The Role of Risk Analysis to Support Cost Models for Digital Preservation

Diogo Proença (1), José Borbinha (1), Neil Grindley (2)

(1) IST/INESC-ID, Lisbon, Portugal

(2) JISC Digital Preservation & Records Management, United Kingdom

Abstract

Cost can be viewed as the amount required to resource inputs into an activity, and there will usually be a need to evaluate these inputs against the outcomes or benefits. Cost modeling techniques exist in many areas, to help to calculate and anticipate the costs of a given activity.

However, there are areas where there is still insufficient knowledge about potential costs, and this is relevant for the domain of digital preservation. Particularly in this area, costs can only be understood in relation to benefits, and the benefits of investing in digital preservation have to be assessed against the potential threats that organizations face. For example, an analysis of the risks of format obsolescence or of a failure of business continuity should motivate organizations to undertake digital preservation when it becomes evident that the potential loss of value of not doing so overcomes the cost.

In order for cost models to be useful, it is critical that they are linked to an analysis of the risks in the given domain. Risk analysis should be the foundation of cost modeling so that all costs can be traced back to a specific set of threats that are applicable to the relevant domains and contexts. For some domains where extensive risk analysis has already been done and the benefits are well defined, this may be fairly straightforward. For other domains it might be more difficult. That is the case of digital preservation, where the costs and benefits are not currently widely and well understood but there is also anxiety about not engaging with digital preservation. Many organizations struggle to understand its value and the reason for this is that there is an insufficient focus on the role of risk in decision-making and preservation planning.

The European Commission's FP7 funded 4C project aims to address a number of issues that relate directly to the cost determinants of digital preservation, one of the most important being an assessment of risk. The objective is to define a clearer economic landscape within which organizations can operate with confidence and where commercial and community-driven organizations can provide solutions that are reliable and effective, but also economic, timely and sustainable. Large memory and archiving institutions are dealing with an ever increasing amount of digital data and there is a concomitant need for scalable and effective solutions and services to emerge that enable them to tackle that challenge. Medium and smaller organizations, as also entities in the private sector, face different types of challenges and a diversity of solutions are required. The cost of digital preservation needs to be understood through a number of different lenses. Benefits, value and sustainability are three different perspectives that must influence cost but a fundamental understanding of risk – and risk in relation to organizational mission – is of crucial importance. This can bring a new view to digital preservation, making it possible to analyze scenarios and take decisions based on perceptions of objective added value, and not only as objective costs and subjective values.

Introduction

Organizations throughout their operation encounter different influences and factors which bring uncertainty to the achievement of their goals. The uncertainty these different factors and influences bring on organization goals or objectives is called risk. In this way, Risk can be defined as the "effect of uncertainty on objectives" [1]. This effect can either be positive and/or negative and is regarded as a deviation from the expected objective.

An organization has associated risks in all its activities, and these risks must be identified, analyzed, evaluated and treated. The evaluation phase evaluates whether a risk should be modified, using risk treatment techniques in order to satisfy the risk criteria. There are several options to treat the analyzed risks as the ones defined in [1]. All these phases together constitute the risk management process depicted in Figure 1.

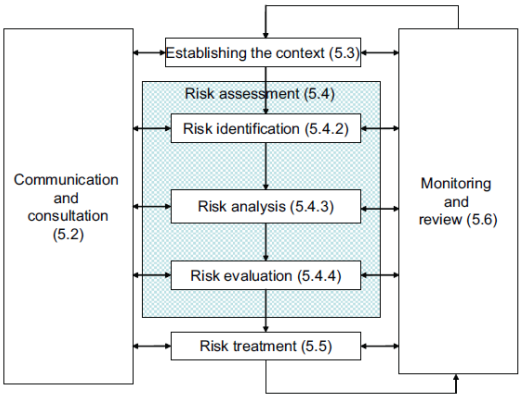

Figure 1. *Risk Management Process [1]*

The risk management process in Figure 1 begins by establishing the context where the organization defines its objectives and the external and internal parameters that should be taken into account in the risk management process while setting the scope and defining the risk criteria which will be used throughout the process. This step includes three phases, the establishment of the external context, the internal context and finally the context of the risk management process.

