Risk Management Based Approach For Long-Term Digital Preservation

Anand Kumar Sinha, Santosh Kumar, H. M. Singh

Abstract: Digital Technology poses new opportunities as well as new threats and problem and its functionality comes with complexity. The digital technology slowly morphed the task of record keeping of business of all kinds. And hence, the digital technology was now better named as information technology. Long term digital preservation is needed because the Technologies on page Digital information depends, change with period of time. This paper presents Limitations and Vulnerability of Digital Documents and Preservation of Documents their Business needs as well as Statutory Requirements. Potential risk of digital data and the factors causing the risk has been summarized. A Digital Preservation Management Process (DPMP) to deal with potential risk has been proposed.


1. INTRODUCTION

SINCS the inception of human civilization, the expression of thought, concern, ideas, created intellectual knowledge, policies/procedures, discourse, events etc. across social, educational, cultural, academic & Research, health, judiciary and governing institution from highest to lowest level of society at individual or group level, has been documented by the means of using (the appropriate & available then) the physical medium, may be stone, leaves, Tampatra or plain hand written or printed paper, gramophone records, chemical coated plastics films, etc. These physical mediums, of course have been changing /being replaced by development /invention of newer or better medium. Societies around the globe from respective countries have shown a vital interest in preserving intellectual materials that document issues, concern, ideas, knowledge, cultural memory etc. Accordingly the custodians of well being of society have shown tremendous amount of concern & effort in creating adequate means and where withal of preserving such physical material carrying created knowledge & cultural memories for the benefit of their future generation in distant future. The available information & artifacts as documentary materials since generations on all varieties of physical medium in India or in other countries, is the live witness of the great concern of our forefathers in preserving the created knowledge over generations. The use of acidic wood pulp paper for documenting works has been a major share among various varieties of physical medium since year 1850.

1.1 DOCUMENT

- Mr. Anand Kumar Sinha is research scholar in the Department of Computer Science & Information Technology, SHUATS, Allahabad, anandk719@gmail.com
- Santosh Kumar is presently working as Professor in department of Computer Science and Engineering at ABES Engineering College, Ghaziabad. After completing his bachelor and master degree, He has completed PhD degree from Jawaharlal Nehru University, New Delhi, India. His research interests are Database, Materialized Views, Web Mining, Distributed Systems etc.
- Dr. Hari Mohan Singh is working as Assistant Professor in the Department of Computer Science & Information Technology, SHUATS, Allahabad. His Area of Specialization are Object Oriented Technologies, Network Survivability; Mobile Computing, Parallel Algorithms.

1.1.1 The Paper Document

In last 150 years or so, the acidic wood pulp paper became the universal standard physical medium /means for recording (write/type/print) any kind of knowledge, ideas, information either for personal, institutional or public purpose. Obviously a document always meant to be a paper document only Consequently, all statutory requirements of every national government as well as International organizations presumed and accepted that all documents would be in paper form itself. In Indian context all the documentary records were maintained by all the organizations in early British period through the use of Hand written/typed/printed paper records/files in the records offices of concerned organizations. Maintenance and preservation of old records/transfer have been an important responsibility for each organization for smooth administrative function as well as dissemination of knowledge in posterity to future generation users. And the system of information recording & sharing has been successfully functioning with reasonably full proof preservation measures while using ‘Paper’ as physical medium However, these had been a traditional apprehension/conflicts in the minds of community of custodian of cultural memory & other official records in term of allowing people to access the books/manuscripts/ recordings as well as ensuring such documents in paper form are safely preserved for future. Generally it is estimated that the shelf life of acidic wood pulp paper to be in decades not centuries. None the less in spite of traditional conflict between ‘Access’ and ‘Preservation’ of the paper document, the system of knowledge sharing has been going on smoothly by and large to the satisfaction of both users as well as custodian community. The migration is set of tasks used to achieve the periodic transfer of paper documents to software files in digital form then digital materials from one hardware/software configuration to another or from one generation of computer technology to a subsequent generation [3].

