Large Scale Integrating Project

Grant Agreement no.: 216736

D14.4 - REPORT ON DEMONSTRATION AND EVALUATION ACTIVITY IN THE DOMAIN OF E-SCIENCE

SHAMAN –WP14-D14.4

Project Number	FP7-ICT-216736
Due Date	30 November 2011
Actual Date	30 November 2011
Document Author/s:	SSLIS, HATII, INESC
Version:	SHAMAN-WP14.4 Version 0.5
Dissemination level	Public
Status	Final
Contributing Sub-project and Work Package	SHAMAN WP14
Document approved by	

Co-funded by the European Union

Document Version Control

Version	Date	Change Made (and if appropriate reason for change)	Initials of Commentator(s) or Author(s)
0.1	10 November, 2011	First suggested outline with draft text	EM, TDW
0.2	24 November 2011	Developed and corrected text	EM, TDW, GA, JB
0.3	29 November 2011	Final text of the D14.4	EM, TDW, GA, JB
0.4			
0.5			
0.6			
0.7			

Document Change Commentator or Author

Author Initials	Name of Author	Institution
EM	Elena Maceviciute	SSLIS
TDW	Tom Wilson	SSLIS
GV	Giusepina Vullo	HATII
GA	Gonçalo Antunes	INESC-ID
JB	José Barateiro	INESC-ID

Document Quality Control

Version QA	Date	Comments (and if appropriate reason for change)	Initials of QA Person
V0.5	2011-11-30		
V0.7			

Catalogue Entry

Title	*Report on demonstration and evaluation activity in the domain "Memory institutions"*
Creators	*Elena Maceviciute, Tom Wilson, Gonçalo Antunes, José Barateiro, Giuseppina Vullo*
Subject	ISP3 demonstration scenarios, demonstration and evaluation by R&D community, customers, end users
Description	Report on the activities carried out to present the ISP3 to the evaluators and the results of these activities.
Publisher	SHAMAN
Contributor	
Date	2011
ISBN	
Type	Public project deliverable
Format	Text
Language	English
Rights	

Citation Guidelines

SHAMAN Project. (2011). D14.4 - Report on demonstration and evaluation activity in the domain of e-science. SHAMAN –WP14-D14.4.

Table of Contents

1. SUMMARY

1. This deliverable reports on the demonstration and evaluation of ISP3, which was designed to demonstrate the potential of the SHAMAN framework for digital preservation in the context of e-science institutions and for the research and development community.

2. The demonstration process was carried out by means of presentations of the training material to the scientists from LNEC and LIP in Lisbon, Portugal. The audiences for the demonstrations consisted of scientists working with different data and IT support staff.

3. The evaluation is based on the reports of focus groups held with two groups of scientists, together with structured data from self-completed questionnaires, administered on the same occasions.

4. The SHAMAN 'consciousness raising' event appears to have had useful consequences. In both institutions it appeared to fulfil the aim of drawing attention to the problems of digital preservation and to the role of the SHAMAN framework in finding answers to the problems. To a degree, the participants were already sympathetic to the ideas, since they were working in situations that daily brought their attention to the need for preservation policies. However, the discussions were also useful in extending the concept of digital preservation beyond the existing perceptions within the SHAMAN teams. In particular, the notion of the 'digital object' needs to be expanded to include not only digitized physical objects such as documents and images, but also 'born digital' phenomena such as data streams, software programs and analytical results of experiments.

2. INTRODUCTION

The work undertaken in WP14 addresses three different communities with a stake in the project's outputs; that is, members of the three previously identified SHAMAN Domains of Focus (DoF). These are, memory institutions (DoF1); industrial design & engineering (DoF2); and e-science (DoF3).

The following tasks are outlined for the WP14:

Task 14.1 Demonstration to, and evaluation by, researcher and developer communities (to show the integration of Digital Library and Persistent Archive technologies in a Grid environment, the Multivalent technology and the use of context representation and information extraction within advanced digital preservation applications).

Task 14.2 Demonstration to, and evaluation by, customer communities (aims to involve relevant communities that aim to set up their own digital libraries and persistent archives with the benefits of a Grid environment, Multivalent technologies, context support and information extraction functions in demonstration and evaluation activities).

Task 14.3 Demonstration to end-user communities (to enlist end-users from the previous two communities who may participate in demonstrations and early evaluations of the use of the demonstrators).

Task 14.4 Application of the SHAMAN Assessment Framework (to summarize the results of the evaluation activities, and to evaluate progress and impact on the representative target domains of the SHAMAN outputs on the basis of the criteria specified in Task 1.4).

Task 14.5 User evaluation plan (to set up the principles behind the entire user evaluation process and link it to the SHAMAN Assessment protocol, define the selection of users and user groups, the evaluation objects, methods and metrics, etc.).

This document is the outcome of the collaboration between the work packages on Tasks 14.1, 14.2, 14.3, and 14.5 in relation to memory institutions domain of focus. It pulls together the results of evaluation of the SHAMAN outcomes as demonstrated to the representatives of the customers and end-users.

