

# Long Term Preservation - Research Study

## 1.1 Introduction

This document describes how the study into the field of long term digital preservation will be undertaken, principally in terms of scope and approach. This research study is part of the DNEP project. The purpose of the study is to investigate in more detail the issues surrounding long term digital preservation and its impact on current efforts within the DNEP project to establish a working first release of an electronic deposit. The research study initially will run for one year.

## 1.2 Objectives and Scope

The objective of this study is the identification/development of methods and techniques, which make digital objects more persistent over time and thereby raising the knowledge in the field of long term digital object preservation. The National Library of the Netherlands (KB, Koninklijke Bibliotheek) is faced with the problem of preserving large amounts of digital documents for the long term. These documents come from two sources: from digitisation of paper documents and from new media types submitted directly in digital form. The digital form offers many advantages; it also comes with a challenge of its own: how can we insure that such documents (now, digital files) may be preserved for a long time, surviving changes in storage, computer hardware, software, formats, etc.

We will use the term digital object to identify all digital information stored in the broader multimedia meaning, including images, sound, video, interactive scenarios, and even programs. Preservation of digital objects needs to be looked at from at least three perspectives: intellectual preservation, media preservation and technology preservation.

### Intellectual Preservation

Intellectual preservation addresses the integrity and authenticity of the information as originally recorded. The need for intellectual preservation arises because the ease with which an identical copy can be made, quickly and flawlessly. Digital objects are rendered by computer programs and peripheral hardware and therefore have no unique physical characteristics. It is, therefore, more problematic to identify what constitutes authenticity of content and lay-out in a digital object context.

### Media Preservation

Media preservation addressed the preservation of the medium on which information is stored, such as tapes, disks, optical disks, CD-ROMs and the like. Often this is simply a matter of rewriting the data on the same medium, i.e. media refreshment. In addition, media migration also allows to make full use of new storage performance improvements.

### Technology Preservation

Technology obsolescence is even more of a problem than medium decay and has to be addressed by technology preservation. Rather than simply refreshing, we also need to make sure that digital objects will be accessible on emerging new technology platforms. Rather than simply refreshing, we also need to speak of migration and emulation:

- Migrating information forward through technology / format stages as they become available and as the old technologies / formats cease being supported by vendors and the user community.
- Emulating old and obsolete technologies / formats on current technology platforms

An effective preservation strategy must take all three aspects into account. The scope of this research study is aiming to cover all aspects. Our initial goal:

*“The identification/development of methods and techniques in all three aspects which make digital objects more persistent over time and thereby raising the knowledge in the field of long term digital object preservation”.*

It has been recognised that the issues surrounding the archiving of Web pages is a research field in its own right. In addition to the above mentioned aspects we also have to deal with issues characteristic for the electronic medium called Web pages:

- What is the original version of a web page in light of dynamic personalised pages
- How to deal with versions of pages
- Selection criteria for web page preservation
- Tooling to extract web pages from the Internet and checking of new versions
- How to deal with references to other web pages

These are only a few of the web page specific issues that potentially could be addressed. Within this study it has been decided to restrict our focus on a tool selection process to harvest static Web pages for preservation purposes.

### 1.3 The Project Plan, Phases and Deliverables

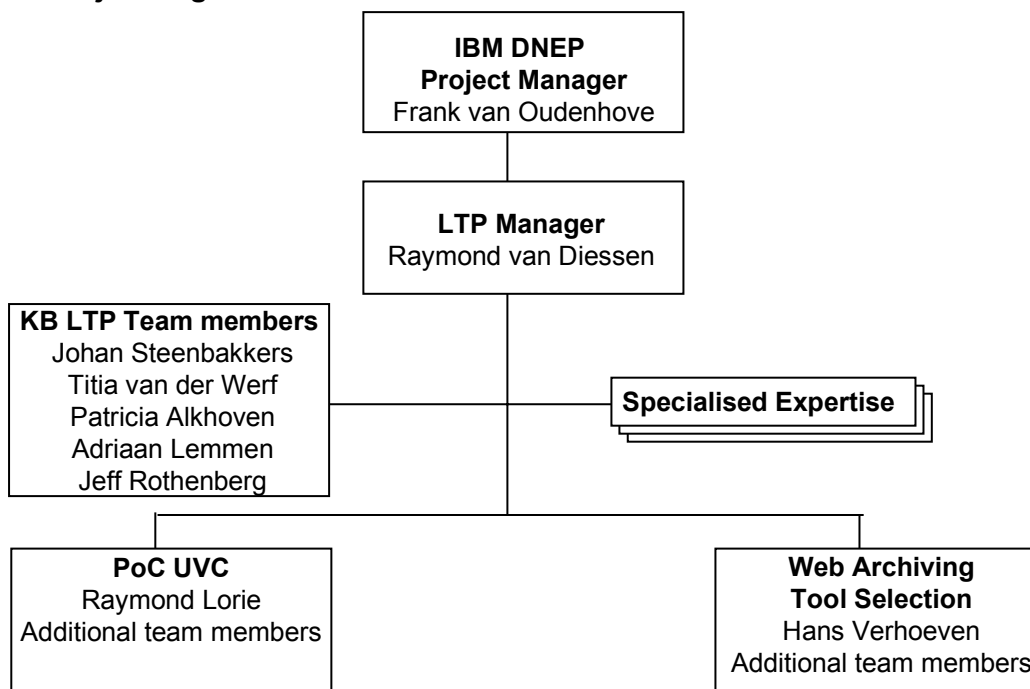
The elapsed time for the project will be some twelve months starting from the start of planning work in early November 2000 and ending in December 2001.