1.1.2 The Digital Document

As per United States Census Bureau 1993, about 45% of US workers were using a computer and the numbers must be now around 100% Similarity the percentage of computer usage has gone up drastically high even in developing country like India or even in underdeveloped countries. Now in most of the organizations, the information is created and stored using digital technology. An increasing amount of writing /typing/ printing work is being performed with computers. Now a day professional sound and video recording takes place virtually through digital technology. As a means of recording, storage
and providing access to stored information using digital technology has enormous advantage over any traditional physical medium based information. Here there is no conflict or fear of deterioration/loss of original document due to user access vis-à-vis the responsibility of preserving them safely [4]. Unlike in the case of traditional form of document, the concept of original document is nullified for the user as he/she can access a digital stored information in its exact image of original work on his/her computer, that too it can be accessed by multiple users simultaneously. Additionally the user can read/retrieve information with flexibility in the display format, manipulate digital information at his desktop in terms of viewing, storming or printing them. Briefly the means of record keeping and access using digital technology over physical medium based document is in comparable. Actually since 1940’s when the first digital computer made its appearance, the digital technology has made a tremendous progress by leaps and bound from mid-1990s onwards. Added with the proliferation of cheap desktop computers along with epidemic kind of spread of Internet, the digital technology has virtually transformed almost all aspects of our life at all kind of offices and institutions. Almost all organizations across the world, large or small, govt. as well as non-Govt., embraced digital technology as integral part of their business processes in a stage manner.

- Computation and compilation tool
- Text / information processing tool
- Information storage device.

The digital technology slowly morphed the task of record keeping of business of all kinds. And hence, the digital technology was now better named as information technology. All business information and records that were usually kept exclusively in paper document form, are now slowly giving way to digital document form. It is mainly due to later’s extremely attractive characteristics like its high storage density, ease of access and convenience to operate on and manage etc. Now almost all organizations have their business records maintained partly in paper document form and partly in digital document form or completely in digital form with ethos of paperless office.

1.1.3 Limitation and Vulnerability of Digital Documents

No doubt, Digital Technology poses new opportunities as well as new threats and problem and its functionality comes with complexity. Reading, writing, and understanding of information on a paper based document is comparatively easier than accessing information in digital form as it requires hardware equipment like computer and associated software. Usage based on sole dependence on digital technology based equipment and software is a key limitation to note i.e. Paper document is physical medium/material dependent however digital documents are device dependent, That too, a high tech equipment which demand combined high end hardware & software related technical skill to operate/use and repair & maintain it, unlike conventional mechanized equipment of day to day use by human society. Over and above, the digital technology related to hardware & software is highly progressive and rate of obsolescence is pretty high, leading to shelf life for such devices may be for couple of years, 3-5 years or utmost a decade. Under such circumstances, the accessibility / usages / understanding of digital documents in long term i.e. even after a decade is questionable. There are numerous stories from our day to day work life in the field of academics, research, defence, government and corporate offices where enormous amount of data stored in 5/4” & 3½” floppies, magnetic tape and spools, hard disks etc. are gathering dust in various organizations and data retrieval or present configuration of PC is tedious task if not impossible since such stored data have been worked upon hardware & software technology of 80’s and 90’s. Also who today has punched card readers, Desk tape drive or a working copy of COBOL or FORTRAN 77? Even newer technology is becoming obsolete in 3-5 years. We cannot save the hardware machine if there is no spare parts available in the trade and we cannot save the software it there is no one who knows how to use it. Rapid changes in the means of reading/writing/typing/printing/recording of information, data storage format and technologies for use threats the life of digital document/information too short and vulgar in the overall interest of any individual or an organization or society at large. To cut short the whole argument for digital technology based digital documents, it is said that while the eligibility of paper documents can be guaranteed for few decades & maybe centuries, the same cannot be assured for digital document even for a decade and some time even shorter than that. Digital documents are vulnerable to loss due to technologies obsolescence and physical damages. The most probable cause of loss of digital information over time, can be foreseen in two ways i.e. due to obsolescence in technologies as well as due to physical damage caused by human negligence, natural calamities or wear and tear due to aging of devices. Some of the real life scenario of technology obsolescence (hardware/software) may be envisaged as below:-

