Institution: King's College London (KCL)

Unit of Assessment: B11: Computer Science and Informatics

Title of case study: Data provenance standardisation [DPS]

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

KCL research played an essential role in the development of data provenance standards published by the World Wide Web Consortium (W3C) standards body for web technologies, which is responsible for HTTP, HTML, etc. The provenance of data concerns records of the processes by which data was produced, by whom, from what other data, and similar metadata. The standards directly **impact on practitioners and professional services** through adoption by commercial, governmental and other bodies, such as Oracle, IBM, and Nasa, in handling computational records of the provenance of data.

2. Underpinning research (indicative maximum 500 words)

The *provenance* of something is "how it came to be as it is": from whence it originated and the processes it has undergone. In the context of computer systems, it has been studied independently in many fields, including geographical information systems, library studies, software engineering (as an aspect of traceability) and bioinformatics. Knowing the provenance of data allows its users to better understand what it means and to judge its reliability. For example, understanding the procedure by which a scientific experiment was conducted, peers may better interpret its conclusions or know how to repeat it; or, by knowing who added facts to a webpage and on what basis, readers may assess what to rely on.

Over the past decade there has been strong interest in re-usable approaches to capturing, storing and processing provenance. This has been driven by research into, first, determining the provenance of database query results (the source of the retrieved data within the database itself and the reason that these particular results were selected), and, second, determining what occurred in a particular run of an automated workflow (the inputs, whether any step failed, intermediate data products).

The latter work led to interest in interoperability of provenance. For example, in executing a bioinformatics workflow, there could be calls to remote public databanks to search for the function of a protein, and to other local or remote services, to convert data formats, get user input, or aggregate data sets. A complete audit trail requires that each service invoked record some data about the processing it has done in such a way that these records could be combined into a coherent whole to later answer provenance questions.

To achieve this interoperable provenance in practice, a number of research developments were required, which Dr Simon Miles (KCL) led or co-led. Dr Miles also contributed extensively to other parts of the work of the standards team.

Requirements. To develop provenance-supporting systems that are of practical benefit, it is necessary to understand the breadth of questions users have about data provenance and in what contexts. Dr Miles was part of a team that collected and analysed use cases from a range of users concerned with provenance of web-based data [4], building on earlier work he led analysing provenance across scientific disciplines [2].

Model. The next critical element of interoperable provenance is a well-founded model that each element of a system can independently use to record its activities. The first model that gained true widespread and cross-application international use was the Open Provenance Model [A]. This describes what has occurred in a system as relations between the 'artifacts' (data and other entities), 'processes' (activities that have taken place), and 'agents' (responsible parties, such as users) involved. Later, the W3C instigated a working group to develop official provenance standards, PROV, based around a data model, PROV-DM. Dr Miles was an author of OPM and was an Invited Expert on the W3C group, including being a contributor to PROV-DM.

Methodology. A non-trivial step in ensuring that the provenance of data resulting from distributed processing can be retrieved is to create or adapt the systems to record provenance data. It is demanding to understand how existing applications should be augmented, what details to record, how to do so in a way that allows interoperability of provenance between system parts and how to be able to answer the provenance questions users will only consider once data is available. Effective engineering methodologies are required, and this is an area in which KCL has led the field internationally. At the design level, the Provenance Incorporation Methodology (PrIMe) allows designers to determine how to add provenance recording to their systems to meet user requirements [1]. At the modelling level, the W3C PROV Primer [3] guides developers in understanding how to apply the PROV-DM to applications. At the implementation level, the SourceSource system allows programmers to automatically introduce provenance recording into their code [6]. Each of these efforts has been led by Dr Miles.

