Published in: International ACM Conference on Management of Emergent Digital EcoSystems (MEDES'09), Proceedings of the International Conference on Management of Emergent Digital EcoSystems, France 2009, ISBN:978-1-60558-829-2

Assessing Digital Preservation Frameworks: the approach of the SHAMAN project

Perla Innocenti, Seamus Ross
HATII at the University of Glasgow
11 University Gardens
G12 8QQ UK
+443304453
P.Innocenti@hatii.arts.gla.ac.uk

Elena Maceciuvite, Tom Wilson
Swedish School of Library and Information Science
University of Borås, 50190
Sweden
+ +46334354000
Elena.Maceviciute@hb.se

Jens Ludwig, Wolfgang Pempe
Goettingen State and University Library
Papendiek 14
37073 Goettingen, Germany
+495513912121
ludwig@sub.uni-goettingen.de

ABSTRACT
How can we deliver infrastructure capable of supporting the preservation of digital objects, as well as the services that can be applied to those digital objects, in ways that future unknown systems will understand? A critical problem in developing systems is the process of validating whether the delivered solution effectively reflects the validated requirements. This is a challenge also for the EU-funded SHAMAN project, which aims to develop an integrated preservation framework using grid-technologies for distributed networks of digital preservation systems, for managing the storage, access, presentation, and manipulation of digital objects over time. Recognising this, the project team ensured that alongside the user requirements an assessment framework was developed. This paper presents the assessment of the SHAMAN demonstrators for the memory institution, industrial design and engineering and eScience domains, from the point of view of user's needs and fitness for purpose. An innovative synergistic use of TRAC criteria, DRAMBORA risk registry and mitigation strategies, iRODS rules and information system models requirements has been designed, with the underlying goal to define associated policies, rules and state information, and make them wherever possible machine-encodable and enforceable. The described assessment framework can be valuable not only for the implementers of this project preservation framework, but for the wider digital preservation community, because it provides a holistic approach to assessing and validating the preservation of digital libraries, digital repositories and data centres.

Categories and Subject Descriptors
H.4 [INFORMATION SYSTEMS APPLICATIONS]

General Terms
Management, Measurement, Documentation, Performance, Design, Economics, Reliability, Security, Standardization, Legal Aspects, Verification.

Keywords
Digital preservation, Digital preservation frameworks, Policy frameworks, Web based Digital Ecosystems, SHAMAN project, Assessment criteria, Data grids, Digital libraries, Digital repositories, Data centres, Persistent archiving.

1. INTRODUCTION

1.1 Context of this work

Digital libraries (together with digital repositories and data centres) represent the confluence of vision, mandate and the imagined possibility of content and services constructed around the opportunity of use. Underpinning every digital library is a policy framework. It is the policy framework that makes them viable - without a policy framework a digital library is little more than a container for content. Even the mechanisms for structuring the content within a traditional library building as container (e.g. deciding what will be on what shelves where) are based upon policy. Policy governs how a digital library is instantiated and run; a digital library without policy therefore is similar to a Ferrari in a world without roads and populated only by blind drivers. The policy domain is therefore a meta-domain which is situated both outside the digital library and any technologies used to deliver it, and with in the digital library itself. That is, policy exists as an intellectual construct, that is deployed to frame the construction the digital library and its external relationships, and then these and other more operational policies are represented in the functional elements of the digital library. Policy permeates the digital library from conceptualisation through to operation and needs to be so represented at these various levels.

As Ross [1] reported, among the nine core research domains in the current Digital Preservation Research Agenda, automation lies at the heart of long term management of digital objects. The investigation of the implementation of policies as automated processes within digital library systems is essential as it forms a crucial component of these systems. The impact of automated policies on the digital objects (e.g. data) and the processes performed on or with this data influences the long term viability of these resources. So while it is easy to define digital object handling policies, the effectiveness of these systems must be measured and the change overtime in them must be anticipated.

The SHAMAN project is delivering a preservation framework and a central element of this is delivering automation of policies. The measurement of the effectiveness of the framework itself and of the systems constructed with the framework is essential. This paper describes SHAMAN and the mechanisms that we have developed to provide for the evaluation of the framework itself and of systems constructed with it.