This step is followed by the risk assessment phase where is performed the risk identification, analysis and evaluation. Finally, there is the risk treatment phase where the risk is finally modified.

During all the phases described previously there should be communication and consultation with internal and external stakeholders and there should also be monitoring and review of the process.

In the end of the whole risk management process we end up with the identified risks associated with the proper treatment to each risk. The risk management process can be applied to the whole organization or to a department or even a process.

Following this line of thought, risk management can be of valuable importance to support the creation of cost models which can be applied to myriad of problems that organizations face, as for example digital preservation.

A cost model is used to estimate the cost of a certain task or service and provides a framework where all costs can be recorded and allocated to organizations.

Digital preservation is a problem that has been widely recognized as a challenge, motivated by the obsolescence of technology and all associated digital objects which might endanger the maintenance of valuable assets over time. Although digital preservation has been mainly a problem faced by memory and cultural heritage institutions, it is also of relevance to virtually all organizations that have to manage information over time.

These organizations often already have information systems which are used for processing and managing information, and a separate system for preservation is not desirable. In these scenarios Digital Preservation is seen as a valuable property of information systems, and not as the main source of requirements. Despite this shift, the main goal of digital preservation is still intact which ensures that information that is understood today can be transmitted into a system in the future and still be correctly understood. Despite the traditional repository scenario there is an alternative scenario that should be considered, where digital preservation is seen as a capability that can be added to existing systems [2], as depicted in Figure 2.

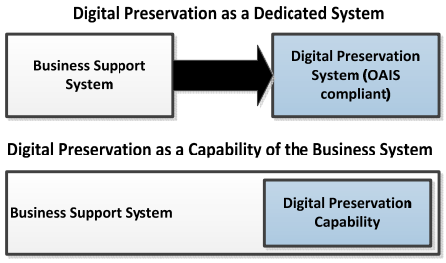

Figure 2. *A Model of Digital Preservation Scenarios [2]*

The first scenario is the traditional scenario as targeted by Open Archival Information System (OAIS) [3] (mainly motivated by the cultural sector); while the second case is a scenario that is believed will emerge soon in many areas of the activities, mainly in the corporate world.

This paper is structured as follows, first we present the related work in the cost modeling domain applied to digital preservation. Then we will formulate the connection between cost modeling and risk management, explain the synergies between the two domains and present the 4C project which will define and study these synergies. Finally, we conclude the paper by formulating some questions as future action points for research.

Related Work

There are innumerous examples of cost models developed either for digital preservation or which can be used for digital preservation as the case of cost models for digital storage or data centers. Some examples are the Cost Model for Digital Preservation (CMDP) [4][5], developed by the Royal Danish Library and the Danish National Archives, the Total Cost of Preservation (TCP) developed by the California Digital Library [6] or the Cost Model for Small Scale Automated Digital Preservation Archives developed by S. Strodl and A. Rauber [8].

The CMDP aims at estimating the costs of digital preservation by using a tool which calculates the present and future costs of digital collections based on users' inputs, which include the type and amount of data. CMDP has three developed phases which are, (1) Preservation Planning and Digital Migrations, (2) Ingest and (3) Archival Storage. The first phase is concerned with costing the activities regarding preservation planning and digital migrations. The second phase is concerned with the costing of the ingest activities and is based on the [3]. Finally, the third phase deals with the costing of the archival storage entity of OAIS. This cost model was part of a project which aim was to develop a generic model for the estimation of digital material preservation costs. It began by studying existing cost models namely the LIFE Costing Model [9] and KRDS 1 [7] but later concluded the existing cost models could not be used for the project purpose and so the CMDP was developed.

The TCP cost model is based on the assumption that digital assets underpinning web-based commerce, science, education and entertainment are fragile in the current technological landscape due to its disruptive nature. Without proper curation and management activities these assets will no longer be viable in the future. To address this issue, TCP provides an analytical framework for modeling the full economic costs of preservation, depicted as the "total cost of preservation". TCP can be applied to two specific cost models in order to get the total preservation cost, (1) Pay-as-you-go and (2) Paid-up.

