelling language. This technique also relies on an extensive set of provenance fragments describing common programming language patterns. The significance of this technique is that it shows that high-level program specifications can be the source of automatic provenance generation, thus reducing the human effort involved in creating provenance, and hereby facilitating provenance adoption.

[3. Exploiting] Provenance is routinely regarded as a technique by which trust can be endowed to systems, by enabling the tracing of data that flow through them, so that it can subsequently be inspected by users. All the information, data dependencies and processes underpinning a decision are collectively the provenance of the decision. In grants EP/R511559/1-EP/S027238/1, in collaboration with the Information Commissioner's Office (ICO), Moreau and his team have developed an approach that exploits provenance to construct human-consumable explanations of decisions made by systems [4]. The King's team demonstrated that some GDPR-related questions pertaining to automated decisions can be answered from their provenance: the solution relies on semantic mark-ups embedded in the provenance, which are exploited by queries to extract the relevant elements that have influenced a decision; these elements are then used in declarative specifications of explanations realised by a natural language generation engine. In addition, Moreau and collaborators have developed a technique called "provenance analytics" [6], which demonstrates that Machine Learning techniques can be applied to analyse provenance in an automated manner so that "quality" or "trust" assurance can be derived without manual human inspection.

To sum-up, research at King's has consisted of *(i)* requirement engineering linking design decisions to their rationale, which help improve PROV understandability, *(ii)* declarative methods reducing human effort involved in producing provenance, and *(iii)* techniques to derive explanations and measures of data quality from provenance. This research has helped transition provenance, originally seen as a laboratory experimental concept, to a fundamental data governance technique, deployed in practical applications.

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

1. Groth, P., Gil, Y., Cheney, J., & **Miles, S.** (2012), Requirements for Provenance on the Web, Int. Journal of Digital Curation 7(1), 39-56. https://doi.org/10.2218/ijdc.v7i1.213

2. **Moreau, L.**, Groth, P., Cheney, J., Lebo, T., & **Miles, S**. (2015), The Rationale of PROV. Journal of Web Semantics. https://doi.org/10.1016/j.websem.2015.04.001

3. **Curcin, V.**, Fairweather, E., Danger, R. & Corrigan, D. (2017) Templates as a method for implementing data provenance in decision support systems, J. of Biomedical Informatics, vol. 65, pp. 1-21. https://doi.org/10.1016/j.jbi.2016.10.022

4. Huynh, TD., Stalla-Bourdillon, S. & **Moreau, L.** (2019), Provenance-based Explanations for Automated Decisions: Final IAA Project Report. (Accepted in ACM journal Digital Government: Research and Practice, Nov. 2020).

5. Sáenz Adán, C., Pérez Valle, B., García Izquierdo, F.J. & **Moreau, L.** (2020), Integrating Provenance Capture and UML with UML2PROV: Principles and Experience, IEEE Transactions on Software Engineering. https://doi.org/10.1109/TSE.2020.2977016

6. Huynh, D., Ebden, M., Fischer, J., Roberts, S & **Moreau, L.** (2018), Provenance Network Analytics: An approach to data analytics using data provenance, Data Mining and Knowledge Discovery. https://doi.org/10.1007/s10618-017-0549-3

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

Research on provenance at King's, along with the influence of standardisation of provenance at the World Wide Web Consortium (W3C) (Miles, PROV-DM), technology transfer to business (Curcin), toolkits and services to promote take-up (Moreau), influence of guidelines for data protection (Moreau), and, later significant world-wide adoption without King's direct involvement, have led to the global recognition that provenance is a critical facet of good data governance and an essential function of an IT system.

Given the openness of W3C standard PROV, it is impossible to track all the usages of PROV. Furthermore, as PROV is used in the background, generally, as part of a data management function, it is challenging to identify its impact in isolation to the rest of the functionality. Thus, below, we explore key strands of impact, including explicitly listing publicly documented usages of PROV or PROV concepts [G].