The results of this investigation have been presented in detail within the second part of the project deliverable *SHAMAN Requirements Analysis Report (public version) and*

Published in: International ACM Conference on Management of Emergent Digital EcoSystems (MEDES'09), Proceedings of the International Conference on Management of Emergent Digital EcoSystems, France 2009, ISBN:978-1-60558-829-2

Specification of the SHAMAN Assessment Framework and Protocol [2].

1.2 The SHAMAN Project

The SHAMAN (Sustaining Heritage Access through Multivalent ArchiviNg) is a Integrated Project [3], part of the European Union's 7th Framework Programme. The aim of this project is to investigate the long-term preservation of large volumes of digital objects in a distributed environment, by developing a preservation framework that is verifiable, open and extensible. Our research addresses digital preservation from ingestion to dissemination in an environment where the collections, producers, consumers and curators are geographically distributed and the content of the collections is of a dynamic nature.

SHAMAN is developing associated preservation tools for analyzing, ingesting, managing, accessing and re-using digital objects across libraries and archives. Three prototypical applications will support evaluation and validation of the results in memory institutions, in industrial design and engineering settings and in the domain of e-science.

To achieve this aim, the project is investigating data grid, digital library, persistent archive and information knowledge and content representation technologies, to create preservation system prototypes that characterize the preservation process in ways that make it feasible to replace preservation services without impact upon the digital objects, or access to it and re-use of it (Figure 1).

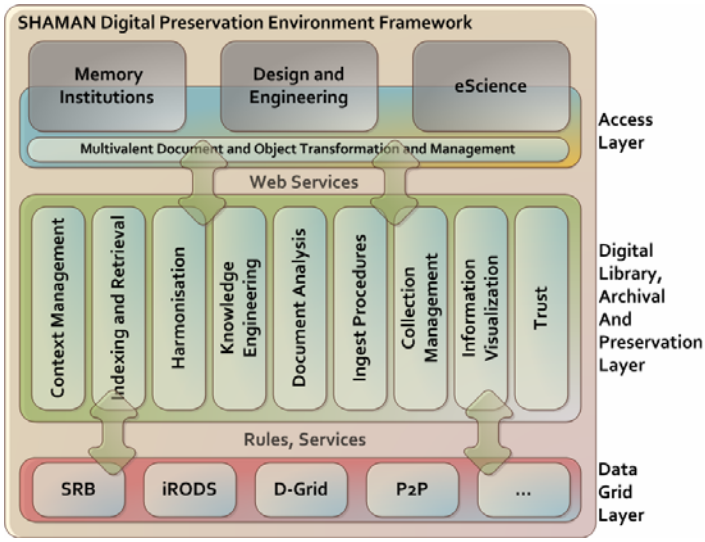

Figure 1. The envisioned SHAMAN Framework

SHAMAN's Digital Preservation Framework is based upon a combination of technological, organizational, and R&D methods. This paper introduces the main principles applied for evaluation of the integrated elements of the SHAMAN framework from the point of view of user's needs within all three domains of focus.

2. ASSESSMENT OF A PRESERVATION FRAMEWORK

2.1 Project goals and outputs to be assessed

The SHAMAN project seeks to:

- provide a vision and rationale to support a comprehensive Theory of Preservation that may be used to develop systems for the storage of and access to any type of digital objects, based on the integration of digital library, persistent archive, and data management technologies;
- supply an infrastructure that provides expertise and support for users requiring the preservation and re-use of digital objects; and,
- develop and implement a grid-based production system, which will support the virtualization of digital objects and services across archival, scientific and engineering domains.

This hierarchy of tasks determines the levels of evaluation of the project and outputs.

2.2 Assessing criteria derived from TRAC and DRAMBORA

SHAMAN will deliver a preservation framework, which will supply the infrastructure for users requiring long-term preservation services, and which will develop and implement a grid-based production system to support the virtualisation of digital objects and services in a variety of user domains. The process of evaluation of its success might therefore be considered in terms of the benchmarking and risk assessment tools that have been proven in other projects and initiatives.

As a crucial part of this process the two noted TRAC (Trustworthy Repositories Audit & Certification: Criteria and Checklist) [4] and DRAMBORA (Digital Repository Audit Method Based on Risk Assessment) [5] were mapped to iRODS (i Rule Oriented Data Systems) [6] and the objectives of the SHAMAN work packages.

