Risk Analysis of Digital Library Material

Deborah Woodyard
Woodyard-Robinson Holdings Ltd.
Christchurch, New Zealand

Abstract

There are multiple risks posed to the longevity of digital materials and the risk analysis method presented here provides a simple but effective way to assess the current state of your digital collections.

This semi-quantitative risk analysis was developed at the British Library (BL) and its application to the BL digital collections is used as a case study.

Introduction

Managing digital library collections has become increasingly challenging as digital materials have grown in size, number and complexity. They exist in a variety of environments and different formats and each face a range of threats to their longevity. Optimum care of collections will be achieved when all the risks that occur throughout the life-cycle of digital collections are minimised. Consideration needs to be given to managing these objects from selection and acquisition, through description, discovery, and delivery, to long-term storage and preservation.

An urgency to engage solutions exists because providing long-term access to digital materials requires earlier intervention in their life-cycle than for their paper counterparts. For example, the effect of media degradation on digital media has a more severe impact in a shorter time frame. More importantly the essential nature of digital material is a terminal reliance on intermediary devices to be accessed. This translates to a vital chain of computer storage technology, software and hardware requirements. Regular changes in these technologies threaten to break the links in the chain, rendering digital materials useless.

The British Library (BL) acknowledged these issues during development of its Digital Object Management Programme and called for a risk analysis of digital materials to enhance understanding of this new breed of library objects. The aim was to examine the risk of loss of access to digital materials to produce a ranked order of predicted life expectancy and a hierarchy of vulnerability and urgency for attention.

This paper describes the development of this risk analysis. It explores the risks posed to digital material and discusses a method to assess and quantify each risk factor. Interdependencies between risk factors are evaluated to provide suitable equations for meaningful comparison of results between digital material in each category of risk.

As this analysis preceded the implementation of consistent digital management procedures across all types of digital materials in the BL, it was designed to work with high level information about collections. However it could work equally well with detailed data at the level of individual items.

The resultant values for each risk can be compared across digital collections to demonstrate which materials and what risks require the most urgent attention.

Complementary Risk Analyses

Risk analysis is a common management tool, however background research showed that the focus of a risk analysis may vary widely and that it was still a new area of exploration for digital materials.

Risk analysis for collection materials is being used or developed by other organizations such as the Canadian Museum of Nature, Cornell University Library, OCLC, Stanford University Library and the National Library of Australia. However, each of these institutions adopted a different focus in their risk analysis to the BL method described in this paper.

Traditional risk analyses evaluate the probability a risk will occur and the impact that risk would have if it does occur. For example, The Canadian Museum of Nature has been perfecting and teaching their methods of risk analysis for conventional museum collections for many years. They have developed a useful model for assessing Managed Risk[1] that evaluates the fraction of the collection susceptible to the risk, the loss in value from the result of the risk, the probability of the risk, and the extent of the result of the risk.

Several risk analyses being designed for digital materials still maintain a similar traditional risk approach to the analysis.

However, when these risks are assessed for digital materials it emerges that comparison is meaningless because the most significant risks are all inevitable, completely destructive and destined within a short time frame. A new risk analysis model is required.

OCLC have been developing INFORM,[2] a detailed risk assessment specifically tailored for digital materials that carefully analyses more than sixty different risk factors. This method combines traditional assessment with an approach to address digital issues. However, it assumes that the digital materials are already stored in an archiving system and therefore would apply later in the life cycle of digital materials than the method described in this paper.

Work at Stanford University Library has been developing an approach to assessing the risks to specific file formats considered to be their standard supported file types in their collections. Prioritisation based on other issues such as value or technology change have not yet been formalised.

The National Library of Australia is discussing the specific aspects of risks as they are manifested in their work on digital collection material. The emphasis so far has been more qualitatively focused and providing a traditional risk management assessment of the severity versus the probability.

The Risk Management of Digital Information: A File Format Investigation[3] report is based on an investigation conducted by Cornell University Library to assess the risks to digital file formats during migration. While this risk assessment is specifically aimed at digital materials its focus is on one specific aspect later in the life-cycle of digital materials.

This method developed at the British Library differs from the other risk analyses in that it can be applied to digital materials early in the life-cycle. It may also be adopted for continued use throughout the life-cycle of digital materials. It provides very broad and high level information for general management. A key factor is that it assesses when the inevitable risks may occur unless they are mitigated.

Risk Assessment

Many factors threaten our ability to continue to access digital material in the long term. Each of the most significant issues for digital materials can be attributed to three significant risk categories: media deterioration, technology obsolescence and adverse management conditions. Each of these categories will be discussed in detail later.

Traditional risks posed to physical collections, such as fire or loss, are of negligible effect compared to the inherent weaknesses and challenges of digital materials. Therefore in this analysis they are only attributed minor importance and subsumed into larger categories.

