Project no: 511085

STREP

Thematic Priority

Provenance: 6th Monthly Interim Report

Deliverable D1.1.1

"Enabling and Supporting Provenance in Grids for Complex Problems"

Period covered: from 01-09-2004 to 28-02-2005 Date of preparation: 28-02-2005

Start date of project: 01-09-2004 Duration: 24 Months

Project coordinator name: John Ibbotson
Project coordinator organisation name: IBM United Kingdom Limited Revision draft

Project co-funded by the European Commission within the Sixth Framework Programme (2002-200C)		
Dissemination Level		
PU	Public	
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	**X**

The PROVENANCE project receives research funding from the European Community's Sixth Framework Programme

File name: GRID_PROVENANCE-GEN-M3-SixMonthReport-WP2-Input-final.sxw

CONTENTS

The PROVENANCE project receives research funding from the European Community's Sixth Framework Programme

1 Project objectives and major achievements during the reporting period

The overarching aim of the Provenance project is:

> *To design, conceive and implement an industrial-strength open provenance architecture for Grid computing, and to deploy and evaluate it in complex grid applications (aerospace engineering and organ transplant management).*

Specifically, the objectives of the project are:

1. To specify the contents of provenance in relation to workflow enactment.
2. To design and implement a scalable and secure distributed co-operation protocol to generate provenance data in workflow enactment.
3. To conceive and implement tools to navigate, harvest and reason over provenance data, also in a scalable and secure manner.
4. To design and engineer a scalable and secure software architecture to support provenance generation and reasoning.
5. To deploy and evaluate the provenance system in two different grid applications, namely aerospace engineering and organ transplant management.
6. To propose a draft provenance specification for input to an open standardisation process thereby contributing to the standardisation efforts in this area within the Grid and Web Services architecture domains.

1.1 Objectives for the Period

According to the project plan provided in the Provenance Technical Annex, objective 1 from section 1 above would be achieved at the end of the first six month period. To meet this objective, the Provenance project would provide:

1. A set of *User and Technical Requirements for provenance, for the two selected applications, and also for related applications with provenance needs, so as to ensure the genericity of the effort* .
2. *A pre-prototype of a provenance system; the pre-prototype will consist of an assemblage of existing software, including a exemplar provenance service described in [SM03] and a prototype implementation of a provenance recording protocol of the pasoa project (www.pasoa.org).*

1.2 Major Achievements during the Period

The following are considered the major achievements during this six month period:

1. A project kick-off meeting was held at IBM United Kingdom Laboratories during October 25-27 to initiate the project and set the management processes and project plans for the first six months.
2. The Architecture workpackage WP3 was started early to provide direction and design for a pre-prototype demonstrator. The demonstrator would provide a testbed for a provenance architecture. The design of the demonstrator would be described in a document included as part of the WP9 deliverable at the end of month six.

3. A development system was established as part of workpackage WP9. This included a source code management system hosted by UWC that allowed access by all project partners.
4. A pre-prototype demonstrator was developed that will be available on the Provenance public website together with its specification. The purpose of the demonstrator is to illustrate the use of provenance to the general public and to provide a proof of concept testcase for later releases of the functional software.
5. A management report for the first three months of the project was provided to the EC Project Officer.

2 Workpackage Progress

The following sections describe in more detail the activities in each of the non-management workpackages WP2 to WP10. This also identifies the workpackages that have not yet started. The project management activities are described in section 3 – Consortium Management.

2.1 WP2: Requirements Capture

The main objectives for this workpackage in the period were to investigate provenance generation, access, navigation and use, collect a set of scenarios and usage options, and to derive user and software requirements. This work package delivers requirements for design, development, deployment and evaluation. All partners were involved. The specific objectives were:

- to define a questionnaire to elicit requirements from the different users,
- to distribute the questionnaire within the project and among all FP6 grid projects,
- to infer user requirements by collating scenarios and results from the returned questionnaire,
- to infer software requirements from the questionnaire results and scenarios.