Demonstration and evaluation is based on the work done previously in SHAMAN and reported in earlier deliverables. It builds on identified user requirements and their analysis as well as the SHAMAN Assessment Framework developed in WP1 (SHAMAN 2008; SHAMAN 2009b). The third round of evaluation and demonstration focuses on the assessment of the digital preservation reference architecture principles developed in the context of WP2 and applied to the scenarios described in WP13 and presented in the SHAMAN Reference Architecture document, and on the application and extension of the SHAMAN Core Infrastructure to e-science scenarios reported inD13.1. D14.1 Demonstration and evaluation plan has outlined the structure of the demonstration and evaluation acitivities. It has also defined detailed relations with other work packages (SHAMAN 2009) and each of them has contributed to the evaluation process to some extent. The methods and outcomes of the evaluation done by memory institutions and industrial design and engineering users presented in D14.2 and D14.3 were used in preparing the evaluation round with the e-science representatives.

2.1 The objectives of the demonstration and evaluation in the domain of e-science

1) WP14 has a shared responsibility with WP15, WP16, WP17 *to disseminate the results of SHAMAN development among the researchers and developers of digital preservation and to the potential customers and users.* Within WP14 this objective was achieved by organizing demonstration events in the e-science that complemented the activities of WP17 reported in D17.2 and simultaneously with the activities of the WP15.

2) WP14 has a shared responsibility with WP13 and other research WPs as well as WP18 to *evaluate the outcomes of the SHAMAN project for e-science institutions*. This objective was achieved through a number of activities planned in WP14.1 and reported in the deliverables produced by WP18. Therefore this deliverable concentrates on the assessment of an integrated sub-project ISP3 that includes the elements developed in other research WPs into one presentation of the digital preservation principles and research architecture.

3) As at this stage, evaluation is summative in nature as the SHAMAN project comes to an end. However, it is intended *to aid the design and implementation further* digital preservation attempts directed towards e-science data. In general terms, WP14 identifies how SHAMAN's reference architecture and principles can be better aligned with the current and future expectations of scientists and whether the ideas underlying SHAMAN are approved by the e-science community as a valid research direction. Taken as a whole, data gathered will also supplement the internal assessment activities taking place within Research & Technical Development (RTD).

4) In addition to informing RTD efforts, evaluation of the presented principles (as opposed to the prototype) offers a unique *insight into how the SHAMAN project is perceived by, and how it can be explained to those within its targeted Domains of Focus,* especially the third one.

2.2 The presentation/training and evaluation activities for the e-science

The nature of the ISP3 outcomes has dictated the presentation mode and the evaluation methods that could be applied in science institutions dealing with large amount of data that needs preservation. The choice of the focus group discussions was prompted by the fact that SHAMAN is not a fully fledged digital preservation system and no test-bed could be presented to the end-users for hands-on testing. The technological ideas could be best assessed by a discussion among informed end-users considering the relevance of the presented principles and features to their domain.

Three training/evaluation events were organized in Lisbon (Portugal) for scientists working with digital preservation problems in two big areas – engineering research and particle physics. There were two institutions involved in evaluation of the SHAMAN ISP3:

1. The Laboratório Nacional de Engenharia Civil – LNEC (Portuguese National Laboratory for Civil Engineering) – a state owned research and development (R&D) institution founded in 1946.
It works in the various domains of civil engineering, giving it a unique multidisciplinary perspective in this field. The main goals of the LNEC are to carry out innovative research and development and to contribute to the best practices in civil engineering. LNEC also plays a key role in advising the government in technical and scientific matters of civil engineering, as an unbiased and independent body.

2. The Laboratório de Instrumentação e Física Experimental de Partículas – LIP founded in 1986 and working in the field of experimental particle physics and associated instrumentation. LIP has played a key role in major international scientific collaborations.

The events in Lisbon consisted of the presentation and training event related to the digital preservation principles and ISP3 reference architecture (this event took place in Lisbon Institute of Engineering on October 27, 2011. The next day consequent focus group discussions were organized in LNEC and LIP with the members of the audience. The presentation/training/evaluation events were prepared and conducted together with the R&D team that has developed ISP3. The WP14 team helped in designing the presentations to the e-science members, developed the evaluation instruments, conducted the focus groups, analyzed the evaluation data, and produced recommendations. The partners working on WP13 and WP15 developed the presentation and participated in focus groups as observers. In addition, the partners from INESC-ID have selected the participants of the event and evaluation focus groups and organized the events in Lisbon.

2.3 The structure of the report

The report presents the preparation of the presentation/training and evaluation activities (Chapter 3) with an emphasis on the presentation approaches that influence the perception of the audience of the SHAMAN development outcomes. The methods used for eliciting feedback from the potential end-user community representatives are presented.

Chapter 4 includes the results of the end-user evaluation from the presentation/training/evaluation events.

The final Chapter 5 concentrates on the conclusions and recommendations to the developers, especially, with regard to the potential improvement of the digital preservation principles and the reference architecture of the ISP3.

3. PREPARATION FOR PRESENTATION AND EVALUATION IN E-SCIENCE

3.1. Preservation requirements of e-science institutions

The third SHAMAN Domain of Focus is concerned with e-science institutions. The term e-science is generally used to describe the application of information technology to academic research (originally science) domains enabling them to maximise the usefulness of their data.

