IJCAT - International Journal of Computing and Technology, Volume 6, Issue 12, December 2019
ISSN (Online) : 2348-6090
www.IJCAT.org
Impact Factor: 0.835

# Open Provenance Model: A Systematic Review

**Shridevi Erayya Hombal**

Data Analytics, IIIT Bangalore
Bangalore, 560100, India

**Abstract -** Open Provenance Model resulting from a community effort to achieve inter-operability in the provenance challenge series. Recently the open provenance model (OPM) has been developed as a consensus exchange format for representing the provenance graphs. OPM is a directed acyclic graph; it is used to represent casual dependencies between the set of processes and products. In this paper the open provenance model has been defined and described through example.

***Keywords:*** *Provenance, Open Provenance Model (OPM), inter-operability*

## 1. Introduction

The terabytes of data and metadata can be generated by the workflows. One kind of metadata is provenance (also referred to as lineage) which tracks the steps from which the data is derived. For different purposes, the scientific domain uses different form of provenance.

Depending on the domain where it is applied, provenance can be described in various terms. Buneman et al [1], who describes the data provenance in the context of database systems, defines it as the origins of data and the process by which it arrived at the database. Lanter [2], who describes the derived data products in GIS, characterizes the provenance as information describing the materials and transformations applied to derive the data. Provenance not only associated with just the data products, but also with the process (es) that enabled their creation as well. Lanter's definition was expanded by Greenwood et al [3] and views it as metadata recording the annotations, notes about the experiments and recording the process of experiment workflows.

Against this background, the International Provenance and Annotation Workshop [4, 5], includes set of participants who had their queries on the provenance. Along with the queries the provenance research community needed to understand different provenance representations, capabilities of different provenance systems, process documentation, data annotation [6], data derivation and issues of data provenance.

Hence, in order to provide a forum for the community to understand the capabilities of different provenance systems and expressiveness of their provenance representations, the first and second provenance challenges were set up. There were several discussions on a core representation of provenance; the Open Provenance Model (OPM) [7] was put forward as a data model by which systems can exchange provenance information. Such agreed model is being the focus of third provenance challenge [8], where in which the aim is to evaluate the efficiency of Open Provenance Model in representing and exchanging the provenance information in the provenance system.

The key structure defined in Open Provenance Model is an OPM graph, a directed acyclic graph aimed at representing data and control dependencies of past computation. Furthermore, OPM introduces the concept of an account, a description of a past execution. In a same graph multiple accounts are allowed to co-exist, hereby representing different explanations or observations at different levels of abstraction of a same execution.

Open Provenance Model is a model that is designed to allow provenance information to be exchanged between systems, to allow developers to build and share tools that operate on such provenance model, to support a digital representation of provenance, to define core set of rules that identify the valid inferences that can be made on provenance representation**.**

## 2. Related Work

The "Open Provenance Model" (OPM) [7] became the first broadly followed specification for storing and modeling provenance. Building on this effort, a provenance working group [9] was formed by the W3C standard body which established a standard data model

for provenance information called PROV. PROV represents provenance as a directed acyclic graph, or DAG. In the PROV model [10] vertices may correspond to entities, activities or agents. The edges connecting these items are annoted with the possible relationships between the entity, agent or activity. The following Figure 1 gives the summary of entities and relationships in the PROV model. The term entity refers to activity or agent.

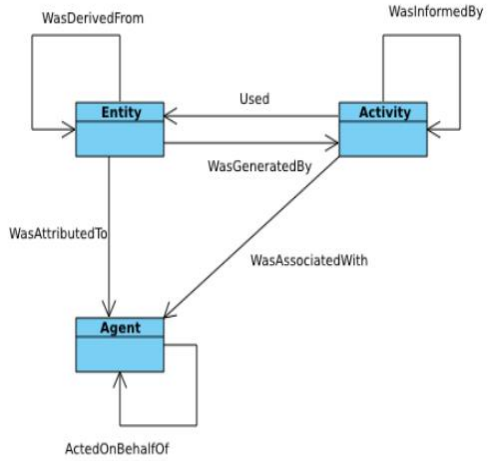


Fig 1: Summary of the core node and relationship types in the PROV model

## 3. Basics

Open Provenance Model describes about how things depended on others and results to specific set of states. It is a generic data model for provenance that allows domain and application specific representations of provenance to be translated into such a data model and interchanged between systems. Thus, heterogeneous systems can export their native provenance into such a core data model, and applications that need to make sense of provenance can then import it, process it, and reason over it. It consists of directed acyclic graph which expresses such dependencies. The listed below are basics of such a graph.

