Provenance Aware Service Oriented Architecture

www.pasoa.org

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

Omer Rana, David Walker, Ben Carroll
School of Computer Science
Cardiff University

About PASOA

Project Summary

PASOA 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 eScience. The problems of determining the origin of a result or deciding when results of analysis are no longer valid become important concerns in an open Grid environment, where providers are dynamically organised in virtual organanisations to offer services to the community. In this context, provenance data is an annotation able to explain how a particular result has been derived.

PASOA Objectives

* To define execution and service provenance in relation to workflow enactment.
* To conceive algorithms to reason over provenance data, in order to help scientists to achieve better utilisation of Grid resources for their specific tasks.
* To design a distributed cooperation protocol to generate provenance data in workflow enactment.
* To investigate value-added properties that can be deduced from provenance-based data.
* To engineer a proof of concept software architecture to support provenance generation and reasoning in Grid environments.

About Provenance

Provenance Definition

Main Entry: prov•e•nance
Pronunciation: 'präv-n&n(t)s, 'prä-v&-"nän(t)s
Function: noun
Etymology: French, from provenir to come forth, originate, from Latin provenire, from pro- forth + venire to come -- more at PRO, COME
Date: 1785
1 : ORIGIN , SOURCE
2 : the history of ownership of a valued object or work of art or literature

Uses of Provenance

* aide memoir
* re-enactment
* notice of change to source data
* patenting
* proof
* auditing
* relating result data to source data
* debugging
* hazard detection
* provides context for result data
* usage pattern detection
* scientist's lab-book

Two Types of Provenance

Execution Provenance relates to data recorded by a workflow engine during a workflow execution. It identifies what data is passed between services, what services are available, and how results are eventually generated for particular sets of input values, etc. Using execution provenance, a scientist can trace the "process" that led to the aggregation of services producing a particular output.

Service Provenance relates to data associated with a particular service, recorded by the service itself (or its provider). Such data may relate to the accuracy of results a service produces, the number of times a given service has been invoked, or the types of other services that have made use of it. A service provider may make such parameters available to other users to enable them to select services that are more likely to produce the output they desire.

Importance of Provenance

All data processing in scientific computing requires some level of judgement (Bob Man)

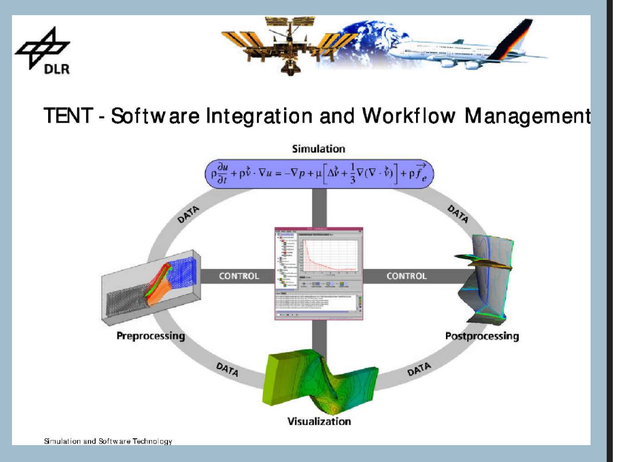
Technical Concern:
How can we provide annotations that reflect these judgements?

Sociological Concern:
Where does data curation stop, and scientific research begin?

Provenance is a crucial concern in Scientific Problem Solving, as the accuracy of the find is significantly based on the origins of data. This is especially true in multi-disciplinary science -- as made possible via eScience -- where data from different sources needs to be fused.

Applications of Provenance

Aerospace
Provenance requirement: to maintain a historical record of inputs from each sub-system involved in simulations. Aircrafts' provenance data need to be kept for up to 99 years when sold to some countries. Currently, there is little direct support available for this.