The SHAMAN project is especially interested in the preservation of e-science digital data (documents, structured data files, etc) so that it continues to be of use to future generations of researchers. Typically, data sets are stored in more than one location and one data set may span many locations. Researchers involved in a particular study may also belong to many different institutions. These issues need to be addressed when tackling the preservation of e-science data. In addition, research programmes have a lifetime and ensuring enough information is captured such that the objects can still be used in a meaningful way in the future is of significant importance.

At the beginning of the project, the requirements have been extracted from the use cases based on interviews conducted by WP1. The institutions interviewed consisted of a civil-engineering laboratory, an academic research grid, a humanities research institute, and a climate data centre. The climate data centre and the humanities research institute were engaged in long-term preservation of some of their data and had developed sophisticated procedures mainly in collaboration with the IT institutions or departments. The other institutions were aware of the need to preserve their research data and were preparing for these developments. The academic research grid was using the grid capacity for variety of research purposes and calculation and also perceived its possibilities for long-term digital preservation of data.

3.2. Developing presentation and training approaches

The presentation and training process at this stage is closely associated with the evaluation, as the end users can assess only what has been presented. However, the combination of the presentation of the SHAMAN framework and reference architecture for the e-science and the training in digital preservation concept and SHAMAN DP policies can be perceived as a separate activity that can be carried out without any subsequent evaluation. This part of the event was developed and conducted in close collaboration with the WP15.

According to the specification of components provided by the RTD Work Packages for the ISP3 prototype, D14.4 focuses on the reference architecture and digital preservation policies as well as Digital preservation life cycle phases.

The presentation and training session on October 27 included the following:

1. Basics of digital preservation
2. SHAMAN digital preservations principles
3. The SHAMAN Reference Architecture and the ISP3

The SHAMAN reference architecture was the main element in the session. The SHAMAN reference architecture follows a holistic approach to digital preservation. The approach taken both to the life-cycle view of digital content and to the digital preservation architecture itself is concerned with the big picture and has an inclusive view of all the aspects relevant for digital preservation (SHAMAN 2011).

By reconciling and harmonizing concepts and terminology of different reference models, the proposed approach enables the reuse of best practices and improves common understanding, providing a multidimensional view on the concepts covered on these models.

The SHAMAN reference architecture should help developing concrete preservation-ready architectures, accommodating preservation-specific concerns in the architecture vision, while bridging into the development of business architecture aligned with that vision. The concrete scenarios can be addressed through the incorporation of contextual concerns in the architecture vision. This distances the SHAMAN approach from traditional approaches existing in the DP domain, as can be depicted in Figure 1. In this sense, it can be argued that this capability-based approach can deliver value to organizations in which the preservation of contents is not a main business requirement, but required to enable actual delivery of value in the primary business, which is the case of e-Science scenarios.
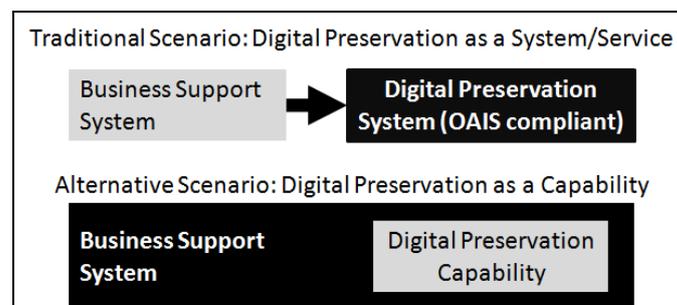
Figure 1 – Digital Preservation as a Capability

The approach taken with this reference architecture enables the transfer of DP know-how into a non-traditional repository-based DP scenario, since it is itself agnostic to concrete scenarios. In other words, this capability-based approach can deliver value to organizations in which the preservation of contents is not a main business requirement, but required to enable actual delivery of value in the primary business (SHAMAN 2011).

3.3. Selecting audiences

The selection of the participants for the event and focus group discussions was quite straightforward. The researchers working at LNEC and LIP who were most interested in digital preservation issues and depending on the long-term preservation of data were invited to the event. Some of the senior researchers took their doctoral students and some researchers combined responsibility for the IT services or data archiving.

3.4. Preparing evaluation methods and instruments for end-user evaluation

Combining evaluation with the presentation/training event, it was natural to select a focus group discussion as a main evaluation method. The WP14 team decided to supplement this

with questionnaires for soliciting personal feedback on the presented material to ensure that some structured information is obtained.

A focus group is a qualitative research method that involves asking a group of people about their perceptions, opinions, beliefs and attitudes towards a product, service, concept, or whatsoever. Focus groups can also be used to subject ideas to review to determine their viability, usefulness or functional applicability.

Using this method in relation to the evaluation of the SHAMAN preservation framework and resulting technological approaches would fall into the latter category of determining the viability of ideas, suitability and usefulness of technologies for certain functions in memory institutions. The technique was known to most of the evaluation team and this was also an argument for using it. In addition, modern sound recording technology provides possibility to get high quality recordings and in combination with observation notes the data capture is very reliable.

