Long-tem digital preservation in e-Science domains

- Risk Management for digital preservation -

e-Science

Scientists are facing an eminent data deluge, which imposes several challenges on the way that data is managed and analyzed. Several communities like in biology, medicine, engineering or physics, manage large amounts of scientific information. It usually includes large datasets of structured data (e.g., data captured by sensors), physical or mathematical simulations, and several highly specialized documents reporting the work and conclusions of researchers.

Usually, scientists make use of isolated information systems to produce, manage and exploit large amounts of heterogeneous data. The absence of integrated solutions not only imposes several barriers to scientists conducting their research activities, but also limits future achievements on the analysis of unrepeatable data.

These required collaborative environments for scientific communities and associated services and infrastructures, usually known as e-Science (or enhanced Science) [1], involves the requirement of interoperability and the respective data sharing. In a broad sense, e-Science concerns the set of techniques, services, personnel and organizations involved in collaborative and networked science. It includes technology but also human social structures and new large scale processes of making science. It also means, on the same time, a need and an opportunity for a better integration between science and engineering processes. Thus, long-term preservation can be thought as a required property for future science and engineering, to assure communication over time, so that information that is understood today is transmitted to an unknown system in the future.

For example, the safety of large civil engineering structures like dams, bridges or nuclear facilities, require a comprehensive set of efforts. Typically, we must consider the structural safety; the structural monitoring; the operational safety and maintenance; and the emergency planning [2]. The consequences of failure of one of these structures may be catastrophic in many areas, such as: loss of life; environmental damage; property damage (e.g., dam flood plain); loss of power plant and electricity production; socioeconomic impact; political impact, etc.. These risks can be reduced by a number of structural measures, allowing the detection of abnormal behaviors. Recently, a substantial technical and financial effort was made in order to implement or improve automatic data acquisition systems that are able to perform real-time monitoring and trigger automatic alarms. This paradigm accomplishes an imminent deluge of data captured by automatic monitoring systems (sensors) and generated by large mathematical simulations (theoretical models). Along the fact that these

Dagstuhl Seminar Proceedings 10291
Automation in Digital Preservation
http://drops.dagstuhl.de/opus/volltexte/2010/2769

monitoring systems can save lives and protect goods, they can also prevent costly repairs and help to save money in maintenance.

As a first assumption, one can consider that the main reason to preserve data is to preserve its value. Consequently, it does not make sense to preserve invaluable data. However, determine and assess the value of data is a difficult and error-prone task.

Moreover, it could be an error to consider that data that cannot be used today will have no value in the future. For instance, today's grid infrastructures allow the simulation of mathematical models with a much higher resolution and volume of simulated data.

In [3], the authors discuss the major reasons to preserve data in the generic context of e-Science. Limiting the context to the example of the monitoring of large civil engineering structures, we consider the main motivations as:

- Retention of unique observational data, which is impossible to recreate and crucial to assure the structural safety of the monitored structures.

- Compliance with legal requirements (in the case of concrete dams), or compliance with contracts and/or service level agreements established with the owners of the structures.

- Retention of expensively generated data (e.g., generated by mathematical simulations) which is cheaper to maintain than re-generate.

- Re-use of data for new research (e.g., development of new theoretical models that better approximate the real behavior of the structure within its entire life-cycle).

It is important to stress that the ultimate motivation of the preservation of data in the scope of large civil engineering structures is to provide "tools" for a better structural safety assurance, avoiding or reducing the potential consequences of undetected and unpredicted structural anomalies.

Data Grids

An already common technology to handle e-Science collaboration and data management, for scenarios like large civil engineering structures' control, is the use of data grids [4]. Data grids such as iRODS [5] are able to manage large digital objects and use middleware that makes file management, user management and networking protocols transparent. However, the daily increase of data and potential changes on components, require an adaptation of current strategies and policies. In other words, an optimal strategy for an actual version of a collection may no longer be an acceptable option in the future, due to changes in components or the increasing size of the collection.

Risk Management

The problem of the long-term preservation of digital assets can be seen as a challenge of Information Security. As a matter of fact, information security is all about protecting and preserving the confidentiality, integrity, authenticity, availability, and reliability of information assets. The ISO/IEC 27001 [6] is the information security management standard created to help organizations to establish and maintain an information security management system. The application of this standard depends on several factors as, for example, the goals, business processes, size and structure of specific organizations.