The Trustworthy Repositories Audit and Certification (TRAC) Criteria and Checklist is configured as a checklist, meant to help institutions objectively to evaluate responsibilities against capabilities and identify potential risks to digital content. TRAC takes OAIS (Open Archival Information System) [7] its foundation, and the benchmark for measuring success in terms of trustworthiness.

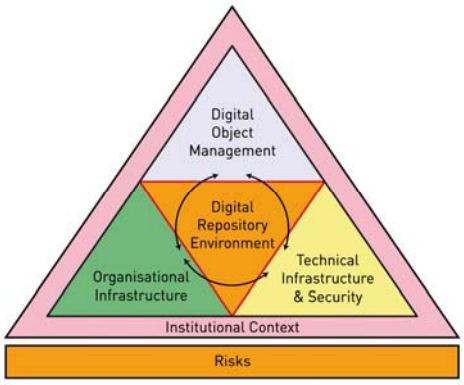

Figure 3. DRAMBORA: Interrelationships within a digital repository environment © HATII at the University of Glasgow

The main goal of DRAMBORA, which is an interactive online support management tool at repository level, is enabling evidence-based risk management for digital repositories. The DRAMBORA assessment process focuses on risks, and their

Published in: International ACM Conference on Management of Emergent Digital EcoSystems (MEDES'09), Proceedings of the International Conference on Management of Emergent Digital EcoSystems, France 2009, ISBN:978-1-60558-829-2

classification and evaluation according to individual repositories' activities, assets and contextual constraints (Figure 3), to determine a particular repository's ability to contain and avoid the risks which might threaten its ability to receive, curate and provide access to authentic and contextually, syntactically and semantically understandable digital information.

As a result of the evaluation conducted as part of this work, we concluded that it would be appropriate to set as a target evaluation criteria the question: Will we be able to prove that a system designed and deployed according to the SHAMAN framework properly supports the TRAC/DRAMBORA rules and criteria? This evaluation will be possible when the 'integrative sub-projects' of SHAMAN have been delivered and are available to test against the TRAC/DRAMBORA criteria.

2.3 Assessing criteria derived from iRODS rules

SHAMAN will use Integrated Rule-Oriented Data System (iRODS™) data grid technology as a storage substrate for digital preservation.

iRODS implements rules for grid-based data management. The whole process of digital objects ingestion, manipulation, access and use can be managed over a grid-based system through the application of the "iRODS Rule Engine". Consequently, the nature of the rules and the ability of digital preservation systems to operate under those rules will guide the development of the SHAMAN preservation framework. Evaluating whether or not the individual elements of the framework satisfy this requirement will be a feature of the assessment process.

Finally, the individual work packages of SHAMAN will produce a variety of outputs, including software and conceptual schemes, designed for grid-based operation. Evaluation of the outputs will be according to the general criteria for successful information systems development and should be tested against the iRODS rules.

2.4 Assessing criteria derived from information system models

The outputs of the individual work packages will be assessed using information systems success criteria developed by DeLone and McLean [8,9]. The framework proposed by DeLone and McClean [8,9] is widely accepted as appropriate for the evaluation of information systems. We intend to use their original model, rather than the more recently developed one, because the latter appears to be related more closely to commercial systems applications such as e-commerce systems.

Each of the areas in the model has associated evaluation criteria:

1. System quality criteria.
2. Information quality criteria
3. Use criteria
4. User satisfaction
5. Individual impact
6. Organizational impact
7. Essential properties

Another set of evaluation criteria for information systems can be derived from the IEEE's *Recommended practice for software requirements specifications* (1998) [10]. The criteria given are intended, as the title suggests, providing guidance on writing specifications. However, some also relate to the evaluation of software. Thus, the basic issue addressed in writing requirements can be directly related to the SHAMAN outputs:

1. Functionality
2. External interfaces
3. Performance
4. Attributes
5. Design constraints imposed on an implementation:

Having in mind the complexity of the SHAMAN project outputs, the team working on the Assessment Framework decided to build the evaluation process on a combination of the approaches described above. We recognize, however, that evaluation of project activities will be formative, rather than summative as we envisage a process of continual refinement and further evaluation of the component elements of the framework.