The focus group discussions concentrated on the issues of perceived usefulness, suitability for the DP policies of the participants and innovativeness, possibility of implementation of demonstrated principles, incentives and conditions of application (Annex 1). They were supplemented by a questionnaire in two parts (Annexes 2). One part presented at the beginning of the demonstration event helped to capture the data on the participants' jobs and experience with digital preservation. The second one provided an opportunity for everyone to express their personal attitudes towards the presentation and SHAMAN outputs. It also helped to record the expectations of digital preservation technologies and check the level of comprehensibility of the presentation. This questionnaire concentrated on essential demonstrated processes and functionalities.

4. EVALUATION BY CUSTOMERS AND END USERS

4.1 Evaluation of actions in the e-science domain of interest

A presentation, by the Training Work Package, on the nature of digital preservation and on the activities of SHAMAN took place in Lisbon at the Instituto Superior Técnico to an audience of engineers and scientists from the Laboratório Nacional de Engenharia Civil (LNEC) and LIP (Labóratorio de Instrumentação e Física Experimental de Particulas) on 27th October, 2011. This was followed by interviews with the same participants, in the respective organizations, on the 28th October. What follows is a report on those focus group interviews (and the associated survey questionnaires).

There were seven participants in the group from LNEC, mainly researchers from different areas of engineering (Concrete dam monitoring, Maritime and coastal engineering, Scientific instrument and systems development, Roads and accident, Information technology). They were working in the Laboratory from one year to 39 years (four for 1-7 years, three for 19-39 years).

Of four participants from the LIP, one was a post-doc student working for one year at the Institute, three were senior researchers working from 11 to 23 years.

4.1.1 Focus group in Laboratório Nacional de Engenharia Civil

LNEC is a state-owned research organization working in seven areas of civil engineering: concrete dams, buildings, structures (e.g., bridges), geotechniques, hydraulics and environment, materials, and transportation. The participants represented the concrete dams, structures, hydraulics, and transportation.

The participants welcomed the SHAMAN initiative and the presentations of the previous day. They noted that LNEC did not have a preservation policy, and that such a policy was needed. In particular, the representatives of the concrete dams department noted that they have statutory responsibility for maintaining the data received from the large dams in Portugal, but no policy on how those data are to be preserved exists.

A database of dam monitoring data is maintained (GestBaragens), using data supplied by the dam owners, some of which is by direct feed, other data are supplied on paper and have to input within LNEC. So far as the use of the data is concerned, each engineer can use analysis programmes provided by the Gestbarragens system, but can also write his or her own programs for analysis, and legacy systems exist (which have been used to run programs that may need to be rerun) and have to be maintained.

One participant pointed to a specific problem, that of trying to test a model of dam behaviour, with forecasts into the future, but finding that the programs that had been written for data analysis needed to be ported from the original PC versions to use on the grid. In other words, there is an issue not only of maintaining original data but also of migrating software programs used in analysis from older computing environments to the latest.

Another participant pointed to the need for adequate metadata to describe the data received, so that the nature of the data could be properly understood and so that any errors in the data (which is not unknown in the case of data coming directly from sensors in the dams) may be accounted for before analysis is undertaken.

In the case of road transport data (LNEC has a database of 20 years of road accident reports), it was noted that one of the issues regarding preservation relates to the re-use of old files, where format differences exist.

A representative of the hydraulics department noted that he had a personal interest in digital preservation, in that he maintained his own databases for various purposes. Currently, he preserved the data in two forms: ASCII records and binary data records, and was trying a third mode of preservation in GPD files (Microsoft's Generic Printer Description files). He was attempting to compensate for the changes in databases (Access is used) over time, which may result in earlier formats not being recognized by later versions.

The representative of the structures department noted that in respect of other structures, such as bridges, no statutory responsibility to maintain the data existed and, consequently, the department could decide what to keep and for how long. However, there remained the problem of retaining the results of analysis of the data (for comparison with future analyses) and a policy was needed to cover this problem. Although the needs of the departments might differ, a global policy, setting guidelines for action, was needed.

In response to a question from Elena Maceviciute on how the impact of SHAMAN might usefully be continued, there was general agreement that education or consciousness raising was key. One participant: "*All engineers must know about preservation, not necessarily do it, but know about it*". He also suggested that LNEC needed to establish a cross-department team to develop policy and consciousness raising.

In summary, the participants from LNEC were all advocates for the development of a digital preservation policy within the organization, recognizing that some departments might go ahead and develop their own policies and attendant strategies. Three levels of preservation potentials were described:

1.Original data, received from various sources, relating to various kinds of structures (e.g, dams and bridges), road accidents, water resources, etc. for which LNEC might or might not have a statutory responsibility. In any event, regardless of legal issues, policy was needed to guide retention, preservation and destruction activities. In addition to data received from external sources, data generated by experiments in the laboratories of LNEC also featured under this heading.
2.Software programs used in the analysis of data, which needed to be maintained so that they could be rerun; for example, to enable new theories to be tested by modifying and re-running the software on the latest computing environment. At present, the possibilities were constrained by the need to maintain legacy systems, upon which such programs could be run and, consequently, a policy for the successful migration of programs from one computing environment to another was needed.
3.Finally, the results of analysis (and any intermediate data generated by the analysis) needed to be preserved for comparison with future analyses.

