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

# Examining the Challenges of Provenance

**Shridevi Erayya Hombal**

Data Analytics, IIIT Bangalore
Bangalore, 560100, India

**Abstract -** In Scientific workflows provenance is a critical concept, since it allows the scientists to understand the origin of the results, to repeat their experiments, to validate set of processes that were used to derive the data products. During a discussion on provenance standardization at the International Provenance and Annotation Workshop (IPAW'06, www.ipaw.info), the community decided that it needs to understand the different representations used for provenance, its common aspects, and the reasons for its differences. As a result, the community agreed that a "Provenance Challenge" should be set to compare and understand existing approaches. This paper describes about the challenges of provenance.

***Keywords -*** *Provenance, Workflows*

## 1. Introduction

Provenance refers to basic documentation of processes in digital object's life cycle [1] or documented history of an art object. Provenance is perceived as crucial component in the workflow systems [2] which basically help the scientists to ensure the reproducibility of scientific processes. In an open and inclusive environment such as the Web, where users find information that is often contradictory or questionable, provenance can help those users to make trust judgments. Against this background, the International Provenance and Annotation Workshop(IPAW 2006) [3,4], includes provenance related queries participants, whereby the provenance research community needed to understand better the capabilities of the different provenance systems and their representations, issues of data provenance, process documentation, data annotation[5,6] and data derivation. Open Provenance Model is a model of provenance, it basically used to allow developers to build and share tools, to define set of rules on provenance representation, to allow provenance information to be exchanged between systems, to support a digital representation of provenance.

Following these discussions the first challenge was born, and the challenge was set up to be informative rather than competitive. The first challenge was aimed to provide the community to understand capabilities of different provenance systems and their representations. The first provenance challenge was followed by second provenance challenge, aiming to provide inter-operability of systems, by exchanging the provenance information. Thirteen teams [7] responded to this second challenge. The consensus that followed led to a proposal for the Open Provenance Model (OPM)(v1.00) [8], a data model for provenance. The second provenance challenge was followed by third provenance challenge, aimed at evaluating the efficiency of Open Provenance Model in representing and exchanging the provenance information in the provenance system. Following the success of these challenges, for OPM an open-source governance approach was adopted, which led to revision of OPM v1.1

## 2. Motivation

There are three considerations which motivates the launch of novel challenge:

- So far, the Provenance Challenge activity has had a strong recognition on scientific workflows. In order to keep the involvement of scientific workflow community, it is needed to illustrate the wider applicability of provenance technology. For instance, it would be suited to do not forget eventualities that contain users, wherein computations take region on the desktop and inside the cloud, wherein various kinds of artifacts are manipulated, example records sets, files, documents, databases, and wherein artifacts are posted and downloaded from the Web.
- If provenance has not used then there is no point of capturing the provenance. Hence it is needed to capture the provenance, to demonstrate the functionality that would have been impossible to implement without provenance.

- Broader scenarios wherein provenance is captured, and higher exploitation of provenance to demonstrate functionality make use converge in the direction of an end to end scenario, in which many technologies are involved, and absolutely justifies the need for an interoperable solution

## 3. The Provenance Challenges

There are four provenance challenges which have been described below.

### 3.1 The First Provenance Challenge

To achieve the goal of understanding capabilities of different provenance systems and their representations the following points to be examined:

- The capabilities of different provenance systems which needed to answer for many provenance related queries.
- The representations of different provenance systems that shows the documentation of different processes which has occurred.
- Consideration of each system should be within the scope of the provenance.

To form the basis of the challenge a simple workflow was defined. The workflow has built based on the real experiment [9] which is in the area of Functional Magnetic Resonance Imaging (fMRI) and answering a set of queries over the provenance derived. Here, the workflow [10] is used to denote a series of procedures being performed in a system, each taking some data as input and producing other data as output. Instead of restricting to any particular technology (e.g., Web services, compiled executable, batch file, EXE files) the participants can use only one technology to implement the procedures and workflows. The main focus is on the provenance but not on running the experiment. All participants can execute the workflows after installing the necessary libraries. Different representations are used by different systems for provenance information. In order to verify the representation of different provenance systems, the challenging team has defined set of queries and asked participants to show how they addressed those queries.

All participants are allowed to upload the following information to the provenance challenge TWiki [11]

- For the example of workflow, representations of provenance.
- Representations of the workflow in their system.
- Representations for the results of core queries.
- Contributions of queries vs. systems
  - The query can be answered by the system.
  - The system cannot answer the query now but considers it relevant.
  - The query is not considered relevant to the project.

