The Open Provenance Model: An Overview

Luc Moreau[1], Juliana Freire[2], Joe Futrelle[3], Robert E. McGrath[3], Jim Myers[3], and Patrick Paulson[4]

[1] University of Southampton
[2] University of Utah
[3] NCSA
[4] PNNL

1 Background

Provenance is well understood in the context of art or digital libaries, where it respectively refers to the documented history of an art object, or the documentation of processes in a digital object's life cycle. Interest for provenance in the "e-science community" [12] is also growing, since provenance is perceived as a crucial component of workflow systems that can help scientists ensure reproducibility of their scientific analyses and processes [2,4].

Against this background, the *International Provenance and Annotation Workshop* (IPAW'06), held on May 3-5, 2006 in Chicago, involved some 50 participants interested in the issues of data provenance, process documentation, data derivation, and data annotation [7]. During a session on provenance standardization, a consensus began to emerge, whereby the provenance research community needed to understand better the capabilities of the different systems, the representations they used for provenance, their similarities, their differences, and the rationale that motivated their designs.

Hence, the first Provenance Challenge [1] was born, and from the outset, the challenge was set up to be *informative* rather than *competitive*. The first Provenance Challenge was set up in order to provide a forum for the community to understand the capabilities of different provenance systems and the expressiveness of their provenance representations. Participants simulated or ran a Functional Magnetic Resonance Imaging workflow, from which they implemented and executed a pre-identified set of "provenance queries". Sixteen teams responded to the challenge, and reported their experience in a journal special issue [9].

The first Provenance Challenge was followed by the second Provenance Challenge [1], aiming at establishing inter-operability of systems, by exchanging provenance information. During discussions, the thirteen teams that responded to the second challenge found out that there was substantial agreement on a core representation of provenance. As a result, following a workshop in August 2007, in Salt Lake City, a data model was crafted by the authors and released as the *Open Provenance Model* (OPM v1.00) [8].

On June 19th 2008, some twenty participants attended the first OPM workshop, held after IPAW'08 [3], to discuss the OPM specification. Minutes of the workshop and recommendations [5] were published, and led to the current version (v1.01) of the Open Provenance Model [10].

J. Freire, D. Koop, and L. Moreau (Eds.): IPAW 2008, LNCS 5272, pp. 323–326, 2008.

2 Scope

The *Open Provenance Model* (OPM) is a model for provenance that is designed to meet the following requirements:

- To allow provenance information to be exchanged between systems, by means of a compatibility layer based on a shared provenance model.
- To allow developers to build and share tools that operate on such provenance model.
- To define the model in a precise, technology-agnostic manner.
- To support a digital representation of provenance for any "thing", whether produced by computer systems or not.
- To define a core set of rules that identify the valid inferences that can be made on provenance graphs.

While specifying this model, we also have some *non*-requirements:

- It is not the purpose of OPM to specify the internal representations that systems have to adopt to store and manipulate provenance internally; systems remain free to adopt internal representations that are fit for their purpose.
- It is not the purpose of [8,10] to define a computer-parsable syntax for this model; model implementations in XML, RDF or others are being specified in separate documents.
- OPM does not specify protocols to store provenance information in provenance repositories.
- OPM does not specify protocols to query provenance repositories.

3 Technical Overview

The foundations of the Open Provenance Model can be traced back to the Second Provenance Challenge 'community agreement', summarized by Miles [6]. It is assumed that the provenance of objects (whether digital or not) can be represented by an annotated causality graph, which is a directed acyclic graph, enriched with annotations capturing further information pertaining to execution.

In OPM, provenance graphs consist of three types of nodes. *Artifacts* represent an immutable piece of state, which may have a physical embodiment in a physical object, or a digital representation in a computer system. *Processes* represent actions performed on or caused by artifacts, and resulting in new artifacts. *Agents* represent contextual entities acting as a catalyst of a process, enabling, facilitating, controlling, or affecting its execution.