4.1.2 Labóratorio de Instrumentação e Física Experimental de Particulas

Although LIP, as a fundamental physics research institute, is very different from the applied technology interests of LNEC, the discussion there took a rather similar form. Initially, the participants noted that the previous day's presentations had been very useful: they were already aware that preservation was a problem for LIP, but the presentations had confirmed this and had given ideas for how the situation might be improved. For example: "*We don't have a digital preservation policy and it's what we need, so for us, anything we can learn is useful. It was very valuable and we have ideas from that*".

In further discussion, a number of issues arose: first, in some of its international collaborations (e.g., with the Large Hadron Collider experiments at CERN), there was no need to preserve the original data, since this was done through existing collaborative agreements. However, capturing the workflows of the experiments conducted by LIP researchers and preserving these, their associated data and software, was essential. Standards did exist in the organization but there was a lack of policy to enforce these standards across the different groups in the Institute.

The situation was similar to that in LNEC: individual researchers, like the engineers in LNEC, had a great deal of autonomy over how they conducted their experiments (which tended to consist of running programs they had developed themselves and what they did with the original data, the software and the intermediate data and results. Several obsolete machines were retained, although powered down, in the event that proprietary data format needed to be re-used. Re-use of old data or old programs arose when, for example, a new PhD student found a need to apply new theory to existing data analyses. This might involve revising the original program to explore new parameters in the existing data.

The participants, collectively, identified what was needed of a digital preservation policy:

1.The user of the data should be able to decide what should be preserved.
2.Any system for preservation should be "plug and play", involving little or no action on the part of the researcher.
3.Collegiality should be maintained: that is, the decisions over policy should be agreed by all groups operating in the Institute.
4.Any preservation system should have the capability of collecting the necessary data from wherever the researcher was undertaking the work, since they might be operating within the institute, at a partner organization elsewhere, or at home on their personal computer.

One participant noted that, "Looking at the analysis stage will give us 80% of the problem".

Asked about how their use of the grid affected the situation, it was noted that in using the grid, one would also be using whatever software is available there, and that whether using remote or local data, the analysis in both cases would be interactive.

4.2 Responses to the questionnaire

The first part of the questionnaire was intended to collect the data about the practices of data archiving and needs for preservation in respective institutions. None of the two organizations has developed a digital preservation policy, but both are concerned about preservation of their research data.

The LNEC collects and archives a wide range of data and data sets: sensor data, field data, satellite remote sensing data, road and accident data, dam monitoring data, mathematical and physical models data, etc. Though there is no policy that sets for which period the data is retained, the need for fairly regular and frequent re-use of the archived data was reported by five participants and two needed it very rarely and occasionally. Therefore, most of the data is stored "permanently", "for the life-time of the structure", "very long", for 50 or 30 years, though this mainly depends on the nature of the data. Some of it may be retained only for several months or deleted the same day. Six participants access the data from their PC and one uses a separate system to access files. Search capabilities in archiving systems are varied and not well developed, though two participants have identified three and four search options (file name and version, time range and storage place, product identifier, metadata). So far only one person has missed older data, which could have been helpful. Only two of the participants are involved in digital preservation work, one of them claims that he has developed personal digital preservation procedures over 30 years. His main activity is selecting and adopting digital preservation systems and software development. The other – information system developer – is involved in data curation, managing and providing digital preservation systems, software development.

The LIP also collects and archives a wide range of data: high energy physics datasets, CERN/LAC data, data from biomed and earth science, bitstream backups of the recorded user activities in home file system, etc. All the participants are acquainted with their archiving system, though some call it backup system rather than archiving one. There is not clear policy for how long the datasets should be kept, but the information system service tries to keep it as long as possible. It is up to research teams to decide when the data should be disposed off. Two participants need to use the archived data fairly regularly and one occasionally. Two participants access data from their PCs and one uses a separate system. The search capabilities are not very well developed: the data can be searched by file name and version, time range and storage place, or using once experience by scanning own repository. Two participants are very recently involved in digital preservation activities, which have been started this year. They are working with policy formulation and procedures, data appraisal, defining need for technology, data curation and management of IT services. One of them in addition is working on selection and adoption of digital preservation systems and software development.

Thus for both organizations and participants the SHAMAN presentation/training session and evaluation were relevant and they could provide informed input on the issue of SHAMAN DP framework.

The second part of the questionnaire sought the respondents' feedback on the presentation of SHAMAN digital preservation framework and its features. A separate question related to the importance of the SHAMAN reference architecture.

Six participants have rated the SHAMAN approach as very useful for their organisations and five as useful.

The relevance of particular digital preservation problems from the organizational perspective could be rated as high (3), moderate (2) and low (1). There was an option Unknown that was marked by one participant in relation to the issue of "Collection of collaborative work context data".

The highest ranking was given to the problems "Integration of long-term preservation into workflow" (average – 2,9), Search and access capabilities for preserved materials" (2,9), and "Re-use of preserved information" (2,9). "Collection and enrichment of data for preservation" followed the three most important problems closely (2,8), and "Collection of collaborative work context data" was rated 2,6.

Table 1: Importance of the SHAMAN features to the organizations

Features	Average rank
Distributed ingest	2,77
Data migration	2,63
Multivalent technology	2,44
Grid infrastructure	2,4
Authenticity validation	2,3

Table 1 shows the average ranking of the importance SHAMAN features for e-science organizations. As one could expect, distributed ingest and data migration were ranked highest, while authenticity validation does not seem to be of the same importance in the e-science environment.