#### 3.1.1 The Functional Magnetic Resonance Imaging Workflow

The FMRI workflow is the provenance challenge workflow. It comprises data items and procedures flowing between them, respectively shown as rectangles and ovals in Fig 1. The workflow consists of five stages; each stage gives the description of the workflow. In addition to the data items, there are other inputs to procedures, details of which can be found on the challenge TWiki [11]. The input to the workflow is a set of brain images (Anatomy Image 1 to 4) and single reference brain image (Reference Image). All images are related to a brain of varying resolutions, each different feature. Each image consists of real image and metadata information for that image (Anatomy Header 1 to 4).

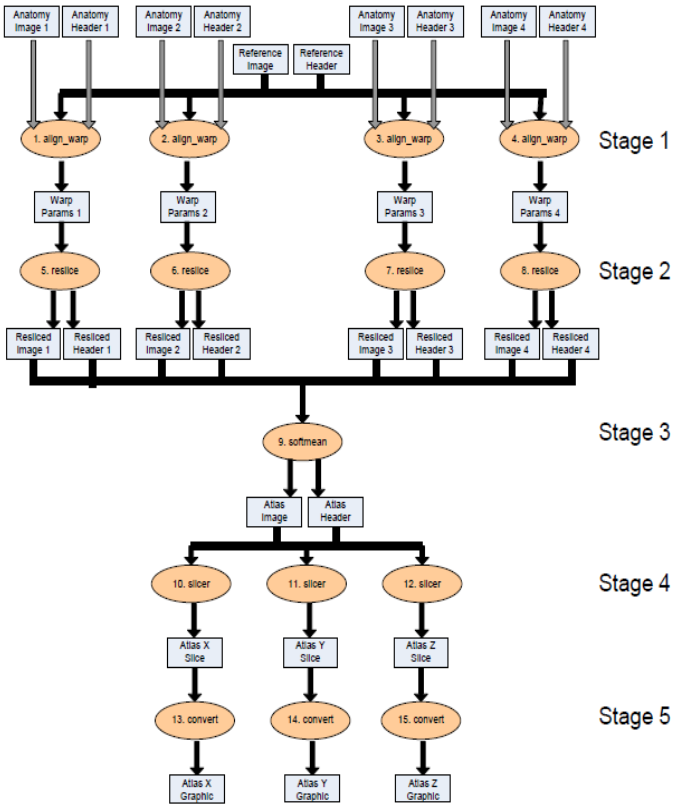


Fig 1: The workflow of Provenance Challenge

Workflow stages are described as follows:

1) align_warp compares reference image with the new brain image, to determine the new brain image shape and position, to match with the reference brain image. At the stage 1 the output of each procedure is a warp parameter set which defines the spatial transformation to be performed.
2) reslice is used to perform the actual transformation of the brain image for each warp parameter set which creates the new version of the original brain image. The output of stage 2 is the replica's image.
3) Using softmean all the replica's of images are averaged into one single image.
4) With slicer, average image is sliced, for each dimension(x,y and z) to give an atlas data set.
5) Using convert each atlas data set is converted into graphical atlas image.

### 3.1.2 Contributions of the First Provenance Challenge

There were 17 teams responded to the challenge and submitted an entry to the challenge TWiki [11]. The preceding sections describes about the broad characteristics of provenance systems, such as technologies they used and the environment in which they are embedded. The main purpose of the provenance systems is to build a computer based representations of provenance which can be queried, the next section is about describe to such representations.

- **Characteristics of Provenance Systems**

The taxonomy establishes six dimensions for comparing provenance systems,

1) Representation Technology: Provenance is stored and represented using a range of technologies such as semantic web technologies (RDF, OWL), relational databases (RDBMS), and internal private formats. Provenance can be represented in XML view by several systems.
2) Execution Environment: In specific execution environment only the provenance systems are embedded. The most common environments the provenance systems are embedded are operating systems and workflow systems.
3) General Aspects: For provenance systems which describes the general background
4) Data Capture: Which describes the way in which provenance data that can be captured on the existing provenance systems
5) Data Access: Which refers to way in which the user can access the provenance data repositories
6) Subject: This refers to the levels of detail in which the provenance can be represented.
7) Storage: Which describes the approaches used by provenance systems in order to register the provenance information
8) Non-Functional requirements: This refers to the non-functional requirements of provenance systems, such as security.

- **Properties of Provenance Representation**

Provenance system basically captures the casual graph, in order to produce the data product during execution. The following gives the criteria which basically extract some of the fundamental concepts underpinning provenance representations.