### 3.1 Entities

It is recommended to represent "things", it may be physical objects such as car, digital data such as simulation results. Many things can be stateful. For example car may contain different passengers, it can have tank full or empty and it may be at various locations. With respect to provenance, a new concepts such as artifact has introduced, which is an immutable piece of state. Likewise another concept has introduced such as process as actions resulting in new artifacts. Agents are like catalyst of process taking place.

The Open Provenance Model is based on three primary entities, which are:

- **Artifact:** Immutable piece of state, which may have digital representation in a computer system or may have a physical embodiment in a physical object.
- **Process**: Actions or series of actions performed on or caused by artifacts and which results to new artifacts.
- **Agent:** Entity acts as a catalyst of a process, enabling, controlling, facilitating, and affecting its execution.

Open Provenance Model is a model of processes which occurred in the past, which means they have already completed their execution, or is a model of artifacts in the past, explains how they were derived. Hence OPM never describes about the activities of future of processes or state of future artifacts.

In the Provenance graph artifacts are represented by circles, which represent the elements of the set Artifact. Likewise, processes are represented as rectangles and denoted as elements of the set Process. Finally, the agents are represented by octagons and denoted as elements of the set Agent

### 3.2 Dependencies

Provenance graph describes the casual dependencies between the entities. Provenance graph is a directed acyclic graph, in which nodes are entities, processes and agents and edges belongs to one of the following categories as shown in Figure 2, An edge representing the dependency between its source, denoting the effect and its destination, denoting the cause.

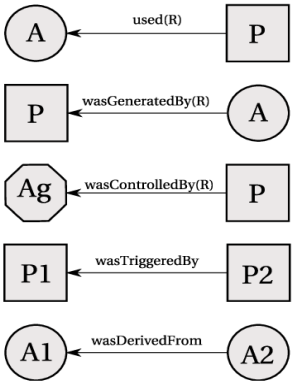


Fig. 2: Edges in the Open Provenance Model

The first two edges describes that a process used an artifact to finish its execution and that an artifact was generated by a process, it means a process is required to generate an artifact. It is important to know under which role an artifact was used by a process to accomplish its task since a process may have used several artifacts. Roles are basically used to distinguish the inputs and outputs. The Roles are defined by application domains and are used to distinguish the involvement of artifacts in processes.

The third edge describes about the agent which acts as a catalyst or controller to a process. This dependency is represented by wasControlledBy edge. A process may have been controlled by several agents their roles as controllers.

According to the fourth edge it is also recognized that it might be unknown about the exact artifact that a process P2 used, but that there was some artifact generated by another process P1. P2 is then said to have been wasTriggeredBy P1.

The fifth edge describes about the data flow oriented view of provenance. A2 was generated by process which used some artifacts, this dependency of which artifact has been used by process to generate A2. Hence to make this dependency explicit, it is required to maintain that artifact A2 wasDerivedFrom another artifact A1.

- Relationship: A relationship is represented by an arc. It denotes the dependency between source of the arc (the effect) and the destination of the arc (the cause). There are five relationships has been recognized: a process used an artifact, an artifact was generated by a process, a process was triggered by a process, an artifact was derived from an artifact, and a process was controlled by an agent.

The task of existence of an entity can be demonstrated by the group of entities. The factors which influences to adopt weaker notion of dependency for OPM

- Expressibility: Without the knowledge of exact internal data and control dependencies, systems will produce descriptions of what their components are doing. In order to use OPM in practice, weak notions of dependency are necessary.
- Composability: In a system consisting of multiple sub components, the high level summary of the system requires a weaker notion of dependency than the low level descriptions of its subcomponents.