As for the SHAMAN reference architecture the Table 2 shows the ranking of its main elements according to the importance to e-science. The highest importance is given to the possibility of identifying options to improve digital preservation capabilities, the least interest is raised by the relations with the stakeholders.

Table 2: Importance of the SHAMAN reference architecture elements to the organizations

Features	Average rank
Helps identifying options to improve your DP capabilities	2,99
to raise awareness of DP	2,9
… plan DP strategically	2,9
… identify flaws in your current approach	2,9
… identify the related stakeholders	2,33
… define the stakeholders responsibilities concerning DP	2,33

Almost all participants (except one) used the open questions on the expectations and improvements to the SHAMAN framework to express their suggestions. Over 20 suggestions were proposed, among them: provide data flow analysis, support dynamic datasets, provide easy to use triggers for users, integrate the concept of cooperating objects, support legacy systems, provide new types of support of data (cloud), disseminate information about the importance of digital preservation, find the ways to archive knowledge of staff, provide general and more specific education to scientists on data preservation, etc.

Some of the participants also expressed their appreciation of the SHAMAN event as very timely, good education for scientists about data preservation, interesting and useful event, etc.

This activity of participants shows not only the attitude towards SHAMAN event, but also the presence of active concern that has been touched upon by the training and evaluation.

5. CONCLUSION

The SHAMAN 'consciousness raising' event appears to have had useful consequences. In both institutions it appeared to fulfil the aim of drawing attention to the problems of digital preservation and to the role of the SHAMAN framework in finding answers to the problems. To a degree, the participants were already sympathetic to the ideas, since they were working in situations that daily brought their attention to the need for preservation policies. However, the discussions were also useful in extending the concept of digital preservation beyond the existing perceptions within the SHAMAN teams. In particular, the notion of the 'digital object' needs to be expanded to include not only digitized physical objects such as documents and images, but also 'born digital' phenomena such as data streams, software programs and analytical results of experiments.

7. Bibliography and References

SHAMAN (2008). D1.1 survey of users and providers in Europe and SHAMAN usage scenarios and DP support.

SHAMAN (2009). D14.1. Demonstration and Evaluation Plan.

SHAMAN (2009a). D11.1. Implementation of context capturing mechanisms within production environments and federated server architecture for temporary storage before ingests.

SHAMAN (2009b). D1.2. SHAMAN Requirements Analysis Report (public version) and Specification of the SHAMAN Assessment Framework and Protocol. [Access through the internet 2010-09-02: http://shaman-ip.eu/shaman/document]

SHAMAN (2009c). Project periodic report – Y1. Project objectives, progress and achievements. SHAMAN – WP18-D18.1 (Y1).

SHAMAN (2010). D11.1A. Implementation of context capturing mechanisms within production environments and federated server architecture for temporary storage before ingests. Addendum.

SHAMAN (2010a) Project periodic report – Y2. Project objectives, progress and achievements. SHAMAN – WP18-D18.1 (Y2).

SHAMAN (2011). SHAMAN Reference Architecture. WP13-D13.1.

LIST OF ABBREVIATIONS

ASCII **A**merican **S**tandard **C**ode for **I**nformation **I**nterchange
CERN The European Organization for Nuclear Research
D13.1 Deliverable 13.1
DoF SHAMAN Domain of Focus
DoF1 SHAMAN Memory Institutions Domain of Focus
DoF2 SHAMAN Industrial Design & Engineering segments Domain of Focus
DoF3 SHAMAN e-Science Domain of Focus
DP Digital Preservation
GPD Microsoft's Generic Printer Description files
INESC-ID   Instituto de Engenharia de Sistemas e Computadores Investigação e Desenvolvimento em Lisboa
ISP Integration and demonstration Sub-Project
IT Information Technology
LAC Laboratory of Atmospheric Chemistry
LIP Labóratorio de Instrumentação e Física Experimental de Particulas
LNEC Laboratório Nacional de Engenharia Civil
PC Personal computer
R&D Research and Development
RTD Research and Technical Development
WP Work package

8. ANNEXES

ANNEX 1: PRESENTATION AND FOCUS GROUP GUIDE

SHAMAN — Sustaining Heritage Access through Multivalent ArchiviNg	**Guide to presentation/training event and focus group discussion**

Introduction

A focus group is a qualitative research method that involves asking a group of people about their perceptions, opinions, beliefs and attitudes towards a product, service, concept, advertisement, idea, or packaging. The members of a focus group are selected as being relevant to the evaluation of a phenomenon under investigation. A group is engaged by a trained facilitator in a series of discussions designed for sharing insights, ideas, and observations on a topic of concern. The focus groups can also be used to subject ideas to review in order to determine their viability, usefulness or functional applicability.

Using a focus group method in relation to evaluation of the SHAMAN preservation framework and resulting technologic approaches would fall into the latter category of determining the viability of ideas, suitability and usefulness of technologies for certain functions in e-science institutions.