3. ASSESSMENT PLAN

3.1 Key Performance Indicators

All SHAMAN Work Package leaders were invited to indicate and describe relevant assessment criteria and Key Performance Indicators (KPIs) for their WP, according to the following fields:

- Title of the KPI
- Definition of the KPI
- Measurement criteria for the KPI
- Target to be achieved.

Information systems may be interpreted as having three dimensions: a technical dimension, a management dimension and a user dimension and assessment criteria are needed for all three, in order to provide a full evaluation of the systems produced within the framework. The feedback from the project's partners [3] was analysed in the light of the combined approaches and groups of criteria (Technical, Management, and User-related KPIs) were defined for the evaluation process.

4. MAPPING TRAC, DRAMBORA AND iRODS

Since the SHAMAN data grid implementation will build on iRODS, a set of iRODS rules can be used to ensure the most comprehensive compliance with the various criteria/requirements defined by DRAMBORA and TRAC. These rules are small units of software, which can execute server-side operations, so called micro-services, in several ways (e.g. triggered by certain events, under certain conditions, manually, or according to a user-defined time schedule). Policies especially for digital objects and data management can therefore be expressed by iRODS rules. Reagan Moore and Adil Hasan have composed a set of rules to enable/support the compliance of an iRODS-based repository with the TRAC criteria [11] .

In a second step, these rules have been prototypically assigned to DRAMBORA risk mitigation strategies and transferring the iRODS rules of a TRAC criterion to its corresponding DRAMBORA risk. The SHAMAN assessment framework has to verify the implementation and proper functioning of the complete set of rules.

The assessment workflow as derived from Moore [12] can be described in six steps:

1. Definition of assessment criteria (in our case: TRAC, DRAMBORA)

Published in: International ACM Conference on Management of Emergent Digital EcoSystems (MEDES'09), Proceedings of the International Conference on Management of Emergent Digital EcoSystems, France 2009, ISBN:978-1-60558-829-2

2. Definition of policies enforcing the assessment criteria
3. Definition of rules that apply the policies (iRODS rules)
4. Definition of capabilities that implement the required (preservation) functions (microservices)
5. Definition of (preservation) metadata that capture information about the application of the preservation functions (persistent state information, e.g. audit trails)
6. Query the (preservation) metadata to assess whether the assessment criteria have been satisfied.

Step 1 has already been performed by choosing the TRAC, DRAMBORA and information systems' success criteria as assessment criteria for SHAMAN, steps 2 and 3 by developing the TRAC-iRODS and DRAMBORA-iRODS mappings [3].

Steps 4 and 5 will be performed by the responsible workpackages and can be seen as project outputs that are the elements of the SHAMAN technologies.

Step 6 is the actual process of performing the assessment. It can only be performed in the context of an implementation of the SHAMAN preservation framework, i.e. a running system like the demonstrators to be developed by the ISPs.

Since iRODS is a data management system, the implementation of iRODS rules focuses on management and technical aspects. Organisational and financial aspects as also addressed by the TRAC and DRAMBORA checklist can be supported, but only on the level of digital objects management. This restriction may be acceptable because SHAMAN is developing a preservation framework and not a ready-to-use preservation system. It is not within the scope of the project to develop business plans, mission statements etc. Cognizant of these limitations, the implementation of TRAC/DRAMBORA-related iRODS rules is a convenient instrument for self-assessment and an important component of the SHAMAN framework for enabling the building of trustworthy repositories.

A preliminary attempt has been made [3] to classify the possible contributions of the several workpackages to the implementation of the TRAC criteria and DRAMBORA risk mitigation strategies and their corresponding iRODS rules.

Some SHAMAN workpackages are developing functions that can enable the implementation of the rules. While almost all workpackages contribute in one way or another, by defining or enabling features needed to execute policy-related rules (e.g. definition of appropriate metadata), some workpackages are directly involved in implementing the rules and developing the related micro-services such as:

- Automation of preservation management policies
- Data grid implementation
- Provision of rules and services/micro-services for automating advanced management policies
- Characterising the management policies that are needed to enforce authenticity, integrity, access restriction, digital objects placement.

The first implementations of the SHAMAN preservation framework will be carried out by three workpackages. They can therefore be regarded as *test-beds for the rule-based self-assessment*, as well as a starting point for evaluating the related KPIs.