Activities carried out to achieve these objectives were:

- At the project kick-off meeting two sessions were dedicated to workpackage 2. All partners actively participated in the WP2 session. The results of the WP2 sessions at the kick-off meeting were:
    1. a schedule for WP2 until the next face to face meeting was accepted,
    2. initial work on requirements questionnaire was presented,
    3. the content of the requirement questionnaire was discussed,
    4. contact persons were assigned for each project to be interviewed.
- Draft versions and the internal final version of the questionnaire were produced on schedule. Project partners contributed through the twiki collaboration space and a telephone conference call on 18/11/04. The on-line questionnaire was produced with an in-house developed form generator.
- The internal final draft of the questionnaire was filled in by five projects: the TENT system by DLR, the Organ Transplant Management application by UPC, the eDiamond project by IBM, Combechem and MyGrid projects by Soton. The partners filling in the questionnaire also gave feedback on the questionnaire. The answers and the feedback from these projects were analysed and the modified final version of the questionnaire was published on 17/12/04.
- The list of projects to be contacted for requirements input was collected. In addition to the project mentioned in the previous paragraph, the list contained the FP6 grid projects and the projects listed in the recent grid special issue of the ERCIM News. The first contact email was sent to 47 projects on the week of 20-23 December.

- Answers to the questionnaire were received and processed to produce the draft versions of the user requirement document. Project partners took part in collating the user scenarios through the twiki collaboration space and a telephone conference call on 27/01/05. Input from the following projects was received for the user requirements document:
    1. Organ Transplant Management application
    2. TENT system
    3. eDiamond project
    4. Healthcare and Life Sciences Framework
    5. CombeChem
    6. myGrid
    7. GENSS
    8. Traffic Management Application (K-WFGrid project)
    9. DataMiningGrid
    10. UniGridS
    11. Diligent
- The deliverable D2.1.1 User Requirements Document was finalised and approved by the project by 28/02/05.

Deliverables and milestones completed in the workpackage in the previous period are shown in the following table.

Deliverable Number	Deliverable Name	Date Due	Actual/Forecast delivery date	Lead Partner
D2.1.1	User Requirements Document	31/01/05	15/03/05	STA
D2.2.1	Software Requirements Document	31/03/05	28/04/05	STA
M2.1.1	Questionnaire issued	31/10/04	17/12/04	STA
M2.3.1	Draft of software requirements for application specific components and tools	28/02/05	31/03/05	STA

Deviations from the Work programme and or Corrective Actions (If any)
Because of the late start of the project, the workplan to reach the objectives of this workpackage has been changed. The schedule for WP2 was accepted at the kick-off meeting:

Draft 1 of questionnaire to include partner input		05/11/04
Draft 2		12/11/04
Internal final draft		19/11/04
UPC and DLR complete questionnaire		10/12/04
eDiamond and myGrid complete questionnaire		10/12/04
Publication of questionnaire		19/12/04
Identification of user requirements		31/01/05
User requirements doc (D2.1.1)	(31/01/05 in plan)	28/02/05
Internal draft of Software requirements		28/03/05
Software Requirements (D2.2.1)	(31/03/05 in plan)	28/04/05

This means that the work package does not deliver the software requirements in this reporting period.

Objectives for the Next Period
According to the revised schedule of workpackage 2, the workpackage will finish 2 months later than planned. In these two additional months the Software Requirements will be inferred.

2.2 WP3: Architecture

While the initial workplan did not start WP3 before month 6, the project agreed that it would be good to produce an overview of a logical infrastructure by Month 6. Hence, we decided that the design of the pre-prototype would fall under the remit of WP3, so as to lead to the logical architecture.

The main objective for the workpackage in the period were: to design the pre-prototype application and architecture so as to prepare for the preliminary specification of a logical architecture. Activity carried out to achieve these objectives were:

- Design of pre-prototype, result: pre-prototype document, partners responsible: UoS/contributing IBM.

The focus of the architecture workpackage has been on the design of the pre-prototype architecture. The goals of the pre-prototype are:

- to be a simple example to understand
- to exhibit interesting examples of provenance use
- to inform the future design of the architecture
- to help build, setup and deploy a development environment for the Project.

We opted for a small example that shows how to bake a cake (Victoria Sponge). It consists of a simple worflow:

1. Whisk together the butter and sugar until light and creamy.
2. Beat the eggs for a certain duration and add to the whisked sugar and butter.
3. Fold the flour into the mixture and add the flavoring (vanilla, lemon).
4. Put the wet dough into the oven and bake for a given time at a given temperature.