1) Naming: To query about the provenance of data products, a name can be used. The names are used to identify data products, some systems require each product to be identified by a unique name, which are created during workflow execution.
2) Time: Most systems prefer a notion of time, so that users can refer to data products or executions according to the time they were produced or took place.
3) Workflow Representations: Important part of the provenance representation is workflow representation. Some systems assume that an explicit representation of a workflow, whereas others do not have such an assumption and hence they depend on other means to describe about the executions.
4) Data Derivation: Some systems describe derivation of data, whereas others describes causal flow of events and some other are capable of characterising both event and data oriented views.
5) Tracked Data: Systems are capable of tracking the provenance of different kinds of data. Some introduce restrictions on the granularity of data they can track the provenance of. For instance, Systems may or may not deal with collections, bytes, files or bits.
6) Abstraction mechanisms: It is useful to describe with the different levels of abstractions when data products or processes are complex. Some

provenance systems provide new concepts or mechanisms in the provenance representations.

### 3.2 The Second Provenance Challenge

The first provenance challenge has led to valuable discussions about the aspects of provenance which were fundamental to all queries; all approaches and the expected results were interpreted differently by different groups. Therefore, there was no systematic way to compare the capabilities of different provenance systems, including the representations of provenance information. It was decided to introduce the second provenance challenge based on the first. With the first challenge, understanding interoperability of systems, by exchanging provenance information becomes a key issue, so this will be the second provenance challenge.

There were thirteen teams [7] responded to this second challenge. Several discussions happened related to the representation of provenance. As a result, in August 2007, a data model was crafted and released as the Open Provenance Model [8]. The provenance of objects is represented by directed acyclic graph, enriched with annotations capturing the information pertaining to execution. Provenance graph is defined as a record of past execution, but not the description of something which could happen in the future.

One way to solve the second provenance challenge is to compose the workflow execution systems, where each workflow system is to execute the part of the workflow and run the provenance queries over the results. The second provenance challenge basically challenges to the workflow systems instead of the approaches to the provenance. The value of provenance comes by tracking the provenance information through the workflow systems. As a result the teams shared provenance data produced by different provenance systems and perform the provenance related queries over the provenance data from other teams as the data has been produced by their own system.

According to the above approach, the second provenance challenge supports the systematic conversions of data between different provenance systems. The main goal is to understand how the provenance of data can be traced across multiple systems.

The challenge is divided into two phases. Each team should create TWiki page, once the challenge is complete, the team should make provenance data, queries and translation programs are available on their website. The first phase allows the teams to run provenance data over the set of workflows in order to answer for the queries. The second challenge is divided into three parts and which is based on the same workflow as the first, which is in the area of Functional Magnetic Resonance Imaging (fMRI).

- Part 1: align_warp and reslice (stages 1 and 2)
- Part 2: softmean(stage 3)
- Part 3: slicer and convert(stages 4 and 5)

There were three different sets of provenance data uploaded to the TWiki as each part is considered a separate workflow with regards to provenance data.

The second phase allows all the teams to use their own approach to combine the provenance data produced by different provenance systems and their own approach to query over it. Each team should download the data for each of the workflow parts. Teams should perform queries, must perform the query operations over the other team's data to have completed the challenge.

- Workflow Parts:

The challenge is based on the workflow definitions which has introduced in the first provenance challenge. In this challenge the workflow definitions is divided into three parts, which is shown in the following figure. The workflow consists of set of procedures, instead of focusing on which environment the workflow runs, the challenge is on the provenance; Hence define only the essentials of the workflow: the types of procedures performed and where the output of one procedure becomes the input for another, the roles of the provenance data in the workflow.

- Part 1: It includes two stages which are align_warp and reslice. It performs the reslicing of images into one referenced new image. This has been demonstrated in below Fig 2

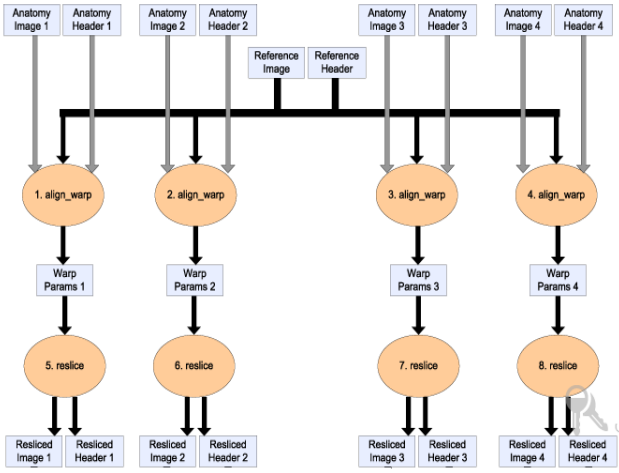

Fig 2: Workflow with Stages 1 and 2

- Part 2: It includes one stage which is Softmean.