- The storage media is replaced by its upgraded and dense version which is smaller, faster, easier to read compared to existing one and whose tech support is being withdrawn by the vendor.
- The storage medium reading device is no longer manufactured.
- A file format is superseded by newer version and the previous one is no longer supported by the original vendor or concerned standard body.
- For creation, management or access of digital content, the software used is being superseded by an upgraded version with higher capabilities and added features while using the state of art latest technology.
- Faster, more powerful machines with higher storage and processing speed is being introduced by vendors for computer hardware of all types of size and capacity.
- Maintenance of digital content over time is always faced with more difficulties due to inherent technology product market dynamics of emergence, merger and fading away.
- Last, but not the least, the present generation user of the preserved data, lacks the operational knowledge about the age old used software details e.g. the lack of operating know how of application software or high level language of 70’s – 80’s among current younger generation.

While most of the organizations, finding so many attractive characteristics of digital technology, started to keep their records in digital form that too perhaps without realizing foreseeing the obvious inherent constraints and complexity to their long term access and eligibility. Consequently many organizations without being aware lost their digital data completely due to
technological obsolescence and once they realized, it was too late to recover causing them huge amount of credibility as well as financial loss.

1.1.4 Preservation of Documents (Records): Business needs as well as Statutory Requirements

An organization Govt./non Govt. generally keeps the records of all its business operations as well as knowledge asset created in process, in document form (may be either in paper document or in digital document form or both) [13]. Additionally organizations & institutions involved in scientific, Research, social, cultural and other professional activities do generate huge amount of knowledge asset that are needed to be preserved for their long term utilization and usage. eg. academics, educational institution, Research & Development laboratories, temples of memory institutions (like library, archives, museum), publishers of academic books & journals etc. Further all Govt./Non-Govt., Profit/Non Profit organizations have to mandatorily comply to all statutory regulations of their respective national/state governments and for this purpose, they have to keep their required business information and records for many years as specified by the statutory requirements. As these records need to be periodically audited for their compliance by designated auditor, the legibility of records needs to be guaranteed for statutorily required period.

In India, there are the set of statutory acts in which the term ‘Document’ play a significant role may be: Official gazette of GOI, Indian evidence Act 1897, Destruction of records Act 1917, Negotiable Instruments Act, Delivery of Books & newspapers (public library) Act 1954, Public Records Act 1933, Public Records Rules 1997, Comptroller & Audit General’s Act 1971, Companies (Preservation & Disposal of Records) rule 1966, Right of information Act 2005 (Public Records Act and Public Records Rules were amended in 1993 and 1997 respectively to accept document in digital form [13]) Further as all statutory requirements presumed all documents in paper form, from mid 1980’s a strong need was felt among most of the governments of developed countries (where digital technology had penetrated in their business processes) either to adopt new set of Acts and/or modify existing set of Acts so that document in digital form might be legally accepted as well. Subsequently, with the wider usage of legally recognized digital documents and frequent occurrence of loss of archived digital documents due to technological obsolescence. By mid of 1990s, many of the governments and scientific communities in advanced countries started feeling the essential requirement for all organizations to preserve their digital documents for long term access to meet their statutory or legal or business needs. That is, they are needed to ensure maintenance of Bitle stream as well as continued accessibility of content of their digital assets for long term (called Long term Digital Preservation). Similarly while lagging behind by 10-15 years or so, the IT Act 2000 and IT Amendment Act 2008/2009 has been adopted by Indian Govt. for legal recognition of electronic records (ie records in digital form) across various other existing Acts [13]. As far as growth and usage of digital technology in business operation goes, in India also by mid 1990s, both in Govt. as well as in private sector organizational documents started being created and kept in digital form (Partly in paper or in digital form or only in digital form). Since such records in digital form do also face the threat and risk of being lost irrecoverably due to technological obsolescence; thus, by around year 2008/09 or so it was getting universally recognized by higher echelon of Govt. that it is imperative for all organizations to preserve their digital documents for long term eligibility (called Digital Preservation), both for business needs as well as statutory requirements and necessary effort may be initiated for preservation of their digital documents produced by its offices along with substance of cultural heritage in digital form for their long term eligibility [13]. On this front efforts made by private sector organizations, is ‘Business in Confidence’ and is not open to public.