Despite its simplicity, this application leads to interesting questions that require the use of provenance in order to be answered. We have identified a preliminary list of these:

- Were the correct ingredients used at the correct step?
- Where was the longest time spent in preparation?
- Why did the cake taste wrong? (answers can be related to altitude, wrong temperature, eggs not beaten enough, proportion of ingredients)
- Was the proportion of ingredients right for the size of the cake?
- Did the baker follow the users' instructions (regardless of any claims from the baker)?
- Did the services follow the users' instructions (regardless of any claims from the services)?

Such questions can be answered if provenance is recorded, specifically consisting of:

- *interaction provenance*: the trace of all service interactions during workflow execution;
- *actor provenance*: a trace of internal service states at specific moments of execution.

The PROVENANCE project receives research funding from the European Community's Sixth Framework Programme

After specifying the application's behaviour, we have designed its architecture. Specifically, we identified the services involved, their interactions (defined as sequence diagrams), and their interfaces (as wsdl files). The whole application data domain was defined as an XML schema. From a provenance viewpoint, we adopted (as indicated in the proposal) an existing implementation of a provenance service, designed as part of the PASOA project (www.pasoa.org), which is capable of recording actor and interaction provenance.

Implementation work was split between IBM and Southampton. IBM have implemented the different services of the application, while Southampton focused on the implementation of the provenance queries. In order to proceed with the later, without having an application available, a dummy client simulated the whole cake baking process was created in order to submit all the necessary provenance information to the provenance service. Three provenance queries verifying the baking temperature have been implemented. Regular meetings have been held to coordinate the implementation and architecture activities.

Deliverables and milestones completed in the workpackage in the previous period are shown in the following table.

Deliverable Number	Deliverable name	Date Due	Actual/Forecast delivery date	Lead Partner
Internal	Pre-Prototype document	31/03/05	31/03/05	UOS
M.3.1.1	Logical view of architecture	31/03/05	31/03/05	UOS

Deviations from the Work programme and or Corrective Actions (If any)
While the initial workplan did not start WP3 before month 6, the project agreed that it would be good to produce an overview of a logical infrastructure by Month 6. Hence, we decided that the design of the pre-prototype would fall under the remit of WP3, so as to lead to the logical architecture.

Objectives for the Next Period
In the next period, we will analyse technical requirements, and design a logical architecture, for which we have introduced an internal milestone (M3.1.1). This is to be followed by a complete design of the architecture. Issues to be investigated include provenance definition, naming convention, protocol for recording, query functionality.

2.3 WP4: Security

No activity has taken place for this workpackage in the period covered by this report.

Deviations from the Work programme and or Corrective Actions (If any)
No deviations from the work programme.

Objectives for the Next Period

- Analysis of Provenance Generation Security Requirements. This analysis will take as input the User Requirements document (Deliverable D2.1.1) to generate a set of security requirements for the generation of Provenance information.
- From the analysis, a specification will be generated for deliverable D4.1.1 which is due at the end of August 2005.

The PROVENANCE project receives research funding from the European Community's Sixth Framework Programme

2.4 WP5: Scalability

No activity has taken place for this workpackage in the period covered by this report.

Deviations from the Work programme and or Corrective Actions (If any)
No deviations from the work programme.

Objectives for the Next Period

- Analysis of Provenance Generation Security Requirements. This analysis will take as input the User Requirements document (Deliverable D2.1.1) to generate a set of security requirements for the generation of Provenance information.
- From the analysis, a specification will be generated for deliverable D5.1.1 which is due at the end of August 2005.

2.5 WP6: Tools and Setup

No activity has taken place for this workpackage in the period covered by this report.

Deviations from the Work programme and or Corrective Actions (If any)
No deviations from the workprogramme.

Objectives for the Next Period

- Review of existing tools and frameworks for Provenance information management.
- From the review, a document will be produced to review tooling options. This is deliverable D6.1.1 which is due at the end of August 2005.

2.6 WP7: Application 1 - Aerospace

Workpackage 7 is not scheduled to start until month 6 of the project.

Deviations from the Work programme and or Corrective Actions (If any)
No deviations from the workprogram

Objectives for the Next Period
WP7 activities begin at the start of the next period. Major objectives in the next six monts are:

- Development of models and specifications for handling of domain specific provenance in the Aerospace Engineering domain. This will be based on the requirements collected by users of TENT and information already gathered with the existing logging system of TENT.
- Preparation for the development of provenance mapping. This includes an design and implementation plan for integrating the TENT system with the provenance architecture/services.