It describes the averaging of brain images into one. This has been shown in below Fig 3

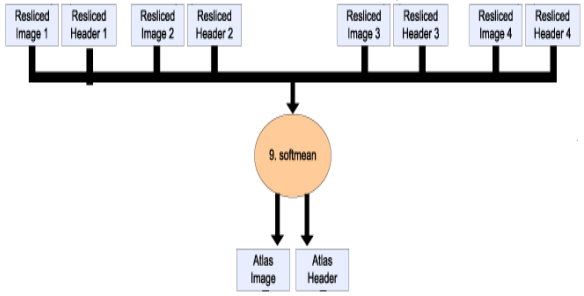

Fig 3: Workflow with Stage 3

- Part 3: It includes two stages which are slicer and convert. This part describes the conversion of averaged image into three graphics files showing the slices of that brain. This has been shown in below Fig 4

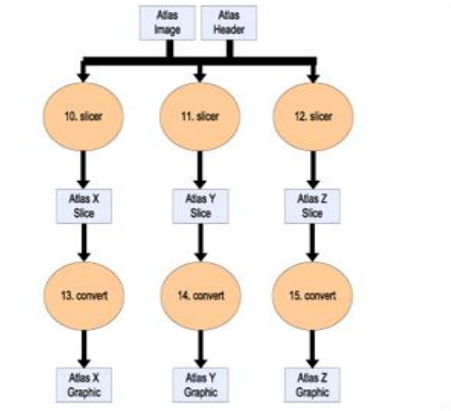

Fig 4: Workflow with Stages 4 and 5

### 3.3 The Third Provenance Challenge

First OPM workshop was attended by twenty participants to discuss the OPM specification v1.00 in June 2008. The Open Provenance Model [8] was actively used during the third provenance challenge. The third provenance challenge, aimed at evaluating the efficiency of Open Provenance Model in representing and exchanging the provenance information in the provenance system and answering the provenance queries.

The third provenance challenge was participated by 15 teams, A series of proposals were put forward, publicly reviewed, and put to vote; the result of participation was the version 1.1 of the Open Provenance Model.

### 3.4 The Fourth and Last Provenance Challenge

The main purpose of fourth provenance challenge is to apply the Open Provenance Model (OPM) for end to end scenario, and demonstrate the novel approach that can only be executed by the presence of an interoperable solution for provenance. The fourth challenge is the ultimate challenge, since it exploits OPM in an end-to-end scenario.

## 4. Conclusion

This paper introduced the provenance challenges and judges that the provenance challenges are highly successful, as measured with the number of participating teams, the quality of their submissions, discussions that resulted during the workshop. Whilst inter-operability is a pragmatic consideration, it entails fundamental studies questions. The fourth challenge remains a research activity, and the main purpose is to disseminate results.

### Acknowledgments

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

## References

[1] P. W. Group, Data Dictionary for Preservation Metadata| Final Report of the PREMIS Working Group, Tech. Rep., Preservation Metadata: Implementation Strategies (PREMIS), URL http://www.oclc.org/research/projects/ pmwg/premis-final.pdf, 2005.

[2] Ewa Deelman and Yolanda Gil (Eds.). Workshop on the challenges of scientific workflows. Technical report, Information Sciences Institute, University of Southern California, May 2006.

[3] 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.

[4] 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.

[5] L. Moreau, I. Foster (Eds.), Provenance and Annotation of Data | International Provenance and Annotation Workshop, IPAW 2006, vol. 4145 of Lecture Notes in Computer Science, Springer-Verlag, ISBN 3-540-46302- X, URL http://www.springer.com/uk/home/generic/search/ results?SGWID=3-40109-22-173681711-0, 2006.

[6] R. Bose, I. Foster, L. Moreau, Report on the International Provenance and Annotation 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] Second:Challenge, Second Challenge Team Contributions, URL http://twiki.ipaw.info/bin/view/Challenge/ ParticipatingTeams, 2007.

[8] 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.

[9] Zhao Y, Dobson J, Foster I, Moreau L, Wilde M. A notation and system for expressing and executing cleanly typed workflows on messy scientific data. SIGMOD Record 2005; 34( 3): 37– 43.

[10] Fox GC, Gannon D. Special issue: Workflow in grid systems. Concurrency and Computation: Practice & Experience 2006; 18(10): 1009– 1019. http://dx.doi.org/10.1002/cpe.1019.

[11] http://twiki.ipaw.info/bin/view/Challenge/FirstProvenanceChallenge [June 2006].