4.1. Impacts on public policy and services

Many governmental organisations have a duty to make data publicly available: open government data is regarded as creating value (worth billions of pounds world-wide) in many areas, including transparency, democracy, participation, innovation and efficiency. We focus on two illustrations of open government. In the UK, the Gazettes are the official journals of public record, whereas, in the US, the U.S. Global Change Research Program (USGCRP, a cooperation between 13 federal agencies) publishes a quadrennial National Climate Assessment. Both have independently adopted PROV as a mechanism to improve public navigation and access to information through the use of linked knowledge, thereby addressing an overarching aim of information transparency.

"*The credibility, trust, and integrity of the [Gazette] data has been strengthened because of PROV"* [D, p.11]. PROV *"works in the background to provide a clear, ethical and transparent data source"* [D, p.1], and *"ensures that the official public record is credible and accurate"* [D, p.2].

In the US, the use of PROV had an impact on the environment, as the *"policy debate on climate change has been influenced through the work of USGCRP: PROV-linked research meant that all information was meticulously documented which led to less controversy"* [D, p.16]. *"There have been a lot of scandals and difficult thinking in the scientific community about reproducibility and about accuracy and integrity, and […] PROV is a structure that can really support advances in general scientific practice to address those concerns"* (Climate Adaptation Lead, USGCRP) [D, p.5]. Overall, PROV helps increase trust in data and processes, and one of its tangible impacts is the reduction of FOI requests, *"as the accuracy of PROV renders many FOI requests pointless"* [D, p. 13].

4.2. Impacts on practitioners

4.2.1. Impact on standards and standardisation bodies

The impact on standards reported for the REF 2014 now goes well beyond the original standardisation of PROV at W3C, which was specified in 2013 as a domain-agnostic ontology for provenance, building on research at King's by Miles on the requirements for provenance. Since then, we can report secondary adoption of PROV, which has been referenced and key concepts directly imported in specifications published by other standardisation bodies (HL7, Allatrope Foundation, RDA, IVOA), reaching out to 500+ member organisations. All these specifications have an impact on a range of practitioners, including software engineers, working on scientific data management systems and health care systems. These examples of secondary standardisation build on King's research but without direct involvement from King's, which is evidence of broad and sustained adoption of provenance. Specifically, Health Level Seven (HL7) is the international body for healthcare standards (500+ corporate members), representing healthcare providers, government stakeholders, payers, and pharmaceutical companies. HL7's Fast Healthcare Interoperability Resources (FHIR) model is the standard for programmatic communication of health data, and it includes a mapping to the core concepts of PROV [E]. The Allatrope Foundation (a consortium of 40 pharmaceutical organisations) defined a universal data format that standardizes laboratory experimental parameters in order to remove human error and enhance scientific reproducibility; its Audit Trail ontology is building on PROV [G].

The Research Data Alliance (RDA, 50+ research organisations, 11,000 individual members) aims to build the social and technical bridges to enable open sharing and re-use of data. The RDA Recommendation [C, p10-11] incorporates PROV as metadata for the management of research objects in data centres to support reproducible research. The International Virtual Observatory Alliance (IVOA), comprising 20 Virtual Observatory (inter-)national programs, has published its recommendation ProvenanceDM applying PROV to astronomical data [H, p.13].

4.2.2. Impact on guidance for AI practitioners

The General Data Protection Regulation (GDPR) is the European wide regulatory framework that codifies some rights for data subjects (the users who have provided data in return for services) and obligations on data controllers (the organisations that are providing these services). A key challenge is that regulatory frameworks remain high-level and do not specify practical ways for organisations to become compliant; this is a particularly salient problem for companies adopting AI in their new products. The Information Commissioner Office (ICO), the data protection regulator in the UK, published guidance around *"Explaining decisions made with AI"*, explicitly referring to provenance: *"Such provenance information provides the foundations to generate explanations for an AI decision, as well as for making the processes that surround an AI decision model more transparent and accountable."* [F, p.59] The ICO report [F, p.59] includes a link (https://explain.openprovenance.org/loan/) to King's demonstrator for provenance-based explanations [4], which in collaboration with ICO was tailored to address seven key requirements of the GDPR. The software system, the first of this kind, provides a tangible artefact to which implementors can refer when implementing high-level guidance. The mark-ups and associated queries over provenance identified in the research were instrumental in delivering fit-for-purpose explanations and present a significant advance for AI practitioners. Appendix 4 [F, p.124] further refers to the King's report [4].