Risk Evaluation

To compare the risk of loss of access between digital collections it is useful to employ a numeric scale and a formula for calculation of total risk.

The method employed here is a semi-quantitative risk analysis as defined by the de-facto international standard for Risk assessment, AS/NZS 4360:1999: Where "… qualitative scales are given values but the number allocated to each description does not have to bear an accurate relationship to the actual magnitude of consequences or likelihood. … The objective is to produce a more detailed prioritisation than is usually achieved in qualitative analysis, not to suggest any realistic values for risk such as is attempted in quantitative analysis."

Each risk factor is assessed and quantified below according to their relevant category. Risk factor inter-dependencies are then evaluated to provide suitable equations for meaningful comparison of results between digital material in each category of risk.

The resultant values can be compared across collections for each risk, but should not be compared between risks for one particular collection, as there is no direct correlation of the numbers.

Scale of Risk

The following five-tier scale in Table 1 is proposed to measure the current risk of losing long-term access to digital collection material.

Table 1. The Scale of Risk Used for this Analysis

Level of risk	Risk value %	Definition
Low	Risk $\leq$ 20	Not applicable or unlikely risk posed
Low to Medium	20 < risk $\leq$ 40	A contributing factor to making access difficult
Medium	40 < risk $\leq$ 60	Moderate current risk
Medium to High	60 < risk $\leq$ 80	Potentially a severe current risk
High	80 < risk	Severe current risk, may have already or will imminently result in complete loss of access

The values on this scale are allocated to the risks in this analysis and specific figures are shown for each risk factor in the following tables.

Media Deterioration

Media deterioration is affected by the type of media, its age, the environmental conditions in which it is stored and used, and the amount of use and handling it receives.

Everybody who has experienced a corrupted floppy disk knows that the deterioration of storage media results in the inability to retrieve data. This is a significant issue when assessing risk, but it must be balanced with the more pressing issues of obsolescence.

Risk values are estimated below but it is not possible to guarantee media longevity figures or predict accurately the life expectancy of digital media because of the significant number of variables involved such as different brands, batches and manufacturing processes used to produce the media over time. The amount of handling, use and storage conditions also contribute to varying the life-expectancy of media, but again only an estimated effect can be derived.

Among other resources, this work refers heavily to a report from the National Media Laboratory (NML) in the USA, prepared under contract for the United States Government for the National Technology Alliance (NTA) Programme. It is an accepted authority on media longevity and an ideal source of information in an area that is frequently compromised by proprietary interests.

Type of Media

There is a wide variety of digital storage media in use, but the majority of library collections can be categorised according to the table below. Each can be described by the type of media which is then used to estimate the stability and fragility of the media. Stability refers to the expected risk of chemical degradation with time, and fragility refers to the likely risk of physical degradation caused by handling and use.

Table 2. Risk Related to Media Type

Media name	Media type	Stability (risk of chemical degradation)	Fragility (risk of physical degradation)
Floppy disks: 5 ¼", 3 ½"	Magnetic	High risk 100%	Medium-High risk 75%
CD : audio, ROM	Optical – physical	Low risk 0%	Low risk 25%
CD-R	Optical – phase change	Medium risk 50%	Low risk 25%
DVD	Optical – physical	Low risk 0%	Low risk 25%
DVD-R	Optical – phase change	Medium risk 50%	Low risk 25%
Video cassettes	Magnetic	High risk 100%	Medium-High risk 75%
Cassettes: audio, data	Magnetic	High risk 100%	Medium-High risk 75%
Hard disks in desk top computers	Magnetic	High risk 100%	Low risk* 0%
Online systems	Magnetic	Low risk* 0%	Lowest risk* 0%
Other (e.g. punch cards)	Paper	Medium risk 50%	High risk 100%

* although magnetic media is actually high risk there is little handling that occurs and a common feature of online systems is backup and redundancy of the data for persistence

Age

The probability or likelihood of occurrence is a common concept in risk assessment, and the measurement of the age of digital materials is used as an equivalent value in this evaluation for risks that are certain to occur eventually.

Age is significant to indicate the likely remaining life expectancy of the media before degradation has rendered it unreadable, and also to indicate the remaining life of the technology generation before it is likely to become obsolete. (See Technology obsolescence) Digital media and formats are unlikely to be usable if unchecked for more than 25 years.[4]

The age of digital materials is assessed in this analysis in four bands as shown in table 3.