The following dependencies have been adopted in OPM.

- Artifact Used by a Process: Process is connected to an artifact by an edge “used” is to indicate that the process requires artifact to complete its execution. Multiple artifacts may also be connected to the same process, all of them required for the process to finish its execution.
- Artifacts Generated by Processes: In a graph an edge “wasGeneratedBy” is used to connect an artifact to a process is intended to mean that process was required to generate an artifact. When several artifacts are linked to a same process by multiple wasGeneratedBy edges, the technique had to have begun, for all of them to be generated
- Process Triggered by Process: An edge “wasTriggeredBy” from P2 to P1 indicates that p1 was required for p2 to be able to complete.
- Artifact Derived from Artifact: An edge “wasDerived From“from A2 to A1 indicates that artifact A1 may have been used by a process that derived A2.
- Process Controlled by Agent: An edge “wasC ontrolledBy” between a process P and an agent Ag indicates that a start and end of process P was controlled by agent Ag.

- Roles

Role describes an agent’s or artifact’s function in a process. The process may be generatedby more than one artifact. Each artifact has unique roles for each of the process. For example, a process may use several files, reading data from another, reading parameters from another.

- Alternate Descriptions

The following Figure 3 describes about how the pair (3,7) can be generated. According to the left hand graph, the pair was generated by a process that added one to all constituents of the pair (2,6). According to the right hand graph, the derivation of (3,7) has done by adding one to 2 and 6,

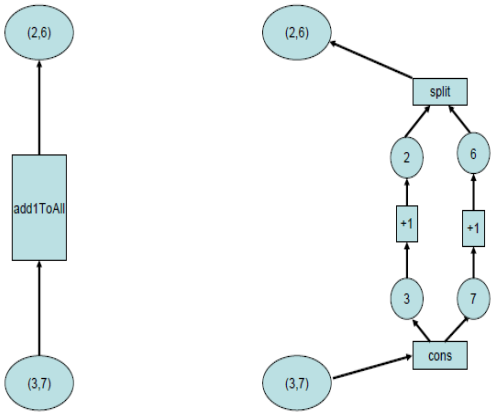


Fig 3:Example of Provenance Graph

These two graphs refer to the same pairs, that is how (3,7) can be derived from (2,6). These two graphs explains about the two different descriptions about the same derivation of (3,7).

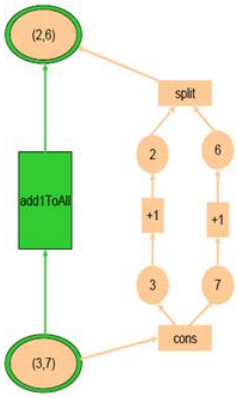


Fig 4: Example of Integration of Two Provenance Graph

The above Figure 4 describes about the integration of two provenance graph, by selecting different colors for nodes and edges. The darker green belongs to left graph of Figure 3, whereas the lighter orange part is the alternate description of right graph of Figure 3. The darker and lighter sub graphs describes about the past execution, offering different levels of explanation for such execution. The artifacts 3 and 7 were required for the process cons to take place.

## 4. Provenance Graph Definition

According to the following rules the provenance graph has been defined.