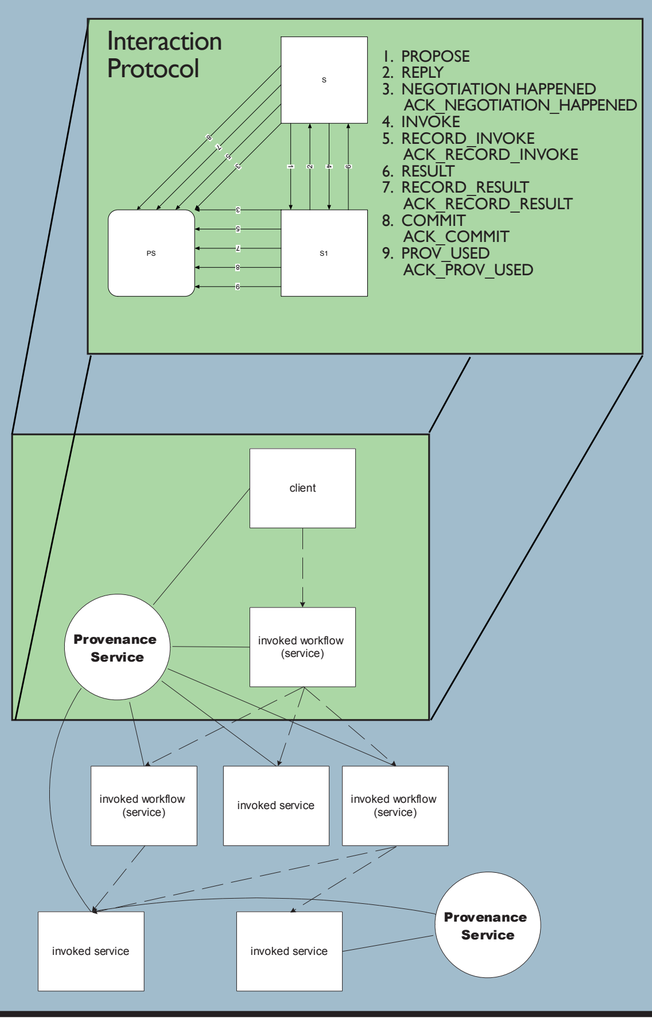


Medical Information Systems
These systems rely on a wide range of data sources, human input and access to patient data. In many cases, domains are highly regulated, must retain careful audit data, and rely heavily not only on information in the system but knowledge added by doctors, surgeons and other individuals using the systems.

Bioinformatics Grid
The myGrid project builds a personalized problem-solving environment that helps bioinformaticians find, adapt, construct and execute in silico experiments. These in silico experiments can have wide ranging provenance requirements, for example, The Food and Drug Administration in the United States requires drug companies to keep a record of the provenance of drug discovery as long as the drug is in use (up to 50 years sometimes).

Service Oriented Architecture

PASOA Architecture



The PASOA architecture is an asynchronous mechanism for recording provenance. The architecture is designed to work with multiple provenance services, as well as multiple workflow enactment engines, i.e. a client calls a workflow enactment engine which then calls a variety of services some of these services (workflow enactment engines) may in turn call other services until a result is returned up the hierarchy.

Even with various services, workflow enactment engines, and provenance stores all interacting, the basic building block of the architecture is the interaction between three entities: the client, the service, and the provenance service. The interaction between these three entities is shown in the Interaction Protocol portion of the architectural diagram. Essentially, the client and service are required to submit all their inputs and outputs to the provenance service for storage. The decision as to which provenance service to use is decided during a negotiation phase before a service is invoked. All submissions to the provenance service can be done in an asynchronous manner allowing for scalability.

Architectural Challenges

Security
Both members of a virtual organisation and users of the provenance data produced by that virtual organisation must trust that any provenance data produced is submitted to a trusted provenance store in a secure manner. The PASOA architecture seeks to ensure this by implementing a protocol that supports mutual authentication and non-repudiation. With mutual authentication, an invoked service can ensure that it submits data to a specific provenance server, and vice-versa, a provenance server can ensure that it receives data from a given service. With non-repudiation, the provenance service can retain evidence of the fact that a service has committed to executing a particular invocation and has produced a given result. The PASOA architecture will also employ crytopographic techniques to ensure these properties.

Scalability
Provenance generation may result in high volumes of provenance data to be submitted to provenance services. This process may be expensive, and we would not want it to delay the execution of workflows. Therefore, it may be desirable to submit provenance data in an asynchronous manner, essentially "staging" provenance data to temporary stores, and transferring it when suitable.

Generality
A wide variety of services and grids should be able to use PASOA's submission architecture in order to record provenance data.

Customization
The PASOA architecture should allow provenance recording and retrieval to be tailored to any specific application domain. This requires customisation of the submission, storage, and reasoning processes.

Related Work

Related work can be found at:

Workshop on Data Derivation and Provenance
(Organised by: Peter Buneman, Ian Foster, October 17-18, 2002, Chicago)

Workshop on Data Provenance and Annotation
(Organised by: Dave Berry, Peter Buneman, Michael Wilde, and Yannis Ioannidis, December 1-3, 2003, Edinburgh)

EPSRC
CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD
University of Southampton