The main project phases are as follows:

- Phase 0: Planning.
- Phase 1: Impact and alignment of findings with DNEP project.
- Phase 2: Identification of core concepts.
- Phase 3: Media preservation – assessment of best of breed processes and techniques.
- Phase 4: Technical preservation - Proof of Concept UVC emulation approach.
- Phase 5: Intellectual preservation - authenticity in a digital environment.
- Phase 6: Web archiving tool selection
- Phase 7: Production final report.

A number of deliverables will be produced during the work of each phase. IBM will drive all activities and the production of the deliverables but will rely heavily on the active involvement of the participating staff of the KB. The final deliverable will be a final report summarising all findings.

### 1.4 Project Organisation



## 1.5 The Main Phases of the Project

The following comprises the activities that will be undertaken in each of the main phases of the project. The activities in all of these phases will be driven by IBM, working in conjunction with staff from KB.

### 1.5.1 Phase 0 – Planning

Planning is concerned with:

- An understanding of all activities, deliverables, resources.
- Confirmation of the objectives and scope of the project.
- Describing project governance and risks and containment actions.
- Preparation of a Project Workplan document.
- Agreement on scope of RSPW document including sign-off.

### 1.5.2 Phase 1 – Impact and alignment of findings with DNEP

Impact and alignment of findings with DNEP is concerned with:

- Agenda and meeting preparation of the monthly LTP team meeting.
- Preparation of weekly and monthly reporting.
- Synergy with main DNEP project.

### 1.5.3 Phase 2 - Identification of core concepts

Identification of core concepts is concerned with extending the knowledge of the project, and creating a shared mind set. This is essential to gain the support and commitment of all the staff to the project. Specifically this phase includes:

- Developing common terminology and definitions.
- Formulation of the initial research question and subquestions that will drive the research activities.

### 1.5.4 Phase 3 – Media preservation: assessment of best of breed processes and techniques

Assessment of best of breed processes and techniques in the area of media preservation focuses on the experience gained in the area of large scale media migration projects and the lessons they have learned:

- Investigating the technological trends over the coming years
- Identification of techniques to estimate media quality as a precaution
- Management strategies for media migration in large volume operational environments
- Impact on the to be defined preservation module

### 1.5.5 Phase 4 – Technical preservation: Proof of Concept UVC emulation approach

Within the field of technology preservation we already identified two major streams:

- migrating information forward through technology / format stages as they become available and as the old technologies / formats cease being supported by vendors and the user community.
- emulating old and obsolete technologies / formats on current technology platforms

From a practical perspective probably both approaches will be used in an electronic depot. The first task is to identify current initiatives in this area and to formulate the characteristics of each migration and simulation.

The emphasis in technology preservation will be placed on a PoC study conducted jointly by IBM Research (Raymond Lorie) and KB to validate the concepts behind Raymond Lorie's emulation approach based on a Universal Virtual Computer (UVC).

IBM Research is already working on the prototype implementation of the UVC emulation to be combined with the PoC activities:

- Selection of the simple document types to be included in the UVC proof of concept
- Identifying POC evaluation criteria

- Design schema and language definitions for meta data
- Design UVC
  - basic design criteria to incorporate a instruction
  - initial basic instruction set
- Assembler to parse UVC programs
- UVC runtime module
- POC testing of UVC against identified document types

IBM Research will commit itself to take the selected document type, jointly selected in co-operation with KB, as one of the first test cases to validate the UVC approach.

During initial meetings the exact definition of the data structures needed to manage technical preservation, called Preservation Layer Model (PLM), has been identified as a second focus area in this phase. The definition of the PLM is concerned with the following:

- Evaluation of existing PLMs
- Definition of a flexible and complete PLM model to be used by the preservation
- Impact of proposed PLM on DNEP design

#### 1.5.6 Phase 5 - Intellectual preservation

The focus in the area of intellectual preservation will be placed on workable definitions of authenticity in a digital environment. The main issue to be addressed:

- Definition of authenticity in a digital context
- Characteristics / attributes of authenticity
- Specific authenticity attributes values of KB use digital object types
- Implications on preservation module and processes

#### 1.5.7 Phase 6 – Web archiving tool selection:

The focus of this phase is the selection of one or a set of tools to archive static Web pages from the Internet. This selection process has to address the following aspects:

- Selection criteria to guide selection process.
- Identification of initial tools set to be evaluated
- Impact Web archiving tools on DNEP.

#### 1.5.8 Phase 7 - Production final report:

In the final report all the findings in all three aspect areas (media-, technical- and intellectual preservation) are integrated and used to define the requirements specifications of a DNEP preservation module.

## 1.6 Project Deliverables

The exact content and format of all deliverables will be agreed KB prior to preparation and issue. The deliverables, which this Project Workplan anticipates, are as follows:

From Phase 0: Project Workplan, i.e., this document.

From Phase 1: Monthly progress reports and minutes of all LTP team meetings

From Phase 2: Awareness workshop programme held jointly with IBM and KB research study staff members to identify core concepts and creating a shared mind set resulting in a glossary of terms.

From Phase 3: Report containing a overview of best of breed specific media migration processes and procedures.

From Phase 4: Working demo of UVC based on the selected document type and accompanying PoC evaluation report and separately a report on the PLM to be used by the preservation module and associated

impact analyses on current DNEP activities.

- From Phase 5: Report on the authenticity of digital objects with the emphases on building a general framework in which to address authenticity for different digital object types and the impact of such a framework on preservation processes and procedures as to be defined in the preservation module .
- From Phase 6: A report on the results of the tool selection and the potential impact on DNEP.
- From Phase 7: Final report on long term digital preservation combining the research results of all the different phases and a requirements specification of the preservation module to be integrated into DNEP