2.7 WP8: Application 2 – Organ Transplant Management

Workpackage 8 is not scheduled to officially start until month 6 of the project, however during the period several activities were carried out in order to prepare for the start of activities. In particular these are:

- Development of an overall description of the Organ Transplant Application which will form the basis of the scenario, including interviews with medical/technical staff at the

Hospital St. Pau in Barcelona, Spain. Contributors: UPC (lead) with comments and feedback from SOTON and SZTAKI..
- Development of a short overview document for the area of patient care record management based on the EHCR records project in order to prepare for discussion of the development of WP8 demonstrator. Contributors: SZTAKI.
- Contributions to WP2 requirements gathering on the basis of these documents.

Deliverables and milestones completed in the workpackage in the previous period are shown in the following table.

Deliverable Number	Deliverable name	Date Due	Actual/Forecast delivery date	Lead Partner
Internal	Internal document: Outline Organ Transplant Management Scenario: GRID Provenance Project	N/A	31/01/05	UPC
Internal	Internal document: Overview of ECHR patient care records.	N/A	31/01/05	SZTAKI

Deviations from the Work programme and or Corrective Actions (If any)
No deviations from the workplan.

Objectives for the Next Period
WP8 activities formally begin at the start of the next period. Major objectives in the next six months are:
- Development of models and precise specifications for handling of domain specific provenance in the Organ Transplant Management domain. This will be based on the documents already created (I.8.1 and I.8.2) but include more detailed application information as well as decisions on the types of provenance queries to be answered.
- Preparation for the development of provenance mapping – identifying an architecture for intermediate layers between the existing Organ Transplant Management applications and the provenance architecture/services.

2.8 WP9: Implementation, Integration and Test

The major objectives of the of the Implementation, Integration and Test workpackage in the first six months were two-fold.

- To put in place a development infrastructure that can be shared amongst the partners
- Development of a pre-prototype

The development infrastructure includes the following components:

- A CVS server to manage code developed by the partners
- A build server situated in IBM Hursley

The development code is managed in a CVS repository at the University of Wales, Cardiff. IBM does not allow access to internal systems except in extreme circumstances so this site was chosen to provide easy access to non-IBM partners. The machine in Cardiff has been configured and is being used by the IBM developers.

The development process to be used by the project will focus around daily builds of the code together with regression unit testing. Build scripts are being prepared which will pull the development code from the CVS server onto a build machine at IBM Hursley. The build process will then complete and run a set of regression tests on all modules. The regression tests will be written with JUnit. Following the build, a report will be prepared which can be accessed by all developers. The built system together with reports will then be pushed onto a system external to IBM for use by the development team.

To support the development process, coding and information standards have been prepared. A draft copy of the HandBook has been circulated to the project partners for comment.

To test the development process, a pre-prototype has been specified by the architecture workpackage team. During discussions between IBM and Southampton, we established a number of principles for the Provenance implementation:

- The Provenance implementation would not impose a development environment on anyone that wanted to use the released code
- The Provenance code releases would not include third party pre-requisite software. For IP reasons, users of the Provenance code will have to download pre-requisite software (such as Apache Tomcat and Axis) separately
- The Provenance code will make use of Apache Ant build scripts to compile and deploy the system on an application server
- The Provenance code will be developed using the Apache Tomcat application server and Apache Axis SOAP support
- The Provenance code will be developed using the Java language to provide portability across systems

Activities carried out to achieve the workpackage objectives were:

- IBM and UWC have configured a CVS repository at Cardiff for access by the project partners
- IBM has configured a server that will build releases of the development code for the project. In addition, IBM has defined development code guidelines and distributed them amongst the partners for approval.
- UOS has specified a demonstration scenario that is being implemented jointly by UOS and IBM. This forms the internal pre-prototype deliverable D9.1.1 and uses the pre-existing Provenance Recording Protocol (PreP) developed by UOS as part of the PASOA project.

Deliverables and milestones completed in the workpackage in the previous period are shown in the following table.

Deliverable Number	Deliverable name	Date Due	Actual/Forecast delivery date	Lead Partner
D9.1.1	Pre-prototype System	28/02/05	15/03//05	IBM

There were no milestones defined for this six month period in the workpackage plan contained in the Provenance Technical Annex.

Deviations from the Work programme and or Corrective Actions (If any)
No deviations from the workprogram.

