Proceedings of the 11th International Conference on Manufacturing Research (ICMR2013), Cranfield University, UK, 19th – 20th September 2013, pp 151-156

A FRAMEWORK FOR IDENTIFYING UNCERTAINTIES IN LONG-TERM DIGITAL PRESERVATION

Essam Shehab
Manufacturing and Materials Department
Cranfield University
Cranfield
Bedford, MK43 0AL, UK
e.shehab@cranfield.ac.uk

Glory Chuku
Manufacturing and Materials Department
Cranfield University
Cranfield
Bedford, MK43 0AL, UK
g.chuku@cranfield.ac.uk

Mohamed Badawy
Manufacturing and Materials Department
Cranfield University
Cranfield
Bedford, MK43 0AL
m.badawy@cranfield.ac.uk

ABSTRACT

With the current expansion in digital information comes an increasing need to preserve such assets. The ENSURE (Enabling knowledge Sustainability, Usability and Recovery for Economic value) project, a research project under the European Community's Seventh Framework Programme, is the parent project to this research area and its aim is to conduct advanced research to address the challenges of Long Term Digital Preservation (LTDP) to ensure the successful preservation, availability and accessibility of preserved data in the future. Focusing on identifying uncertainties in the LTDP activities and their impact on cost and economic performance of digital preservation systems, this paper discusses a framework to identify uncertainties in LTDP for business sectors interested.

Keywords: Digital Preservation, Uncertainties, Risk, ENSURE

1 INTRODUCTION

Digital data is now dominating the information scene (Charles Beagrie Limited, 2010); most of the currently generated information and records are in digital format or end up being digitized. This is due to the existence of new technologies and the ease of handling these digital data forms by the users. In order to keep these vastly growing population of digital materials safe, useable, meaningful and accessible for a long period of time; preservation methods and techniques have been introduced and employed.

Enabling kNowledge Sustainability Usability and Recovery for Economic value (ENSURE) aims to provide a total long-term digital preservation solution for a new sector in the IT market. The businesses in healthcare, financial and clinical trials sectors are now interested to preserve their data, due to legal obligations and due to the increasing cost of data regeneration, especially in the clinical trials sector. Along with the new business sector, ENSURE is aiming to utilise cloud computing for storage and computing.

ENSURE aims to provide its customers from the three business sectors with a detailed cost and economic performance report. This will enable decision makers to select their long-term digital preservation requirements and have the highest quality preservation possible for the cheapest running cost and ensure ease of access to and security of data. ENSURE's cost model aims to tackle uncertainties and obsolescence issues that may arise. This is due to IT systems that are prone to failures and obsolescence. Failures and obsolesce issues will generate cost to mitigate, and a rigorous study is need to estimate the impact on future cost due to these mitigation strategies. To have a real cost estimate,

that reflects the effect of uncertainties, ENSURE requires that the cost modelling development should include a thorough uncertainty study. This paper outlines a framework developed for identification of uncertainties in LTDP, which will feed afterwards into the cost model reflects cost impacts of these uncertainties.

2 UNCERTAINTY AND RISK

Uncertainty can be described as a state marked by the inability to specify an entity (outcome, event, or occurrence) with precision. It is the lack of certainty, a state of having little or no knowledge about the existing state, or future outcome. Uncertainties of different types influence the successful implementation of DP. It is important to bear in mind that uncertainty does not always imply loss or damage, they sometimes create opportunity for value creation. The rapid nature of change in current times (change in technology, change in economy, change in needs, change in markets etc.) has led to the need for better perception and understanding of the resulting uncertainties created by these changes. Uncertainties lead to risk, which can be negative, in which case there is eminent damage, loss or failure; or positive, in which case they are opportunities. A negative risk would be handled by mitigations, and a positive, by exploitation.

A fair amount of effort has gone into developing strategies to manage the influence of uncertainty. These efforts have encountered challenges due to the variable nature of uncertainty (Erkoyuncu, 2011). Prior research (Erkoyuncu, 2011) reveal that the nature of uncertainty is two-fold. It can be either aleatory or epistemic in nature. Epistemic uncertainty is uncertainty that is associated with limitations of insufficient data. They describe Epistemic Uncertainty as that whose influence can be reduced through an increase in knowledge, understanding, or (relevant) data; and Aleatory Uncertainty on the other hand as that uncertainty which remains unpredictable irrespective of available data.