Importantly, in OPM, a provenance graph is defined as *a record of a past execution* (or current execution); it is not a description of something that may happen in the future, nor a general recipe (workflow) that could be used to derive future data. OPM is a model of artifacts *in the past*, explaining how they *were* derived. Processes may be in the past, or can still be currently running. In no case is OPM intended to describe the state of future artifacts and the activities of future processes.

A provenance graph aims to capture the causal dependencies between the abovementioned entities. Therefore, nodes, whether artifacts, processes or agents, can be connected by directed edges that belong to one of the categories defined in the model. An edge represents a causal dependency, between its source, denoting the effect, and its destination, denoting the cause. Edges can express the following dependencies: an artifact was generated by a process; a process used an artifact; a process was controlled by an agent; an artifact was derived from another artifact; a process was triggered by another process.

A set theoretic model is proposed, and a set of inference rules are defined, allowing reasoning over causal dependencies. While the core model is timeless, it is permitted to annotate a provenance graph with time annotations, which themselves must satisfy constraints regarding causality.

4 Conclusion

The Open Provenance Model is work in progress, as indicated by the issues raised in the OPM Workshop [5]. We hope to capitalize on the community momentum, to keep on evolving the OPM specification into a well-founded data exchange format. It is proposed that OPM be used as a model for an inter-operability exercise, in a third Provenance Challenge. Serialisations are now being proposed for OPM, and libraries to manipulate provenance graphs are being implemented. All material related to OPM can be found from [11].

References

1. The Provenance Challenge Wiki (June 2006), `http://twiki.ipaw.info/bin/view/Challenge`
2. Davidson, S.B., Freire, J.: Provenance and scientific workflows: challenges and opportunities. In: SIGMOD Conference, pp. 1345–1350 (2008)
3. Freire, J., Koop, D., Moreau, L. (eds.): IPAW 2008. LNCS, vol. 5272. Springer, Heidelberg (2008)
4. Gil, Y., Deelman, E., Ellisman, M., Fahringer, T., Fox, G., Gannon, D., Goble, C., Livny, M., Moreau, L., Myers, J.: Examining the challenges of scientific workflows. IEEE Computer 40(12), 26–34 (2007)
5. Groth, P.: First OPM Workshop Minutes. In: Information Science Institute, USC (July 2008), `http://twiki.ipaw.info/bin/view/Challenge/FirstOPMWorkshopMinutes`
6. Miles, S.: Technical summary of the second provenance challenge workshop, King's College (July 2007), `http://twiki.ipaw.info/bin/view/Challenge/SecondWorkshopMinutes`
7. Moreau, L., Foster, I. (eds.): IPAW 2006. LNCS, vol. 4145. Springer, Heidelberg (2006)
8. Moreau, L., Freire, J., Futrelle, J., McGrath, R.E., Myers, J., Paulson, P.: The open provenance model (v1.00). Technical report, University of Southampton (December 2007), `http://eprints.ecs.soton.ac.uk/14979`

9. Moreau, L., Ludaescher, B. (eds.): Special Issue on the First Provenance Challenge, vol. 20. Wiley, Chichester (2007)
10. Moreau, L. (ed.), Plale, B., Miles, S., Goble, C., Missier, P., Barga, R., Simmhan, Y., Futrelle, J., McGrath, R., Myers, J., Paulson, P., Bowers, S., Ludaescher, B., Kwasnikowska, N., Van den Bussche, J., Ellkvist, T., Freire, J., Groth, P.: The open provenance model (v1.01). Technical report, University of Southampton (July 2008), http://eprints.ecs.soton.ac.uk/16148
11. The Open Provenance Web Site (August 2008), http://openprovenance.org
12. Simmhan, Y., Plale, B., Gannon, D.: A survey of data provenance in e-science. SIGMOD Record 34(3), 31–36 (2005)