2. RELATED WORK

In [1], author proposed a Framework for digital preservation of three-dimensional data. The framework is based on emulation and metadata which guarantees The authenticity and usability of data. In [9], author categorized five distinct methods to recognize information quality: transcendent (timelessness); manufacturing-based (consumer preferences); value-based (cost and price); product-based; and user-based. In [2], author has suggested a service oriented architecture using Web Services for migrating the the cultural heritage institutions to preserve the data in digital formats. An approach for Migration of complex formats has been discussed. The proposed system also preserves metadata information which could be used by the clients institutions to preserve the document it and maintains object authenticity. In [5], author has given theoretical foundation for establishing the association between archival quality E and information quality research. Some test measures have also been given to test the quality of digital data preserved in large repository HathiTrust. HathiTrust is a technological environment to address the common challenges of of digital data collection, preservation and accessing [6]. Archival quality is very important to an emergent theory of archival science [7]. An association between archival quality and the trustworthiness that scientists put down to digital data repositories has been documented [8]. In [10], author stated the most noteworthy requirement for long term exercise to digital surrogates which tells that treat digital circuits and archival records that carry traces of their fluid lifecycles and therefore are worthy of management and preservation as archives.

3. AREAS OF DIGITAL PRESERVATION

- Digital Preservation in Grids and Clouds
- Digital Preservation Lifecycle Management for Multimedia Collections
- Digital Preservation of Scientific Data
- Digital Preservation Needs of Scientific Communities
- The Long-Term Preservation of Web Content
- Digital repository: preservation environment and policy
- Research Issues in Migration and Long-Term Preservation

4. ISSUES AND CHALLENGES IN DIGITAL PRESERVATION

Before going into deeper technical depth about ‘Long Term Digital Preservation’ and associated processes, let us refer to the literal meaning of the word ‘Preserved’ as mentioned in Oxford Dictionary i.e. “To keep or maintain things in an unchanged or perfect condition”; and the word ‘Preservation means “Action of preserving’; for example, the preservation of food, one’s health, work of art and culture etc. or degree to which something has been unaffected by age, weather,
environmental condition etc. Digital preservation is the set of activities which ensure that digital objects are accessible for long duration whenever required. In [11], definition of digital preservation is given as “the planning, resource allocation, and application of preservation methods and technologies to ensure that digital information of continuing value remains accessible and usable.” Long term digital preservation is needed because the Technologies on page Digital information depends, change with period of time as outlined in the influential report of a task force set up in 1994 by the Commission on Preservation and Access and the Research Libraries Group [12]; frequent changes in the resources of storing information in their formats and technologies. [11] suggests, the challenges of digital preservation are multifaceted, involving a combination of technical and organisation issues. In [13], author discusses about challenges in digital repositories for long term preservation of their digital assets against various kinds of threats, and their vulnerabilities against perpetual technological obsolescence.