Risk Management is a prominent arena of information security, whose ultimate goal is to define prevention and control mechanisms to address the risk attached to specific activities and valuable assets, where risk is defined as the combination of the probability of an event and its consequences (ISO/IEC Guide 73, 2002) It is recognized that Risk Management is concerned with both positive and negative consequences of risks.

The Risk Management Standard (ISO/FDIS 31000, 2009) defines the principles and implementation of Risk Management to control the behaviour of an organization with regard to risk, and is based on the principle that Risk Management is a process operating at different levels, as shown in Figure 1. The Risk Management process encloses the limitation of the context, risk assessment (identification, analysis and evaluation of risks) and risk treatment. This process requires a continuous monitor and review activity to audit the behaviour of the whole environment allowing, for instance, the identification and treatment of an unexpected vulnerability.

First, defining the context is crucial to identify strategic objectives and define criterions to determine which consequences are acceptable to this specific context. Second, today's organizations are continuously exposed to several threats and vulnerabilities that may affect their normal behaviour. The identification, analysis and evaluation of these threats and vulnerabilities are the only way to decide on the appropriate techniques to handle them. The identification of threats, vulnerabilities and risks is based on events that may affect the achievement of goals identified in the first phase. After that, the risk analysis and evaluation estimates the likelihood and impact of risks to the strategic goals, in order to be able to decide on the appropriate techniques to handle these risks (Treat Risks).

Currently, the digital preservation arena uses Risk Management concepts to assess repositories. The Trustworthy Repositories Audit and Certification - TRAC Criteria and Checklist[1] is meant to identify potential risks to digital content held in repositories. It takes OAIS as its intellectual foundation, and as the benchmark for

[1] The TRAC checklist is available at http://www.crl.edu/PDF/trac.pdf

measuring success in terms of trustworthiness. It establishes appropriate methodologies for determining the soundness and sustainability of digital repositories.

The Digital Repository Audit Method Based on Risk Assessment – DRAMBORA [7] process focuses on risks, and their classification and evaluation according to individual repositories' activities, assets and contextual constraints.

Risk Management can be used to assess existing solutions, but also to conceive digital preservation environments, enclosing three main steps: (*i*) establish digital preservation requirements (context and strategic objectives); (*ii*) identify digital preservation risks, based on vulnerabilities and threats, and (*iii*) treat the risks by addressing digital preservation threats and vulnerabilities.

SHAMAN project

The SHAMAN (Sustaining Heritage Access through Multivalent Archiving) project is funded under 7th Framework Programme of the EU under the contract 216736. It includes an e-Science work package to apply the SHAMAN technologies and use data grids to support the federation of preservation environments to scientific domains. Some of the challenges to be addressed include the use of description languages to represent structural, spatial and temporal relation relationships inherent within a digital entity; and the use of integrated visualization tools to interpret digitally preserved data.

References

[1] S. Miles, S. Wong, W. Fang, P. Groth, K.-P. Zauner, L. Moreau, Provenance-based validation of e-science experiments, Web Semantics

5 (1) (2007) 28–38.

[2] M. Wieland, R. Mueller, Dam safety, emergency action plans and water alarm systems, International Water Power and Dam Construction.

[3] P. Lord and A. Macdonald. E-Science curation report – Data curation for e-Science in the UK: an audit to establish requirements for future curation and provision, 2003.

[4] W. E. Johnston, Computational and data grids in large-scale science and engineering, Future Gener. Comput. Syst. 18 (8) (2002) 1085–1100.

[5] A. Rajasekar, M. Wan, R. Moore, W. Schroeder, A prototype rule- based distributed data management system, HPDC workshop on Next Generation Distributed Data Management.

[6] ISO, ISO/IEC 27001. Information security standard. (2005).

[7] McHugh, A., Ruusalepp, R. Ross, S. & Hofman, H. (2007). The Digital Repository Audit Method Based on Risk Assessment (DRAMBORA). DCC and DPE, Edinburgh. 2007.