The pay-as-you-go model is intended for organizations where there is a predictable and reliable income budget, which is available to purchase preservation services. On the other hand, the paid-up model is intended for organizations that face irregular annual budgets or which are grant-funded. TCP has a defined set of entities and interactions which are based on ISO 14721 (OAIS) [3]. However, the model and some of the terminologies have been modified in order for TCP to be applicable to more scenarios and facilitate understanding by non-specialists, as depicted in Figure 3.

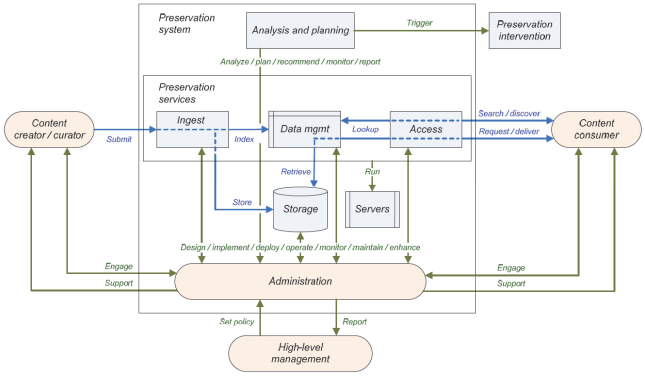

Figure 3. *TCP Entities and Interactions according to [6]*

The cost model for automated preservation archives developed by S. Strodl and A. Rauber is based on the Life model v2 [10] that will be described later on this section. The model is based in a set of assumptions that were then analyzed against the Life model to check to which extent they were applicable to small scale automated preservation systems. When the Life model is found to be lacking the support for automated preservation systems it is extended and adjusted. There are six assumptions for this model, (1) Small scale data collection, (2) Licensing & Rights of the data, (3) (Semi-)Automation preservation system, (4) Outsourcing of knowledge and expertise in digital preservation, (5) No dedicated archiving host system and (6) Internal archive.
The Life model was then applied to the automated archiving system and the model returns the value in Euros for the acquisition, Ingestion, Bit-stream preservation, Content preservation and Access.

The Life model used by [8] in their own model is part of a project which aim is to look at the life cycle of the collection and preservation of digital assets. The project is part of a collaboration between the University College London Library Services and the British Library. It consists of three phases, the first one known as LIFE[1] which was completed in 2006, the second one, LIFE[2] was completed in 2008 and the last one, LIFE[3] [11] concluded in 2010.
The first phase of the project identified the six main individual stages of the cycle, (1) Acquisition, (2) Ingest, (3) Metadata, (4) Access, (5) Storage and (6) Preservation. In order to get the complete lifecycle cost, from time 0 to time T, is accomplished by summing the costs of each of the individual stages. The main stages are then decomposed into smaller stages within the designated stage, which can be found in [9]. The model was then applied to series of case studies in order to evaluate and validate the model.

In LIFE[2] [10], the model was refined and new case studies were added. Finally, LIFE[3] extended the model in order to provide greater accuracy and assurance in the cost estimation.

There was already been work that specialized cost models for digital preservation to specific domains. For example, [13] focused that for the health sector.

The authors defined a process based on the Oltmans [12] definition of Digital Preservation, which considers that migration or emulation are the core tasks of Digital Preservation. As both tasks have strengths and weaknesses the authors chose the task of migration for the design of the process. The process consists of phases as depicted in Figure 4.

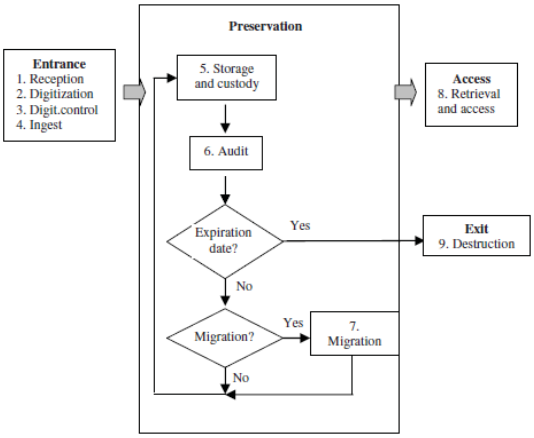

Figure 4. *Productive process for digital preservation of health information [13]*