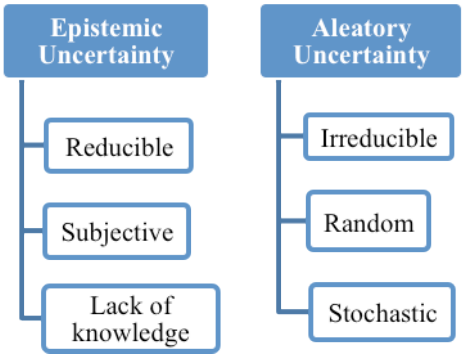

Figure 1: Natures of Uncertainty (Erkoyuncu, 2011)

The terms Uncertainty and Risk have often been used interchangeably; however, there is a blurred line of distinction. Erkoyuncu (2011) agree that the link between both concepts relates to uncertainty as the source of risk. They further point out that risk is a choice and not a fate, adding that undecided things are uncertain. Knight (2002) contributed to the study of uncertainty by describing it to be a complex from of risk. This tallies with other authors who share similar views about uncertainty which they describe as situations that are known only imprecisely or not known at all. They further state that a case of uncertainty can be better or worse than expected. Another more sophisticated definition which states that "Uncertainty is any deviation from the achievable ideal of completely deterministic knowledge of the relevant system." Erkoyuncu's extensive research in the area of uncertainty and risk gave rise to the following definitions: "Uncertainty is the stochastic behaviour of any physical phenomenon that causes the indefiniteness of outcomes meaning the expected and actual outcomes are never the same." He considers risk to be a special case of uncertainty in which the outcomes of a specific event or events have a negative effect on the overall performance of a project. (Erkoyuncu, 2011). Figure 2 illustrates the relationship between concepts. It suggests that risk is embedded in uncertainty.

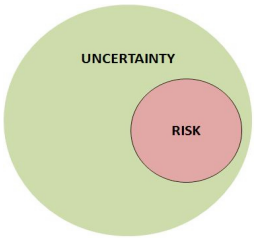

Figure 2: Uncertainty and Risk (Knight, 2002)

3 LONG-TERM DIGITAL PRESERVATION

The concept of digital preservation began as the term digital curation. The term was first used at the seminar Digital Curation: digital archives, libraries and e-science seminar. Both terminologies in terms of definition have one common theme, to make digital material available and usable in the distant future. The Digital Curation Centre (DCC) describes the term as being about "maintaining and adding value to, a trusted body of digital information for current and future use". From Lord and Macdonald's perspective, "This is a relatively new field, and terminologies are not yet stable…preservation is an aspect of archiving, and archiving is an activity needed for curation. All three are concerned with managing change over time." . For the purpose of this work, the definition by Beagrie and Jones is adopted. "Digital preservation refers to the series of managed activities necessary to ensure continued access to digital materials for as long as necessary." (Beagrie and Jones, 2001)

4 FRAMEWORK DEVELOPMENT

A framework within the context of this work can be defined as a structure providing guidance on a step-by-step approach for a goal. There is no doubt that more research has gone into risk handling as opposed to uncertainty. However, it has been successfully established in literature and this work that the handling of uncertainty is not solely about perceived threats, but also opportunity. Therefore, this framework is an effort to addressing the major uncertainties that impede successful LTDP, their corresponding risks, as well as mitigation approaches. The author anticipates that in implementing this framework, uncertainties can be addressed across sectors. It can also be used as a discussion and teaching tool. It has been developed with the aim of providing clarity and structure to the way uncertainty is discussed and handled.

4.1 Key Steps

1. The first and vital step in the development of an uncertainty framework is to actually understand what it means within your context of research. Therefore, upon review of definitions of uncertainty in general, the author has developed a working definition for uncertainty within the DP scenario.
2. The identification of uncertainties was achieved through a review of literature, company reports and interviews. Because of the stochastic and somewhat unpredictable nature of uncertainty, the author has narrowed research down to major uncertainties. Nevertheless, using this framework allows additional uncertainties to be considered. In other words, the framework is not limited to the presented uncertainties alone.
3. A step-by-step process was mapped out to guide the framework user on the identification and handling of uncertainties. Because uncertainty gives rise to risk, the uncertainty handling framework may be used as a precursor to the risk management process framework.
4. Simple tools were identified as useful for the identification of uncertain factors. Some of these tools include brainstorming, check listing etc.
5. The categorisation of major uncertainties was achieved based on descriptions in literature. An initial list was evaluated based on severity using questionnaire results and scenario analysis. This led to a final list.

Having successfully achieved all of the above, a framework was carefully put together (in a process flow format), by placing each objective in a manner that allows their combination to unfold meaningfully and make logical sense.

4.2 LTDP Uncertainties Taxonomy

Taxonomy serves as a map through the concepts of a subject. A taxonomic classification of uncertainties in LTDP has been identified as one of the objectives of this research work. In the taxonomic classification of uncertainties in LTDP, the author takes the perspective of the dual nature of uncertainty as being objective and subjective.

