Applications Made Provenance-Aware with PASOA

www.pasoa.org

Paul Groth, Simon Miles, Miguel Branco, Luc Moreau,
School of Electronics and Computer Science
University of Southampton

Ian Wootten, Shrija Rajbhandari, Omer Rana, David Walker
School of Computer Science
Cardiff University

About PASOA

Definition of Provenance

Merriam-Webster Online dictionary defines provenance as:
(i) the origin, source
(ii) the history of ownership of a valued object or work of art or literature.
Our Definition:
The provenance of a piece of data is the process that led to the data.

Goal

The Provenance Aware Service Oriented Architecture project aims to investigate the concept of provenance and its use for reasoning about the quality and accuracy of data and services in the context of e-Science.

Aims

• To define provenance in relation to workflow enactment.
• To conceive algorithms to reason about provenance, in order to help scientists to achieve better utilisation of Grid resources for their specific tasks.
• To design a distributed cooperation protocol to record documentation of process in Grids.
• To investigate value-added properties that can be deduced from provenance.
• To engineer a proof of concept software architecture to support recording and reasoning over provenance in Grid environments

Contributions

PASOA has developed techniques, use cases, and software that enable developers to create provenance-aware applications.

Techniques

We have developed techniques to:

• map applications to a uniform, general provenance architecture

• wrap software components in order to automatically record documentation of process

• record documentation of process using our protocol PReP

• delimit recorded information in order to distinguish separate experiments

• query provenance using an application independent language

Use Cases

The following domains have provided PASOA with provenance related use cases. These use cases may inform other projects about how provenance might be useful to them. If your project has a use case, we would be interested in hearing from you.

Bioinformatics, High-Energy Physics, Proteomics, Computer Security, Chemistry, Fault Tolerance for Grids, Medicine.

Software

We have developed a suite of software for recording documentation of process compatible with the PReP protocol. Provenance Recording for Services (PReServ) is available at www.pasoa.org.

Functionality

• A Web Service based persistent provenance store.

• Java Libraries for recording documentation of process in the store.

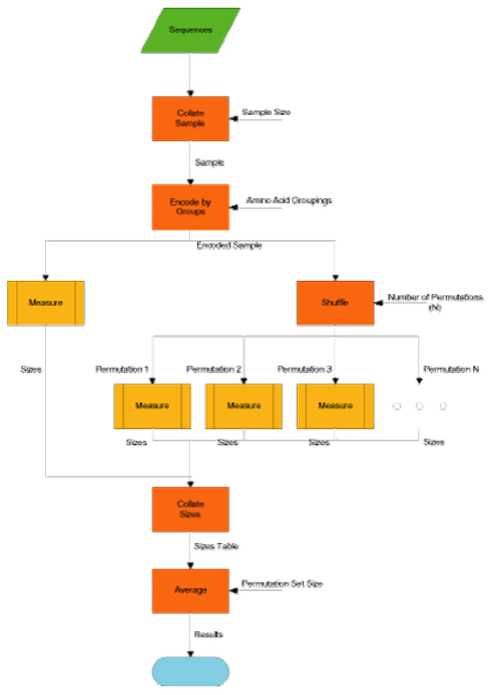
• An Axis Handler for transparently recording interactions in Axis based Web Services.

• Distributed under the MIT open source license.

• Tested in the applications shown to the right.

The Protein Compressibility Experiment

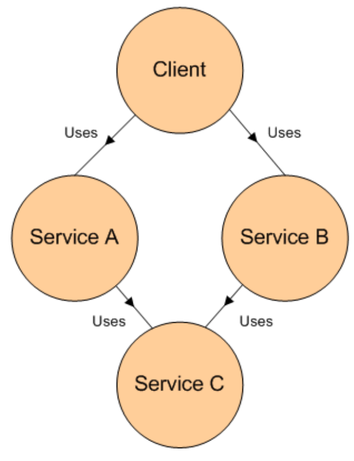
The bioinformatics domain involves the analysis of a massive amount of complex data, and, as experiments become faster and automated to a larger degree, the experimental records are becoming unmanageable. The Protein Compressibility Experiment (PCE) is a bioinformatics experiment to identify the relative complexity of different amino acid alphabet groupings. This experiment was designed by Klaus-Peter Zauner (University of Southampton) and Stefan Artmann (University of Jena). *PASOA converted this application to be grid-enabled and provenance-aware using PASOA techniques.* This included wrapping software components to automatically record documentation of process using PReP.