Risk Analysis for Improving Digital Preservation Cost Models

In the models described in the section before there is no concrete evidence of the use of risk assessment methods when designing the models. This issue can raise several questions, as for example, how can the authors of such models guarantee that the cost model can be applied to any kind of organization and how can they attest that the model takes into consideration all aspects of digital preservation that put in jeopardy the digital assets of an organization.
The relation between risk and cost models might not be clear, but in this paper we aim at bridging the gap between these two aspects, as illustrated in Figure 5.

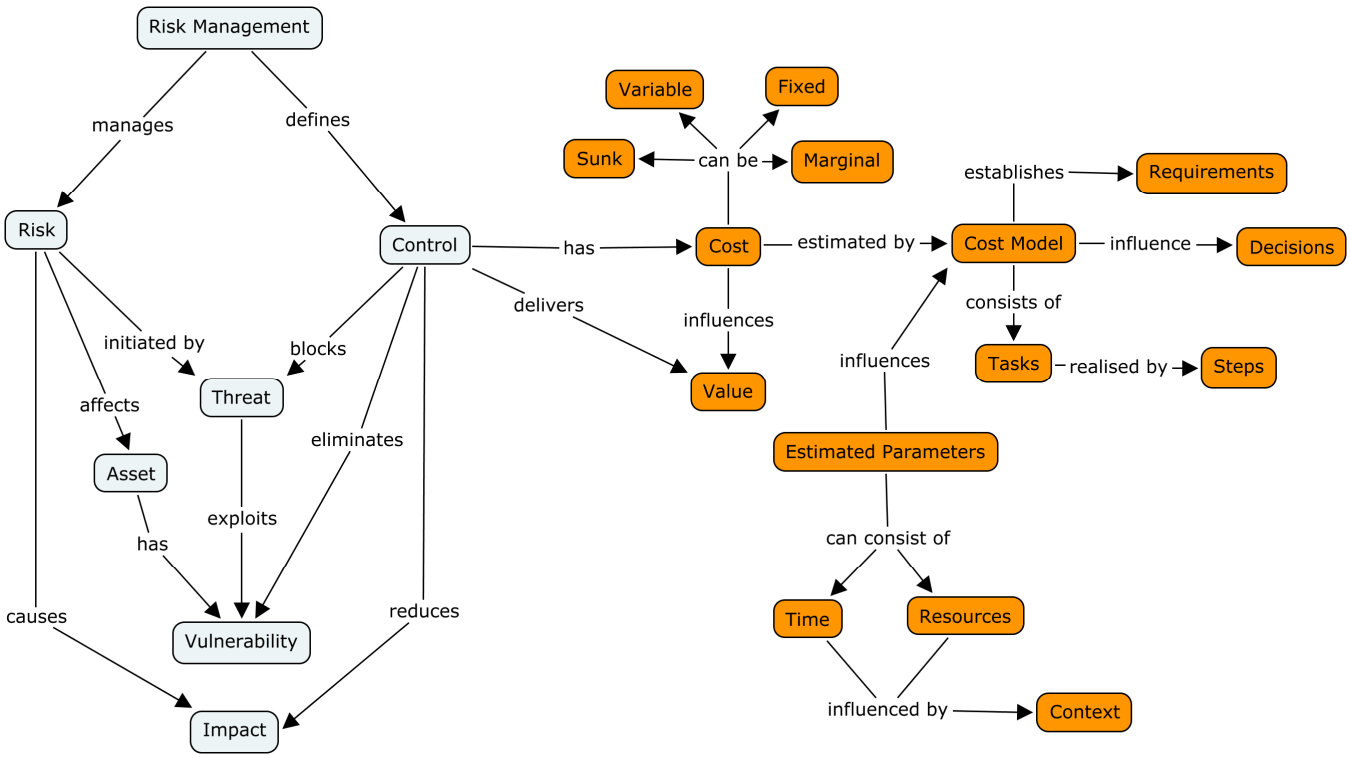

Figure 5. *Conceptual map of the risk assessment and cost modeling concepts*

Figure 5 has been adapted from [14] and depicts the risk management concepts and the connection from the risk management to the cost model viewpoints. The ultimate goal of risk management is to manage risks by defining a set of controls which will block a threat which was initiated by the risk. The risk can cause impact in an asset in which was found a vulnerability. In this way the controls blocks the eminent threat, eliminates the vulnerability and reduces the impact.