Objective uncertainty is associated with that uncertainty that is as a result of the stochastic characteristic of a factor. It is also known as Irreducible (or Aleatory) uncertainty, because in principle, it cannot be reduced through additional investigation. A characteristic of this uncertainty is though expert judgement may be useful in characterising it, it is not likely to be reduced by it.

Subjective uncertainty on the other hand is that which results from some form of knowledge deficiency or another. It is described in literature (Campos et al., 2007), as Reducible (or Epistemic) uncertainty, because it can be reduced by further empiric efforts. In principle, it is expected that this class of uncertainty can be reduced by sufficient study and expert judgement.

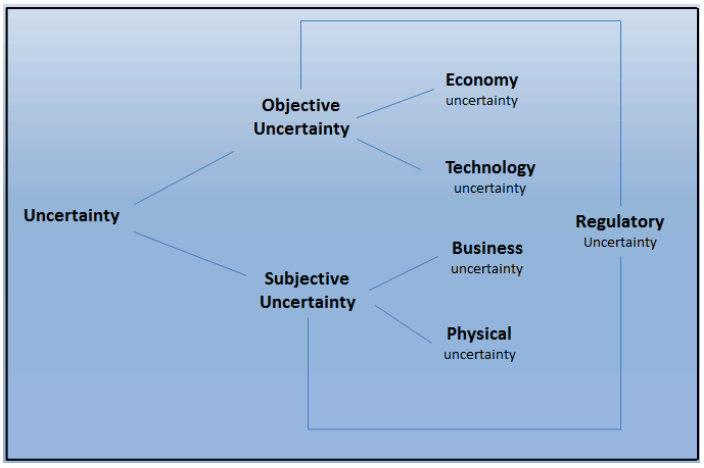

Figure 3: Taxonomy of Uncertainty in LTDP

4.3 Uncertainties Categories

Economic Uncertainty: Economic uncertainty implies the economy is unpredictable. There is a high likelihood of economic downturn, in which case there will most likely be a shortage of funds for LTDP activities; and also the possibility of an economic boom in which case; there will be available financial resources to effectively implement LTDP.

Technology Uncertainty: Technology is dynamic. Technology can be described as being in a constant state of evolution, hence the upgrades, updates, new versions, and the resulting obsolescence. Technology uncertainty is identified as a major category of because there is no guarantee that the preservation technologies used today to preserve data will still be the same ones used in few years to come.

Business Uncertainty: Every business is composed of tangible and intangible constituents. People, who could be stakeholders, employers, employees, vendors etc., are an example of tangible constituents. A businesses reputation is an example of an intangible constituent of a business. All these constituent generate a significant amount of uncertainty.

Physical Uncertainty: Physical uncertainties have to do with every element which can be seen, and whose impact is of a physical type. Physical uncertainties can be broken-down into subcategories of infrastructure, environment, storage etc.

Regulatory Uncertainty: As time changes, so does the economy, as the economy changes so, so do the regulations, laws, and policies.

4.4 Framework Implementation Tools

Figure 4 shows the full framework for identifying uncertainties in LTDP. The framework comprises the following stages:

Checklists: This is a preliminary technique used to provide a starting point to the identification of uncertainties. It guarantees that no known or common source of uncertainty to digital preservation is overlooked. It helps to ensure consistency and completeness in carrying handling uncertainties.

Brainstorming: To implement this tool, a group of knowledgeable stakeholders will come together in a formal process to list both known and new uncertain factors and risks about the prevailing context.

Interviews: This tool is introduced with the aim of identifying "concern-related" uncertainties and risks, and providing further information on risks potentially related to those identified by check-lists and brainstorming.

Structured "what-if" technique (SWIFT): This tool relies on expert knowledge for identifying and evaluating uncertainties, particularly taking into consideration the selection, preservation and dissemination stages of the preservation process, where change can be more influential, to identify potential risks arising from that change.

Failure Mode and Effect Analysis (FMEA): This tool is used to identify objective deviations and associated risks, potential causes, and consequences, regarding both the LTDP process and the digital repository itself, while making sure digital preservation's objectives, needs, and requirements have not been neglected.

Reliability Centred Maintenance (RCM): This is used in order to identify preventive measures and policies that should be put in place to protect the digital repository, especially regarding the ingestion, preservation, and dissemination phases of the DP process, which are the ones which rely on the repository.

Human Risk Assessment (HRA): This is used to assess possible human impact on every stage of the preservation process.

Scenario Planning: Scenario planning may be used to envision potential futures and provide a mechanism for testing strategic assumptions in order to determine their robustness.

5 CONCLUSIONS

Preservation of these digital assets is riddled with uncertainties born from the influence of uncertain factors. The uncertainties then present risks of two main kinds – Opportunity and threats. Effective digital preservation is achieved when risks are addressed proactively, and the first step in identifying these risks is to handle the uncertainties which give rise to them. LTDP strategies must anticipate uncertainties, which is what this framework ensures, and thereby build in resilience to respond to any future events which are at the time unpredictable. It will most importantly ensure that the data preserved remains accessible and interpretable in the future when current technologies for servers, or operating systems, and applications may not be available.