With this purpose in mind, the participants of the focus groups should be recruited according to the following criteria:

a) they should represent e-science institutions, i.e. research institutes, laboratories, research departments of universities;
b) they should be involved or interested in preservation activities in e-science on managerial or operational level, be acquainted with the preservation issues in general;
c) the experience of work with available archiving or preservation technologies would be an asset and should be taken into account when approaching potential participants.

Involvement of the potential focus group members in preservation activities should work as an incentive and motivation to take part in the evaluation. The text of invitation should emphasise the relevance of the presentation and focus group to their own work as well as to the goals of their organizations. Additional incentives could be distribution of presentation materials and research articles related to the project, promise of further involvement in the evaluation process and networking possibilities opened by participation in the SHAMAN

activities. Relation to and with other relevant projects that already gained reputation (e.g., PLANETS) should be mentioned as a part of invitation text. It is also advisable to have small souvenirs and certificates of participation in the SHAMAN project activities.

The participants should form a heterogeneous group in terms of organizational level (top or middle level managers), expertise, and professional interest in long-term digital preservation issues. They can be identified within the existing network of SHAMAN partners or through exploring information on e-science institutions engaged in long-term digital preservation activities. It is advisable to have 5-10 participants in one focus group as this is an optimal number for generating a lively, yet manageable discussion. SHAMAN focus groups should include not less than 4 members (a mini-group) having in mind the purpose of getting a wider span of opinions for feedback into the R&D activities.
The presentation events and evaluation exercise is lead by a team of persons from SHAMAN partner institutions.

The focus group members should be introduced to the achievements of the SHAMAN project for the e-science institutions (as they are at present). As SHAMAN is a very complex and innovative project, presenting these achievements is a very important part of the evaluation exercise. The presentations should be led by technologically savvy partners of SHAMAN. They also should be ready to answer the questions about the demonstrated SHAMAN framework and technologic approaches.

A focus group moderator should have proper moderating skills and experience of leading focus groups. Whenever possible, a moderator should be assisted at least by one person helping to take care of recorders, taking notes of the participants conduct, involvement, body language, and answers.

Participation of several people form SHAMAN partners in the presentation and focus group discussion requires drawing a line of conduct for them:

> 1. The presentation is given by designed persons and the rest of the team should not intervene into it, except when requested by presenters.
> 2. The focus group is moderated by one person asking questions and the rest of the team should not intervene with their questions and comments (*no matter what the group participants might say or the team members may think about needing to clarify issues*), except when requested by a moderator.
> 3. The team members should distribute the roles and functions beforehand and concentrate on them during the whole event.
> 4. The team members should support each other and help to exploit the focus group to the greatest benefit.

Conducting the demonstration event

1. Prepare the room: seating should allow participants view the presentation comfortably and then participate in a discussion without much change but positioned without anyone in dominating position. The focus group leader should avoid dominating position as well, but should have a good overview of all participants.
Provide paper and pencils for participants to be able to record their impressions, comments, questions, etc.
Find proper placements for presenters, observers, focus group leaders, so that they are comfortable but, especially, that their location does not interfere with the focus group participants' vision. Make sure that all of them have necessary equipment and materials for recording participants' reactions. Check that everyone of them knows their function and line of conduct. Discuss the division of the roles among the focus group leader, the presenters and the rest of the research group.
Check the computer, projector, connections.

2. Introduction
Explain what is the purpose of the meeting, why we have invited participants and what we expect from them. Draw their attention to the provided stationary, coffee, screens, recorders etc.
Most probably, it may be useful to apologize for and explain the usage of English in none English speaking countries (EU project, need to use their feedback by international group, etc.). It may help to put the participants at ease and make them less conscious of speaking a foreign language.
Make introduction round.
Distribute the questionnaire and explain its structure.
Ask to fill in the first part of the Questionnaire: previous involvement in preservation activities (QI.x-xx).

3. Presentation

I. Presentation has to be staged as fluently as possible with clear explanations and highlights. The main elements should be demonstrated, their functionality and technological novelty revealed and related to the participants work context. Emphasise the long-term preservation and communication with future features (whatever they are – minor or significant).

II. If there are more than one presenter, they should have clearly assigned role, script and defined interactions between themselves.

III. Questions from participants asked by the presenters. Specifically, request the participants to ask what they have not understood in the presentation, but refrain from any comments at this moment.
One person should record the questions and the reaction to the answers.
When the participants are satisfied with the clarifications and feel that they have understood the essence of the demonstration the focus group discussion starts.

Break

4. Focus group

Focus group leader reminds of the purpose of the meeting, sets the rules of discussion (e.g., asking question, having each participant's opinion round the table, free discussion, and move to another issue). Focus group leader has to ensure that every participant has a possibility and is drawn to express personal opinion, raise doubts, and comment on other participants' input.

The opening question:

I. What is your overall impression of the SHAMAN project results presented yesterday?
> Subsidiaries: *Did you find it useful/not useful?*
> *What was lacking?*
> *Did you see what you expected or there were unexpected things?*
> *What improvements would you suggest?*

III. Remember yesterday's presentations: what are the key features of SHAMAN framework, in your opinion?

II. Are you aware of the long term digital preservation policies in your organization?

If the answer is **yes**: Are you involved in formulating them yourself?
> Do you have any policies in your organizations that the framework would be relevant for?
> Subsidiaries: Can you give examples of preservation policies?
> Which demonstrated functions do you see as most useful in relation to your policies?