4.2.3. Widespread adoption of PROV

Two fundamental characteristics of PROV is the simplicity of its core and its design based on a wide consensus, promoting interoperability, as exposed by King's research [2] and provenance toolkits and services hosted at openprovenance.org. For these reasons, several communities have coalesced around PROV (or PROV concepts) with a view to facilitating the interoperable exchange of provenance. The key driver for these communities stems from their work with complex workflows, involving multiple stakeholders, typically each with their respective IT systems, across which flows of data and decisions need to be documented to ensure their auditability or reproducibility. A growing list of adaptors (in excess of 50, counted November 1, 2020) is maintained at [G]. For instance, the environmental science

community is adopting PROV across the world, as illustrated by a range of projects, in the USA (NASA, JPL, PNNL), Germany (DLR), Austria (EEA), and Australia (Geoscience Australia, CSIRO). There is adoption of FHIR provenance in the HealthCare community, including companies such as Astrazeneca, Smile CDR and Perspecta. In the Astronomy community, Applause, a collection of photographic plates with full provenance, is operationally deployed by the Leibniz-Institut für Astrophysik Potsdam (AIP). Finally, innovative commercial products are exploiting provenance capabilities, Surround Australia, Smile CDR, Perspecta, and Imosphere, which we specifically discuss in Section 4.3.

4.3. Impact on Imosphere Ltd's commerce and practice

Imosphere Ltd is a company of 50 employees commercialising solutions for healthcare organisations. Their flagship Atmolytics product is a tool that produces interactive reports from patient cohort data. Thanks to an Innovate UK grant (KTP-509790), Atmolytics incorporated elements of Curcin's provenance template technology, to enable a range of new functionality in the product, such as a full audit trail of data sets, versioning, analytics and reports providing transparency and increasing the users' trust in the data findings. Provenance templates were deployed to model key Atmolytics behaviours, such as patient cohort management and decision making, and ensure a faithful provenance record in a standardised PROV format. The provenance module differentiates Atmolytics from the competition, by its functionality to maintain the integrity of data, supporting informed and trusted decision-making [A].

The new, provenance-enabled version was launched in the summer of 2018, and has been installed at customer sites in the UK, Europe and USA. There are in excess of 4,000 users in the UK and USA, managing data of over 1,000,000 patients (by November 2020). Among them, 60 councils in the UK are benefiting of provenance functionality to manage their individual care budgets. In the US, the National Cancer Institute "City of Hope" tracks over 100,000,000 medical events, whereas the medical school Meharry Medical College, Tennessee, tracks 250,000 patient's data. For the end-users of the system the *"network graph view of provenance data is a far more natural way of visualising and querying historical relationships between patient cohorts and analytical tasks".* [B]

Overall, due to PROV, Imosphere has benefitted significantly: *"Using the W3C PROV standard and provenance templates for this task, saved us years in design and development time and ensured we are standard compliant for any further extensions, reducing time to market by approximately one year."* Furthermore, *"the introduction of data provenance capabilities in the software, has also improved the software engineering aspect of our data analytics portals, as it promotes good practice in reusing and documenting analytical components across reports, avoiding duplication".* (Imosphere CEO, [B])

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

A. YouTube Atmolytics video, July 2018

B. Testimonial from CEO of Imosphere Ltd.

C. Weigel, T., Plale, B., Parsons, M., Zhou, G., Luo, Y., Schwardmann, U., Quick, R., Hellström, M., Kurakawa, K. (2018). RDA Recommendation on PID Kernel Information (Version 1)

D. Impact Evaluation of PROV - a provenance standard published by the World Wide Web Consortium, by Impact Science, July 2020

E. HL7 FHIR version 4.0.1 (section 6.3.3), October 2019

F. Explaining decisions made with AI, Project Explain, ICO, May 2020

G. Adoption of provenance, page maintained by Luc Moreau, November 2020, *(password: REF2021-kcl)*

H. International Virtual Observatory Alliance Provenance Data Model, April 2020