Key Researcher: Simon Miles (KCL Lecturer 2007-)

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

***[1] S. Miles**, P. Groth, S. Munroe and L. Moreau. *PrIMe: A Methodology for Developing Provenance-Aware Applications*, ACM Transactions on Software Engineering and Methodology 20 (3), pp. 1-42, 2011. DOI: 10.1145/2000791.2000792
***[2] S. Miles**, P. Groth, M. Branco and L. Moreau. *The Requirements of Using Provenance in e-Science Experiments*, Journal of Grid Computing 5(1), pp. 1-25, 2007. DOI: 10.1007/s10723-006-9055-3
[3] Y. Gil and **S. Miles**. *W3C PROV Primer*. http://www.w3.org/TR/prov-primer/
***[4]** P. Groth, Y. Gil, J. Cheney, and **S. Miles**. *Requirements for Provenance on the Web*, International Journal of Digital Curation 7(1), 39-56, 2012. DOI: 10.2218/ijdc.v7i1.213
[5] Y. Gil, J. Cheney, P. Groth, O. Hartig, **S. Miles**, L. Moreau, P. Pinheiro da Silva. *Provenance XG Final Report*, 2010. Online publication at http://www.w3.org/2005/Incubator/prov/XGR-prov-20101214/
[6] S. Miles. *Automatically Adapting Source Code to Document Provenance*. Proceedings of the 3rd International Provenance and Annotation Workshop (IPAW 2010), Troy, US, June 2010, pp. 102-110, published by Springer. DOI: 10.1007/978-3-642-17819-1_13

*Publications indicating quality of underpinning research.

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

The W3C standard [A] is of direct benefit to practitioners, as can be seen from the corroborating evidence and the involvement in the standard's development by companies and NGOs. The standard and supporting research outputs, such as the methodology work led by KCL, allow organisations to adapt their systems in order to document and analyse their processes, or to incorporate such functionality into the software they produce for their customers. Working backwards from the impact to the underpinning research, the following summarises how the former was derived from the latter, and explains Dr Miles' involvement at each stage.

Impact: The W3C standard on provenance (PROV) [A, B], is being implemented, extended and included in applications by a range of commercial, governmental and other organisations with use cases regarding the provenance of their data eg:

(i) Oracle [C]: *"Until PROV, one of the hugest problems we faced was maintaining transaction audit trails in a heterogeneous environment in a standard and compatible way. Audit trails are described with literally millions of different formats in different organizations. This used to mean it was impossible to create a single audit time line. PROV solves this problem. We now provide (and consume) a PROV feed that unifies the audit trails generated by transactions across heterogeneous systems."*

(ii) NASA [D]: *"Earth Science Data Systems across NASA play a critical role in data processing* *and analysis of NASA datasets. However, there is a growing need to provide the provenance of these datasets as scientists increasingly need to assess the lineage of the data products to improve their understanding and trust of the science results. Lessons learned from Climategate show that there is public demand for more transparency and understanding in the science process. Science data systems are key to enabling the capture, management, and use of production provenance information.... The W3C Provenance Working Group ... standard is very general, intended to support the breadth of any domain. To better serve the needs of specific domain communities, the standard has several built in points of extensibility. This working group will participate in efforts to develop an Earth Science PROV Extension (PROV-ES)."*

(iii) IBM [E]: *"We don't know whether the information we find on the Web is accurate or not. The Dublin Core model describes a resource for the purpose of discovery. The W3C PROV model describes entities and processes involved in producing and delivering that resource."*

(iv) German Aerospace Centre [F], who use the PrIMe methodology [1], taking OPM [A] as their provenance model.

The primary evidence for this lies in the results of a survey conducted by the W3C group, during the final stage of the standardisation, examining implementations and applications of the PROV standard. There were already 66 such implementations at that time (April 2013). The implementers include large commercial organisations (e.g. Oracle), SMEs, and academic institutions from around the world. This survey was conducted before the specifications became recommendations (official standards), so captures only the pioneering implementers [G]. Dr Miles was an Invited Expert to the W3C Working Group on provenance. He is co-editor of one of the specifications produced, a primer [3] which provides the introductory steps for those wishing to understand and adopt PROV. He is also a contributor to many of the other specifications, including the core data model for representing provenance data, PROV-DM, on which the other specifications are grounded.

Connection to underpinning research: Standards are, by nature, a collaborative community effort, bringing together adequately mature state-of-the-art to create a published specification which the world can rely on to be stable. However, the standards are directly influenced and, in some cases, derived from research activities in which Dr Miles led, co-led or was a major contributor.