If the answer is **no**:
> Do you think the demonstrated framework could be still relevant to your organization?
> Might the demonstrated framework help you to develop preservation policies?
> If so, which demonstrated functions do you see as most useful for helping you towards shaping possible policies?

IV. What demonstrated methods fit in with existing work operations or procedures?
> Subsidiaries: What impressed you most in the proposed method?
> Where there functions or actions that you expected to see that appear to be missing?
> Does your institution already support the development of digital preservation tools or processes or use them? If yes, which ones?

V. What struck you as innovative in the demonstration?
> Subsidiaries: Would these innovative features be an improvement on existing technologies of which your are aware?
> What have you seen as innovative and would be useful to incorporate into your current practices?
> Would systems based on what you have seen presented serve your needs as users more effectively than your current strategies?

What do you think is missing at this point that would require inclusion in the preservation framework?
Could you suggest any other improvements?
What incentives could help to adopt SHAMAN framework?

V. If systems were developed based on SHAMAN would you be interested in applying them?
    Subsidiary: How well do the demonstrated approaches fit in with your existing systems?
      Sub-sub: *Would the framework extend your existing systems in any way? How?*

5. Ask the participants to fill in the second part of the Questionnaire (impressions from demonstration).
    Thank for the participation and good discussion.

ANNEX 2: QUESTIONNAIRES TO THE FOCUS GROUP PARTICIPANTS

Questionnaire

Dear participants,
We would be grateful if in addition to the discussion you could answer this questionnaire. It is divided in two parts. Please, fill in the first one at the beginning of the meeting.
The second part can be answered only after you have seen the presentation and participated in the discussion. Please, put it aside – your moderator will remind when to fill it.

Part I

1. Which organization do you work in? (please, tick the appropriate box)

LNEC   ☐
LIP   ☐

       What is your role or job in your organization?
       _____
       _____

       How long have you worked in the organization? _____ years

2. Are you familiar with information archiving systems used in your institution?

               No ☐   Not really ☐   Yes ☐

3. What e-science data sets do you collect and archive?
_____
_____
_____
_____
_____
–

4. For how long are these data sets archived?_____

5. Is there any policy that sets the time, for which the data sets are retained?

               No ☐   Yes ☐   Don't know ☐

6. How often do you need to access the archived data sets?

           Very rarely ☐ Occasionally ☐ Fairly regularly ☐ Often ☐

7. When you need to access the archived file sets does it require specific actions?
       .... access them directly from you own PC   ☐
       .... use a separate system to access the files   ☐
       .... need to ask the IT department to get the files for you   ☐

8. If you need access to older data sets, do you have search capabilities?

               No ☐   Yes ☐

       What are the search criteria you can use

I need to know file name and version ☐
I need to know the time range where it was stored ☐
I can search using metadata to identify the files ☐
I need the product identifier or number to start searching ☐
I know it be experience and scan my own repository ☐

9. Have you ever missed older data which would have been helpful for your work?

No ☐   Yes ☐

10. Does your organization have a *documented* digital preservation policy?

No ☐   Yes ☐

11. Are you personally involved in digital preservation work in the organization?

No ☐   Yes ☐

If YES,

How long have you worked on digital preservation matters? _____ years

What is your role in relation to digital preservation? Please tick appropriate functions.

☐ Formulating policies and/or procedures
☐ Appraising data for digital preservation
☐ Defining the organizational need for technology
☐ Data curation
☐ Selecting and adopting digital preservation systems
☐ Managing or providing IT services
☐ Software development
☐ Other (please specify) _____

Thank you for answering this part of the questionnaire.

Part II

1. Having attended the training session, how useful do you find the SHAMAN approach for your organization? *Please tick **one** box.*

Very useful	☐		Not at all useful	☐
Useful	☐		Not known	☐
Not very useful	☐			

2. Having attended the training, how would you rate the relevance of the following problems in general from the organizations' perspective?

Question	High	Moderate	Low	Unknown
2.1 Integration of long-term preservation into workflow				
2.2 Collection, and enrichment of data for preservation				
2.3 Collection of collaborative work context data				
2.4 Search and access capabilities for preserved materials.				
2.5 Re-use of preserved information.				

3. Please, rate the importance of the following features of SHAMAN infrastructure for your organization?

Features	Very important	Medium importance	Not important	Unknown
Distributed ingest				
Grid infrastructure				
Data migration				
Authenticity validation				
Multivalent technology				

4. Please, rate the importance of the SHAMAN reference architecture for your organization?

Features	Very important	Medium importance	Not important	Unknown
Helps to raise awareness of DP				
… to identify the related stakeholders				
…define the stakeholders responsibilities concerning DP				
… to identify flaws in your current approach				
… to identify options to improve your DP capabilities				
… to plan DP strategically				

5. Which other functions would you expected to find in a digital preservation in your environment?

_____
_____
_____
_____
_____
__

6. What improvements would you like to see implemented in the SHAMAN approaches?

_____
_____
_____
_____
_____
_____

ANY OTHER COMMENTS

_____
_____
_____
_____
_____
_____
_____
_____
_____

Thank you for answering our questionnaire and participating in the focus group.