- Artifacts: These are identified by unique identifiers. Two artifacts are said be equal if they are having same identifier. Artifacts can optionally belong to accounts: account membership is declared by listing the accounts an artifact belongs to.
- Accounts: These are entities which can be compared.
- Processes: Processes are identified by unique identifiers. Two processes are identified as equal if they have the same identifier. Processes can optionally belong to accounts: account membership is declared by listing the accounts a process belongs to.
- Agents: Agents are identified by unique identifiers. Agents can optionally belong to accounts: account membership is declared by listing the accounts an agent belongs to.
- Edges: Edges are identified by source, destination and role. According to Figure 2 source and destinations are identified by identifiers for artifacts, processes or agents. Edges can optionally belong to accounts: account membership is declared by listing the accounts an edge belongs to.
- Roles: For edges like used, wasControlledBy, wasGeneratedBy roles are mandatory.
- Edges ensure connection between actual causes and effects, the model assumes that if an edge belongs to account, then its source and destination also belongs to this account.
- OPM graph may contain artifacts, agents, processes, and accounts. OPM graphs may be an empty set. A singleton containing an artifact, agent or process is an OPM graph. The intersection of two OPM graphs is an OPM graph and The intersection of two OPM graphs is also an OPM graph
- Account view can be defined as a view of an OPM graph according to one account, consists of elements whose account membership for artifacts, agents and processes and for edges contain the account.
- An account view is legal if it free of cycle of "was derived from" edges and if it contains at most one was generated by edge per artifact.
- If all account views are legal then OPM graph is a legal OPM graph.
- Legal account views are OPM graphs. The union of two legal account views is an OPM graph. The intersection of two legal account views is an OPM graph.
- A legal provenance graph might not contain time information.

- Edges can optionally describe with time information.
- If the two account view is having some agent, process and artifact in common then it said to be that two account views are overlapping.
- If the set of multi-step dependencies that can be inferred in v1 after application of completion rules is a superset of multi-steps dependencies that can be inferred in v2 after application of completion rules then it is said to be that account view v1 is a refinement of another account view v2.
- OPM graph relations between accounts can be asserted. If two accounts asserted to be in relationship satisfy above relationship definition then account relation assertions are legal.

## 5. Conclusion

This paper has introduced the Open Provenance Model, consisting of graphical notation and technology-independent specification, to represent past executions. OPM specification includes vast amount of potential activity. OPM is the focus of third challenge from a practical viewpoint, where set of provenance queries will help to evaluate its efficiency.

**Acknowledgment**
I will forever be grateful to my parents for their support and encouragement.

## References

[1] P. Buneman, S. Khanna, and W. C. Tan, "Why and Where: A Characterization of Data Provenance," in *ICDT*, 2001.

[2] D. P. Lanter, "Design of a Lineage-Based Meta-Data Base for GIS," in Cartography and Geographic Information Systems, vol. 18, 1991.

[3] M. Greenwood, C. Goble, R. Stevens, J. Zhao, M. Addis, D. Marvin, L. Moreau, and T. Oinn, "Provenance of e-Science Experiments - experience from Bioinformatics," in Proceedings of the UK OST e-Science 2nd AHM, 2003.

[4] Luc Moreau and Ian Foster, editors. Provenance and Annotation of Data —International Provenance and Annotation Workshop, IPAW 2006, volume4145 of Lecture Notes in Computer Science. Springer-Verlag, May 2006.

[5] Raj Bose, Ian Foster, and Luc Moreau. Report on the International Prove-nance and Annotation Workshop (IPAW06). Sigmod Records, 35(3):51–53,September 2006.

[6] R. Bose, I. Foster, L. Moreau, Report on the International Provenance and Annotation Workshop (IPAW06), Sigmod Records 35 (3) Workshop (IPAW06), Sigmod Records 35 (3) (2006) 51{53, ISSN 0163-5808, doi:http://doi.acm.org/10.1145/1168092.1168102, URL http://www.sigmod.org/sigmod/record/issues/0609/sigmod-record.september2006.pdf.

[7] L. Moreau, J. Freire, J. Futrelle, R. E. McGrath, J. Myers, P. Paulson, The Open Provenance Model (v1.00), Tech. Rep., University of Southampton, URL http://eprints.ecs.soton.ac.uk/14979/1/opm.pdf, 2007.

[8] The third provenance challenge. http://twiki.ipaw.info/bin/view/Challenge/ThirdProvenanceChall February 2009

[9] W.W.W Consortium, “Provenance working group”, http://www.w3.org/2011/prov/wiki/Main_Page.

[10] PROV-DM: The PROV Data Model https://www.w3.org/TR/2013/REC-prov-dm-20130430/