New controls bring value to an organization as it reduces the impact of risks, which is an advantage to the organization. However, establishing a new control has costs to an organization and this cost influences the value of a certain control. The costs can be decomposed into several types as, for example, sunk, variable, fixed, or marginal.

Sunk costs are the expenditure that is committed for a period of time and within this period (which can be of indeterminate duration) the organization cannot withdraw the commitment of this expenditure. This is used implicitly by some economists to define long term costs, instead of fixed time interval costs [15].

Variable costs or operating costs are the costs that are committed during the normal operation of an organization, even if the organization is not producing any product. Examples of these costs are electricity and maintenance.

Fixed costs are costs that do not change as, for example, building rent or salaries.

Marginal costs are actually an estimate on how costs would change if there was change in the output of a task [16]. For example, if you have a data center which only replicates data in one site, how would costs change if you replicate data in one site per continent?

These costs can be estimated using a cost model, which consists of tasks which are realized in a set of steps. There are certain estimated parameters that influence the cost model and which inherently differ from organization to organization. Examples are the time and resources needed to perform a certain task. These parameters are influenced by the context of the organization. For example, if an organization does not have skilled personnel to perform a task they can either take more time to perform a task, or there might be the need to allocate more people to the task or even outsource that task.

Other example is an organization in risk of losing all its data due to the lack of proper data replication mechanisms. A solution might be to buy more storage data to mitigate the risk, but if it is an organization that doesn't have any technological landscape, the required time and extra resources for that will be much higher than in the case of an organization with a strong information technology base.

Ultimately, the output of the cost models will influence decisions, as if the cost of a certain task is much higher than the perceived value it brings to the organization, some organizations might opt to not perform this task which would result in an identified risk not being treated and instead ignored.

On the other hand, if the cost is deemed acceptable to the organization, the cost model will help establishing the requirements for performing that task and treating the risk.

Seeing from another point view, risk management can also be of extreme importance to cost models, in order to prioritize and justify expenditure. Through the use of a risk matrix, we prioritize expenditure where it is most needed, in order to treat critical risks identified in the organization. A risk matrix is used during the risk assessment phase and defines the various risk levels in terms of

likelihood and consequences of a certain risk. The result is matrix, which is based in these two parameters and categorizes the risk in low, moderate, high or critical risk as depicted in Figure 6.

Likelihood						
	Almost Certain	Moderate Risk	High Risk	High Risk	Critical risk	Critical risk
	Likely	Moderate Risk	Moderate Risk	High Risk	High Risk	Critical risk
	Possible	Low Risk	Moderate Risk	High Risk	High Risk	Critical risk
	Unlikely	Low Risk	Moderate Risk	Moderate Risk	High Risk	High Risk
	Rare	Low Risk	Low Risk	Moderate Risk	Moderate Risk	High Risk
		Insignificant	Minor	Moderate	Major	Catastrophic
		Consequences				

Figure 6. *Risk Matrix Example*

Through the use of the risk matrix and if the alignment between the risk assessment and the cost model is good an organization can prioritize expenditure in critical risks rather than investing in controlling low risks.

The EC-funded 4C project, coordinated by JISC, has as one of its aims to properly address this synergy between risk assessment and cost modeling in the specific case of digital curation and preservation. The project has defined two sets of activity, (1) Ensure the awareness and uptake of existing tools and knowledge on the costs and benefits/value of digital preservation and (2) Identify the gaps and shortcomings in existing research related to costs with the objective of creating a roadmap for future research and development in the digital preservation area.