If organisational criteria have to be addressed, domain-specific blueprints for mission statements, business plans etc. may possibly be provided/discussed by the dissemination workpackages.

The complete criteria-rule mappings with a tentative assignment to certain workpackages and to which extent the criteria can be implemented have been listed [3]. Based on these mappings, additional KPIs can be derived that measure the extent to which the fully implementable and the partially implementable criteria and risk mitigation strategies are implemented.

5. CONCLUSION

SHAMAN is a project involving eighteen partners in Europe and North America from both private and public sectors. The research conducted in the first workpackage innovates in the underlying approach to defining and validating the SHAMAN preservation mechanisms as well as in terms of how it enriches our knowledge about the characteristics that preservation systems must have. An assessment framework has been produced within this workpackage, providing assessment criteria for the whole project, for individual work packages and for outputs. While the work on the user requirements and on the Assessment Framework was defined with the work in the subsequent phases of analysis, design and implementation in mind, they will be invaluable to others who may be developing preservation systems as they provide an example of well-founded and validated requirements. Through the public release of this deliverable we hope to assist other preservation initiatives.

6. ACKNOWLEDGMENTS

SHAMAN (Sustaining Heritage Access through Multivalent ArchiviNg) Integrated Project is co-funded by the European Union (Grant Agreement No. ICT-216736).

7. REFERENCES

[1] Ross, S. Digital Preservation, Archival Science and Methodological Foundations for Digital Libraries, Keynote Speech at the European Conference on Research and Advanced Technology for Digital Libraries (ECDL) 2007, Budapest, Hungary, 17 September 2007. http://www.ecdl2007.org/Keynote_ECDL2007_SROSS.pdf

[2] Innocenti, P. Aitken, B. Hasan, A. Ludwig, J. Maciuvite, E. Barateiro, J. Antunes, G. Mois, M. Jäschke, G. Pempe, W. Wilson, T. Hundsdoerfer, A. Krandstedt, A. Ross, S. 2009. SHAMAN Requirements Analysis Report (public version) and Specification of the SHAMAN Assessment Framework and Protocol, SHAMAN Project. http://shaman-ip.eu/shaman/document

[3] Sustaining Heritage Access through Multivalent ArchiviNg (SHAMAN) Project website, 2009. http://www.shaman-ip.eu

[4] CRL, 2008. Trustworthy Repositories Audit and Certification (TRAC) Criteria and Checklist, Centre for Research Libraries, Chicago. http://www.crl.edu/content.asp?l1=13&l2=58&l3=162&l4=9

[5] DigitalPreservationEurope & Digital Curation Centre, 2008. DRAMBORA Interactive: Digital Repository Audit Method Based on Risk Assessment. http://www.repositoryaudit.eu/

[6] iRODS Integrated Rule-Oriented Data System. https://www.irods.org/

[7] Consultative Committee for Data Space Systems, Reference Model for Open Archival Information Systems (OAIS), CCSDS, January 2002 http://public.ccsds.org/publications/archive/650x0b1.pdf

Published in: International ACM Conference on Management of Emergent Digital EcoSystems (MEDES'09), Proceedings of the International Conference on Management of Emergent Digital EcoSystems, France 2009, ISBN:978-1-60558-829-2

[8] DeLone, W.H. & McClean, E.R. Information systems success: the quest for the dependent variable. Information Systems Research, 3(1), 60-95. (1992).

[9] DeLone, W.H. & McClean, E.R. The DeLone and McLean model of information systems success: a ten-year update. Journal of Management Information Systems, 19(4), 9-30 (2003).

[10] Institute of Electrical and Electronics Engineers, 1998. IEEE recommended practice for software requirements specifications, Institute of Electrical and Electronics Engineers. (IEEE Std. 830-1998), New York, NY.

[11] The TRAC-iRODS mapping tables were conceived, designed and created by Reagan Moore (DICE, University of North Carolina at Chapel Hill) with contributions from Adil Hasan (University of Liverpool). Moore and Hasan provided SHAMAN WP1 partners with this TRAC-iRODS mapping for a further analysis as part of the investigations of this deliverable.

[12] Moore, R. (2008). Towards a Theory of Digital Preservation. The International Journal of Digital Curation. 1(3), pp. 63-75. http://www.ijdc.net/index.php/ijdc/article/view/63/42