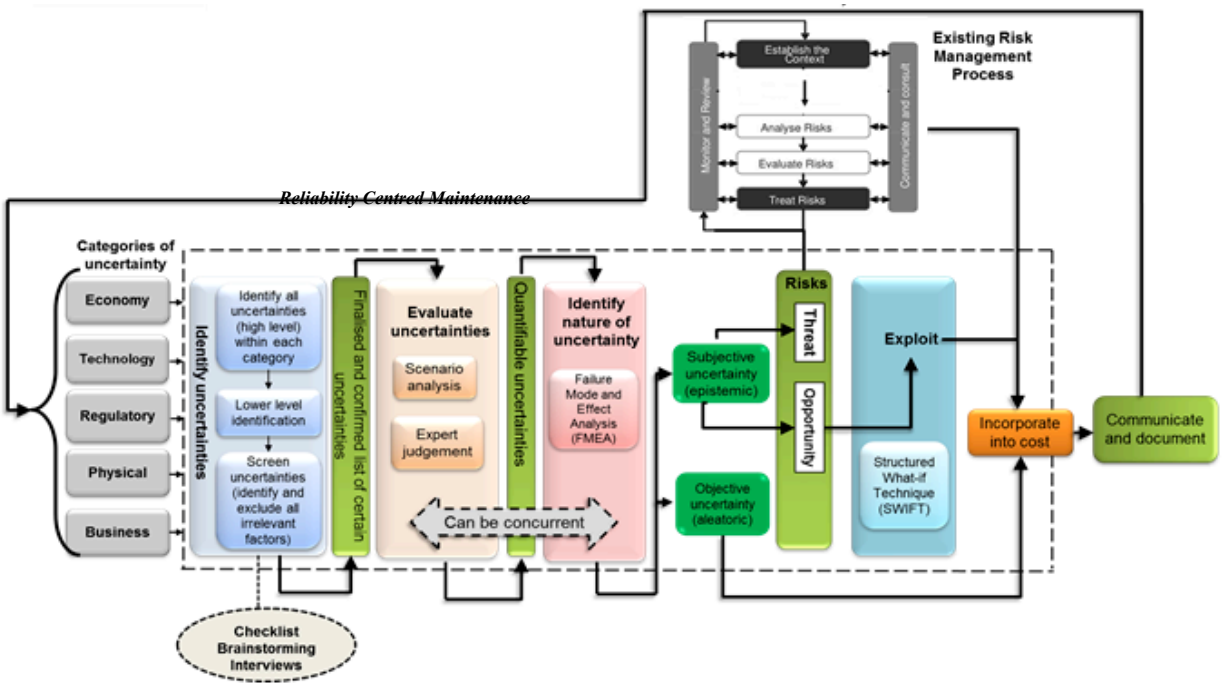

Figure 4: Uncertainties Identification Framework

ACKNOWLEDGMENTS

The research has been conducted as part of an EU-FP7 project titled 'ENSURE'. The project involves 13 research and industrial partners. The project is supported by the Commission of European Community (contract number: 270000) under the ICT Programme. The authors acknowledge the European Commission for its support as well as the other partners in the consortium (www.ensure-fp7.eu).

REFERENCES

Beagrie, N. and Jones, M. (2001), "Preservation Management of Digital Materials: The Handbook", Digital Preservation Coalition (DPC), Autralia, available at: http://www.dpconline.org/advice/preservationhandbook/ (accessed July 21, 2012)

Beagrie, N., Chruszcz, J. and Lavoie, B. (2008), Keeping Research Data Safe: A Cost Model and Guidance for UK Universities, , Charles Beagrie Limited, UK.

Campos, F., Neves, A. and Campello de Souza, F. (2007), "Decision Making Under Subjective Uncertainty", *Proceedings of the Computational Intelligence in Multicriteria Decision Making (MCDM), IEEE Symposium,* April 1-5 2007, IEEE, United Kingdom, pp. 85-90

Erkoyuncu, J. A. (2011), *Cost Uncertainty Management and Modelling for Industrial Product-Service Systems* (PhD thesis), Cranfield University, Cranfield, Bedford.

Erkoyuncu, J. A., Roy, R., Shehab, E. and Cheruvu, K. (2011), "Understanding Service Uncertainties in Industrial Product–Service System Cost Estimation", *the International Journal of Advanced Manufacturing Technology,* vol. 52, no. 9, pp. 1223-1238.

Knight, F. H. (2002), *Risk, Uncertainty and Profit,* Beard Books, Washington DC, USA.