W3C Incubator Group. In order for a W3C working group to be established, the need for a standard and the maturity of the state-of-the-art must each be established. This is achieved through an 'incubator group'. Dr Miles was involved in the incubator group for provenance throughout its operation (2009-2010). In particular, he led the activity of collecting and curating use cases from a variety of applications and projects, so demonstrating the need for provenance and the scope of the requirements. These use cases are available online [H]and a related publication later drew together these requirements into some illustrative scenarios [4]. The research of the incubator group produced the definition of the standards the working group would create [5].

Open Provenance Model. Prior to and in parallel to the incubator group, an international community effort was underway to interconnect different approaches to provenance. This ultimately resulted in the Open Provenance Model [A], a widely used *de facto* standard until W3C PROV was developed. One piece of evidence for the popularity of OPM comes from its citations, 242 on Google Scholar Citations when last checked. The influence of OPM on PROV is direct: PROV takes the same core model and approach as OPM. Where OPM is founded on describing past processes in terms of 'artifacts', 'processes' and 'agents', PROV's data model core is the semantically almost identical 'entities', 'activities' and 'agents'. Dr Miles was a co-author of OPM.

Methodology. One of the key distinct contributions of KCL to the international efforts in provenance research lies in *methodology*, i.e. how to design or adapt distributed applications so that provenance data is recorded that will allow users' questions to later be answered. The W3C primer is the latest incarnation of this effort, but it is founded on earlier work. In particular, Dr Miles led the development of the first (and currently only) software engineering methodology for adapting existing applications to meet provenance requirements, PrIMe [1].

Provenance Challenges. OPM was itself the outcome of prior research activities: the Provenance Challenges. These were exercises in which research teams from around the world, both academic and commercial, applied their approaches to provenance to a single problem to allow comparison and establishment of a mutual understanding. The Second Provenance Challenge, held in late 2007, compared data models used in these approaches, and the concluding workshop for the challenge was the start of the development of OPM. Dr Miles co-chaired the workshop as well as co-devising the challenge exercise itself. The minutes of the challenge workshop, and participating teams are available [I], together with record of the resulting release of OPM [J].

Requirements. The requirements capture work of the incubator group and the methodology research described above drew on earlier work led by Dr Miles into provenance requirements, and techniques for their capture. In particular, a study was conducted to understand provenance requirements across scientific disciplines (e.g. bioinformatics, particle physics, chemistry, proteomics, medicine), which influenced much future work in the field of provenance [2].

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

The following document corroborates Miles' contribution to the OPM standard:

[A] L. Moreau, B. Clifford, J. Freire, J. Futrelle, Y. Gil, P. Groth, N. Kwasnikowska, **S. Miles**, P. Missier, J. Myers, Y. Simmhan, E. Stephan, and J. V. den Bussche. *The Open Provenance Model Core Specification (v1.1)*, Future Generation Computer Systems 27 (6), pp. 743-756, 2011. DOI: 10.1016/j.future.2010.07.005

The following materials corroborate the content and use of the standard. Website materials are available recording snapshots of content at time of submission.

[B] The W3C PROV standard (the link is to an overview with links to all the other documents). http://www.w3.org/TR/prov-overview/

[C] http://www.w3.org/QA/2013/05/interview_oracle_on_semantic_w.html

[D] NASA's work in applying PROV to earth science https://earthdata.nasa.gov/wiki/main/index.php/PROV-ES_Earth_Science_extension_to_W3C_PROV

[E] IBM blog on intended use of PROV and Dublin Core (a library metadata technology) https://www.ibm.com/developerworks/community/blogs/nlp/entry/november_29_2012_3_55_am9?lang=en

[F] H. Wendel, M. Kunde, A. Schreiber. *Provenance of software development processes*, in Proceedings of the Third International Provenance and Annotation Workshop (IPAW 2010), Troy, US, 2010. DOI: 10.1007/978-3-642-17819-1_7

[G] Report on survey of adoption: http://www.w3.org/TR/prov-implementations/.

[H] Report on Incubator Group use cases: http://www.w3.org/2005/Incubator/prov/wiki/Use_Cases.

[I] Minutes of Challenge Workshop http://twiki.ipaw.info/bin/view/Challenge/SecondWorkshopMinutes

[J] Release of OPM: http://twiki.ipaw.info/bin/view/Challenge/OPM.