5. GOALS AND OBJECTIVES OF LONG TERM DIGITAL PRESERVATION

As mentioned earlier over the last 20 yrs of so, the preservation communities in developed countries have been grappling with the challenges pertaining to the task of long term digital preservation and evolving wedge and means of managing them through experiments and research on different approaches and methods. However on the analysis of knowledge acquired so far, it is understood that the task of long term digital preservation has three objectives. They are:

a) To safeguard various kinds of digital assets for preservation over long term against various kinds of endangered threats and risks i.e. physical damage and obsolescence of storage media over period of time.

b) To manage digital infrastructure assets of the designated establishment. may be a Data Center/Archive/Repository, e.g. the system hardware, operating system, storage media, database, application software, standards etc.; against their technological obsolesce i.e managing technological obsolesce of digital resources deployed for the preservation practices.

c) To ensure the integrity, authenticity, confidentiality and reliability of reproduced preserved digital data from archival storage media whenever accessed over long term.

On bird’s eye view of above three objectives, the probable solution for the task of long term digital preservation is taking shape of an inter-disciplinary domain of Information & Archival Science and Computer Science where the first objective may be handled by conventional Information & Archival science professionals but next two objectives needs to be managed by Computer Science scientists. In other words, it can also be said that ‘Bit preservation Tasks’ and ‘Logical preservation Tasks’ are primarily two functions of long term digital preservation. The ability to preserve bits of digital record against storage media degradation is ‘Bit preservation’ where as preservation of access, readability and understandability of the content over long term against technological obsolesce of deployed digital resources, is ‘Logical preservation’.

6. THE PROPOSED FRAMEWORK

DIGITAL PRESERVATION POTENTIAL RISK AND MANAGEMENT PROCESS

We have proposed the revision of threats to digital preservation [14], and the causes of the potential risks in Table 1. Process to deal with the potential risks in digital preservation, i.e. Digital Preservation Management Process (DPMP) is given in Fig.1. Potential causes for digital preservation are obsolescence of resources which occurs hardware and software up gradation, faults in software, old file format not supported in new software. Other causes are natural disasters, economical, and political.

**Table 1**

<table>
<thead>
<tr>
<th>Risk</th>
<th>Factors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Obsolescence</td>
<td>Hardware, software and storage media</td>
</tr>
<tr>
<td>Process</td>
<td>Software faults Software obsolescence</td>
</tr>
<tr>
<td>Transfer failure &amp; corruption</td>
<td>Copying of digital materials to backup tapes, System failure</td>
</tr>
<tr>
<td>File formats &amp; complex digital objects</td>
<td>Some file formats might not be supported accurately by most available software</td>
</tr>
<tr>
<td>Storage media failure</td>
<td>Due to a lack of refreshment policies</td>
</tr>
<tr>
<td>Human error &amp; inappropriate access</td>
<td>During alteration or deletion, Control access to content internally not just externally</td>
</tr>
<tr>
<td>Non compliance</td>
<td>Not meet our legal obligations (Laws, policies (internal and external), and standards), Non compliance may result in financial loss, loss of reputation and trust</td>
</tr>
<tr>
<td>Institutional risk</td>
<td>Lack of funding and political climate can be a risk</td>
</tr>
<tr>
<td>Infrastructure</td>
<td>Hardware faults, Hardware obsolescence, Communication faults Network service failures</td>
</tr>
<tr>
<td>Disasters</td>
<td>Natural disasters</td>
</tr>
<tr>
<td>Management</td>
<td>Economic failures, Organizational failures</td>
</tr>
</tbody>
</table>

The proposed model for Risk Management in long term digital preservation, first identify the digital preservation requirement and its context then the possible potential risks As given in table 1 are identified and level of risk is calculated. Risk treatment process, selects and applies the appropriate measures to modify the risk in long term digital preservation. This should ensure the efficient excess of digital resources for long duration. Risk treatment plan is is created keeping in mind the security, performance, integrity, authenticity, proposed actions, and priorities of the new system. Further plan is implemented and evaluated for any cause and effect.
Digital documents are vulnerable to loss mainly due to technologies obsolescence and physical damages hence there is a need for long term digital data preservation. Evolving technologies and approaches for digital data preservation have been discussed. Further Potential risks of digital data preservation and the process to deal with these potential risks has been proposed.

REFERENCES