The project will have the kick-off in February 2013 and has duration of 24 months and involves thirteen partners large and small from the commercial, non-profit and public sectors. The partners involved are the Higher Education Funding Council for England (JISC) from United Kingdom, Det Kongelige Bibliotek, Nationalbibliotek og Kobenhavns Universitetbibliotek (KBDK) from Denmark, INESC ID – Instituto de Engenharia de Sistemas e Computadores, Investigação e Desenvolvimento em Lisboa (INESC-ID) from Portugal, Statens Arkiver (DNA) from Denmark, Deutsche Nationalbibliothek (DNB) from Germany, University of Glasgow (HATII-DCC) from the United Kingdom, University of Essex (UESSEX) from the United Kingdom, KEEP Solutions Lda (KEEPS) from Portugal, Digital Preservation Coalition Limited by Guarantee (DPC) from the United Kingdom, Verein Zur Forderung Der IT-Sicherheit in Osterreich (SBA) from Austria, The University of Edinburgh (UEDIN-DCC) from the United Kingdom, Koninklijke Nederlandse Akademie Van Wetenschappen (KNAW-DANS) from the Netherlands and Eesti Rahvusraamatukogu (NLE) from Estonia.

Regarding the role of risk, benefit, impact and value on the cost modeling of digital preservation, 4C will look at a range of inter-related issues from a perspective of Risk Management. The principal trade-off between cost is obviously benefit but if that can be measured with some objective relevance in some sectors of activity in the corporate world, in various contexts and for other different types of organizations, this can be a very complicated equation. Using case studies, the role of risk and risk assessment will be considered in relation to curation as one of the principal drivers for governance. In that sense, not only cost but also benefit, impact and value (and its relation to cost efficiency) will also be considered terminologically and by sector to try and characterize the influence of these factors as determinants. For example, one of the cost and risk factors that will be specifically looked into is the issue of loss and recovery from loss, as opposed to preventive curatorial action.

Conclusion

This paper presented a view on the risk analysis importance for digital preservation cost modeling. We presented the related work in digital preservation cost modeling and then formulated our view on how risk management and cost modeling activities can be interlinked.

From the work presented here we outline five major questions (Q1 –Q5) as a future outlook which will guide our future development in the cost modeling domain applied to digital preservation.

Q1: How can we effectively integrate risk management in cost modeling?

In this paper we stated that risk management can guide the cost modeling techniques. However we need to define a method and/or process to properly integrate both these aspects.

Q2: How can we use current risk treatment and assessment techniques to help improving current and future cost models?

This question (Q2) is a decomposition from Q1, the risk treatment and assessment techniques can be used to validate current and future cost models against the risks of a certain domain, such as, digital preservation. These synergies must be properly defined so that there can be a real life application.

Q3: To what extent is risk management useful to justify expenditure?

As stated before, the risk matrix can be extremely helpful to justify and prioritize expenditure when applying the cost model to a certain organization. Despite this, there are certain drawbacks as the cost model must be correctly aligned with the risk assessment of the domain, because if they aren't aligned we won't have any cost model tasks linked to the identified risks in the domain.

Q4: To what extent cost influences the value of risk controls?

As we stated before, in order to reduce the impact of a risk we define controls. However, these controls have associated costs. In this way cost models can help to define to what extent the cost of a certain control will influence the value of the control to the organization, because if the cost overcomes the value there might not be a real benefit in implementing the control.

Q5: To what extent the organizational context is a decisive parameter of cost models?

The organizational context influences the time and resources needed for performing a certain task and will inherently influence the cost of performing the task. However, we need to know how

can context influence the cost, we need to formalize and define all the aspects and relationships of the organizational context that are important to digital preservation and perhaps build a context model to assess these when applying a cost model to an organization.

The answers to the questions presented above will help to create a clear view on the synergies of risk and cost modeling, as both these aspects are important and sometimes decisive to create value in organizations. Without value many boards can't justify expenditure and can't perceive the benefit of a certain task. With this paper we want to create the awareness of this perspective and also create synergies in both domains with the main objective of raising the awareness that risk and cost are very important aspects to improve the value of organizations.