Objectives for the Next Period

- Given the Software Requirements Document (Deliverable D2.2.1) from workpackage 2, a design process for the Functional Prototype (Deliverable D9.2.1) will commence leading to the delivery of D9.2.1 by the end of August 2005.
- Milestones will be established for this six month period as part of the software development process.

2.9 WP10: Collaboration

This collaboration workpackage covers the liaison and co-operation activities with the other IST projects under Strategic Objective "Grid-based systems for solving complex problems" and its successor in WP 2005/2006.

The following activities have been completed to support the collaboration workpackage:

- We have produced our draft collaboration workplan that was delivered to the EU at the end of October 2004.
- Provenance (Luc Moreau) attended the first meeting of the European Grid Standards Co-ordination Group, on January 28, 2005 at European Microsoft Innovation Centre in Aachen, Germany. This resulted in Provenance filling the COPRAS questionnaire.
- The Provenance project will contribute a "Provenance" education session at the core grid summer school
- Provenance (John Ibbotson) has attended meetings of the key indicators working group with the EU in Brussels and has contributed to the working group report.
- The project contributed to the AgentLink III roadmap, overview and consultation report, page 23, identifying provenance as a "specific challenge". *Today's distributed environments (includingGrid, Web Services and agent-based systems) suffer from a lack of mechanisms to trace results and a lack of infrastructures to build up trusted networks. Provenance enables users to trace how a particular result has been arrived at by identifying the individual and aggregated services that produced a particular output. From both an academic and an industrial perspective, the research question is to design, formalise and implement an open provenance architecture. Such a provenance architectureshould be scalable and secure; it must be open and promote interoperability.*

Deviations from the Work programme and or Corrective Actions (If any)
No deviations from the workprogram.

3 Consortium management

The Provenance project kick-off meeting was held on 25-27 October 2004 at the IBM UK Laboratories. At this meeting, the working processes of the project were agreed together with a schedule for the first six months. Key points agreed were:

- All project documents would be managed via a twiki. This has been created at http://twiki.gridprovenance.org with public and restricted areas. The restricted area is for project partners only. The restricted part includes areas for meetings, workpackages, deliverables and management.

The PROVENANCE project receives research funding from the European Community's Sixth Framework Programme

- Monthly management telcons are held with agendas published on the twiki. All meetings are minuted with actions tracked – again this is via the twiki.
- Dates have been agreed for regular face to face meetings during the project lifetime. The next meeting will be in Barcelona at UPC in March 2005.
- Standards were agreed for documents generated by the project. Where possible, they will conform to open standards and use the OpenOffice suite of tools.
- Project planning is managed by actions and dates agreed by the partners. These are published and tracked on the twiki.

The main management tasks to be carried out in this period were the following:
Activities carried out to achieve these objectives were:

- A kickoff meeting for the Provenance project was held at IBM Hursley. Details of the agenda and supporting minutes are at http://twiki.gridprovenance.org/bin/viewauth/Restricted/KickOff
- Management conference telephone calls were held monthly amongst the partners. Agendas and minutes are at http://twiki.gridprovenance.org/bin/viewauth/Restricted/PhoneConferences

In addition the following arose and were dealt with:

- The Coordination workpackage (WP10) proposed by the EC was included in the Provenance project plan. For more details see section 2.9.
- At the kickoff meeting, the delayed start to the project and its consequences were discussed. A revised schedule for workpackage 2 (Requirements) was produced and communicated to the FP6 Project Officer on 05/11/2004. Further details of this revised plan are in section 2.1.
- A report for the first three months of the project was prepared for the EC Project Officer.
- Initial pr-financing was received by IBM from the EC and distributed to partners.

Deliverables and milestones completed in the workpackage in the previous period are shown in the following table.

Deliverable Number	Deliverable name	Date Due	Actual/Forecast delivery date	Lead Partner
Internal	Three Month Report	11/01/05	11/01/05	IBM

4 Annex A: Use and dissemination

Overview table

Date	Type	Type of audience	Countries addressed	Size of audience	Partner responsible / involved
12/12/04	Dissemination Meeting, Hospital de St. Pau, Barcelona, Spain	Organ transplant unit representatives (medical doctors and administrators) + technical staff	1	10 persons	UPC

Date	Type	Type of audience	Countries addressed	Size of audience	Partner responsible / involved
17/01/05	Press Release available at http://www.ecs.soton.ac.uk/news/671				UoS