Highlights of the experiment:
• Thousands of activities per experiment run.
• Large amounts of documentation generated and stored.
• The achievement of two provenance related use cases specified by the scientists.
• Used common Grid middleware, Virtual Data System, Condor and Globus
• Used a combination of scripts and Web Services.
• A scalability analysis.

E-Demand, Investigating Fault Tolerance in Grids

E-Demand's (CS for E-Science supported project) FT-Grid system introduces a replication-based fault tolerance scheme that allows faults occurring in service based systems such as Grids to be tolerated, thus increasing the dependability of such systems.



The E-Demand project investigated a provenance-aware fault tolerance technique in order to try and detect failures originating from shared services. To achieve provenance awareness in their application, they made use of the Axis Handler and Provenance Store components of PReServ. The provenance-aware application used multiple distributed provenance stores. *This new fault tolerance functionality was made possible by the provenance-awareness provided by our techniques and software..*

Provenance and High Energy Physics

PASOA is also *investigating the possible use of provenance for tracking, analysing, verifying data sets in the ATLAS Experiment* of the Large Hadron Collider at CERN. Some use cases derived from this experiment are listed below.

Use Cases:

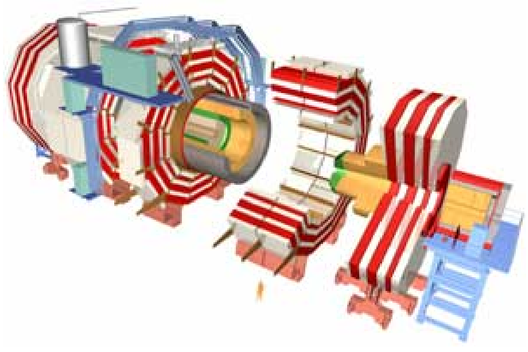
• Users are interested in regenerating previously created datasets. These datasets can be generated and (re)processed repeatedly on physically distinct locations (distributed environment using the Grid). The reproduction of the datasets must be possible both in terms of previous conditions (software versions, ...) to check whether a dataset was correctly created, and in terms of current conditions to allow comparison and detect differences in datasets.

• Users want to quickly discover if a dataset has been generated in the past and still exists.

• A user has to be able to query provenance of a specific dataset regardless of where the data and associated information are physically stored.

• Users must be able to manually add annotations to data according to some user-defined schema.

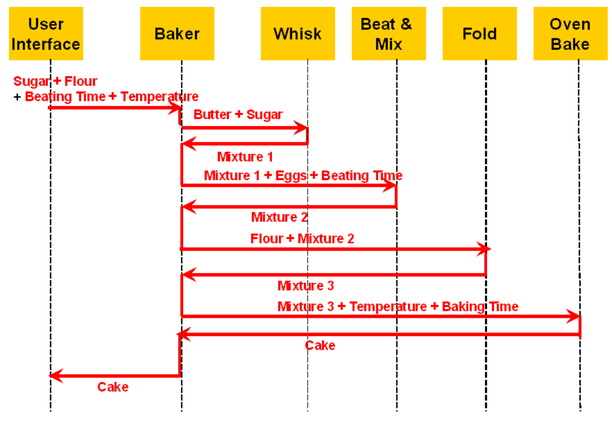
• Users and developers integrating existing services with a provenance infrastructure must be able to use existing services without changes - wrapping services.



EU Provenance

The EU Provenance project used PReServ to develop a simple, intuitive demonstration application. Using this demonstration application they explain and explore the principles of provenance, with a view to eventual standardisation.

The example scenario chosen for the demonstration is a bakery in which Victoria sponge cakes are baked. Each cake is baked on demand for a customer by a baker. The baking of each cake is a process with well-defined results, i.e. a cake, and so is good for illustrating the use of provenance. Using the baking example, a number of provenance questions can be asked including; why did the cake taste bad? What was the longest step in the baking process?



The demonstration was implemented as a set of Web Services, one for a baker and one for each of the steps in the baking process. A Web Service client was also implemented to simulate the customer. All clients and services were deployed using Axis. A Provenance Store was used to hold the documentation of process and each Web Service and client used the Axis Handler to automatically record their interactions without changing their business logic.

University of Southampton
CARDIFF UNIVERSITY
EPSRC Engineering and Physical Sciences Research Council