Acknowledgements

This work was supported by national funds through FCT – Fundação para a Ciência e Tecnologia in context of the pluriannual project PEst-OE/EEI/LA0021/2011 and by the projects TIMBUS and 4C, co-funded by the EU under FP7 under grant agreement no. 269940 and 600471 respectively.

References

[1] ISO 31000 – Risk Management – Principles and guidelines, International Organization for Standardization Std. (2009).

[2] G. Antunes, D. Proença, J. Barateiro, R. Vieira, J. Borbinha. Assessing Digital Preservation Capabilities using a Checklist Assessment Method. In 9th International Conference on Preservation of Digital Objects (iPRES 2012), pg. 266 - 273, Toronto. (2012).

[3] ISO 14721 – Space data and information transfer systems – open archival information system (OAIS) – Reference Model, International Organization for Standardization Std. (2012).

[4] U. Kejser, A. Nielsen and A. Thirifays. The Cost of Digital Preservation – Project Report v. 1.0. (2009).

[5] U. Kejser, A. Nielsen and A. Thirifays. The Cost of Digital Preservation – Project Report for phase 2. (2011).

[6] UC Curation Center and California Digital Library. Total Cost of Preservation (TCP) – Cost Modeling for sustainable Services. (2012).

[7] N. Beagrie, J. Chruszcz and B. Lavoie. Keeping Research Data Safe – A cost model and guidance for UK universities. Charles Beagrie Limited. (2008).

[8] S. Strodl and A. Rauber. A cost model for small scale automated digital preservation archives. *In* 8th International Conference on Preservation of Digital Objects (iPRES 2011), pg. 97 - 107, Singapore. (2011).

[9] R. McLeod, P. Wheatley and P. Ayris. Lifecycle information for e-literature: full report from the LIFE project. LIFE Project: London, UK. (2006).

[10] P. Ayris, R. Davies, R. McLeod, R. Miao, H. Shenton, P. Wheatley. The LIFE^2 final project report. The LIFE^2 Project. (2008).

[11] B. Aitken, P. Ayris, B. Hole, L. Lin, P. McCann, C. Peach and P. Wheatley. The LIFE^3 Project – Lifecycle Information for Literature. (2010).

[12] E. Oltmans. Cost models in digital archiving: An overview of life cycle management at the national library of the Netherlands. *LIBER Quarterly* 2004; 14(3-4):380—392. (2004).

[13] J. Bote, B. Fernandez-Feijoo and S. Ruiz. The Cost of Digital Preservation: A Methodological Analysis, Procedia Technology, Volume 5, pg. 103-111. (2012).

[14] José Barateiro, A Risk Management Framework Applied to Digital Prservation. PhD Thesis, Instituto Superior Técnico, Technical University of Lisbon. (2012).

[15] Incyte Consulting. Economic Aspects and Costing of NGNs. (2012).

[16] R. Turvey. What are marginal costs and how to estimate them? University of Bath – School of Management. (2000).

Author Biography

Diogo Proença received his MSc in computer science from the Polytechnic Institute of Leiria (2011) and is currently a PhD candidate at the Technical University of Lisbon. He is a researcher for INESC-ID Information Systems Group and his focus is on Systems Governance, Process Maturity and Cost Modeling. He is involved in several digital preservation projects, specifically TIMBUS, SCAPE, BenchmarkDP and 4C.

José Borbinha is an Information Systems professor at the Computer Science and Engineering Department of the Lisbon Technical University, and the actual coordinator of the Information Systems Group at INESC-ID. He has a long time interest in digital preservation, digital libraries (projects SHAMAN, TIMBUS, 4C, etc.; conferences JCDL ECDL/TPDL, ICADL; initiatives DCMI and IEEE TCDL – was a previous chair) and archives (will be the chair for iPRES2013, and represents INESC-ID at the DLM Forum).

Neil Grindley is Programme Manager for Digital Preservation and Curation at Jisc, a UK charity that funds and supports technology-related projects and services for the UK Higher and Further Education sector. He is a Board member of: the Digital Preservation Coalition (DPC); the Open Planets Foundation (OPF); and the Alliance for Permanent Access (APA). He is the co-ordinator of the EC-funded 4C Project.