Table 3. Risk Related to Age of Media

Age in years	Risk Level	Risk %
<5	Low: New media, new media technology	0
5 to 15	Medium: Media technology changes being introduced	50
15 to 25	Medium – High : Media technology likely to have changed	75
>25	High: Highest risk, certainly technology has changed, approaching media instability	100

Environmental Conditions

The NML report discusses: "Earlier data storage media exhibited some physical and chemical weaknesses and instability due to environmental conditions…. Contemporary magnetic and optical storage media are highly refined products. Some of them have excellent environmental durability and stability …. under air-conditioned environments designed for human comfort.

"However for extended-term magnetic media storage the preferable temperature ranges are between 12-15°C and 30% ± 5% [humidity], which is considerably colder and drier than would be comfortable for humans."

Also from the same report, in summary on the subject of light conditions: "…recording media… prefer darkness."

Therefore, in general, storage areas that comply to a standard such as BS5454:2000 *Recommendations for Storage and exhibition of archival documents* will have suitable conditions for most digital materials.

On the subject of environmental pollutants the report claims: "the conditions comfortable and healthy for human beings are acceptable environments for the normal storage and operation of the storage media". As it is expected that libraries do not contain environmental pollutants unhealthy to humans, digital material storage locations are expected to be acceptable and will not be further assessed for pollutants in this analysis.

Table 4. Risk Related to Environment

Environment	Risk Level	Risk %
Office, Air-conditioned	Low	0
No air-conditioning	Low - Medium	25

Table 5. Risk Related to Handling and Use

Level of use	Risk Level	Risk %
No use		0
Low use of originals, of high use of copies	Low	20
High use by staff, low use by readers	Low - Medium	40
High use of originals by readers	Medium	60

Handling and Use

The amount of handling and use that digital media are subjected to can affect their life expectancy. Physical media can be marked and scratched accidentally or intentionally. Repeated use of magnetic media wears the media and the reading devices. The relationship is linear, high use equals high risk, low use equals lower risk as shown in Table 5.

Technology Obsolescence

Technology obsolescence is a result of the speed of change in computer technology. The storage media, hardware and software are all frequently upgraded and if the digital materials are not upgraded in synchronisation then access to them may be lost.

Storage Media

The assessment of storage media relates to whether the technology is still available to physically read a disk or tape. Media have been grouped under the media formats listed in Table 6 for assessment.

Table 6. Risk Related to Storage Media

Media formats	Issues	Risk %
Online	Networked access, hardware drives not needed	0
Optical – pressed	Common current technology	25
Optical – writable	Some discs susceptible to variations in drives ability to read writable media	50
Magnetic	Increasingly difficult to source appropriate hardware	75
Paper tape/cards	Obsolete	100

Table 7. Risk Related to Data Formats

File formats and data encoding	complexity, proprietary/open, common usage	Risk %
HTML, image, plain text, CD-audio (e.g. .html, .tif, .jpg, .txt, .rtf, .wav)	Simple, single file objects, long life span, common usage	20
Word, PDF, multi-file web sites (e.g. .doc, .pdf)	Proprietary, but possible to transfer, common usage	40
Databases, spreadsheets (e.g. .xls)	Frequently software dependent	60
Executable, interdependent multiple file objects (e.g. .exe) and rare proprietary file types	Proprietary or complex	80
Paper tape/cards (unknown)	obsolete	100

File Formats and Data Encoding

Unless reliable technical metadata has already been recorded in an accessible way for collection material, generalisations about the types of formats in collections must be used to project the risk involved. Table 7 shows an estimated range of common options for library collections and their allocated risk values.

Hardware and Software Required

Similar projections may also be used to approximate the system requirements across digital collections. See Table 8.

Table 8. Risk Related to System Requirements

Software and hardware examples	complexity, proprietary/open, common usage	Risk %
Image viewer, web browser	Common, open , interchangeable, not platform specific	20
Audio accessories	Generic software but added accessories required, not platform specific	40
PC, Mac (platform specific)	Common usage but subject to extreme changes between versions	60
Unknown but probably older than current PCs	Not common usage	80
BBC Micro, cassette player equipment, paper tape reader	Obsolete or highly proprietary or highly complex	100

Age

Based on the production period and manufacturers support for data storage equipment the NML report states: “the active and useful life of a particular product, therefore, does not usually extend beyond twenty-five years.”

Age will affect the stage of technology obsolescence as it does the stage of media deterioration and roughly equates to the same bands that were employed above for media stability. See Table 9.

Table 9. Risk Related to Age of Technology

Age	Risk level	Risk %
Less than 5 years	Low : New storage and system technology	0
5 – 15 years	Medium : Storage and system technology changes being introduced	50
15 – 25 years	Medium-High : Storage and system technology likely to have changed	75
More than 25 years	High : Certainly technology has changed, approaching obsolescence	100

Adverse Management

Another category of risk has been identified and labelled adverse management, which is similar to the 'custodial neglect' coined by Waller[1] for museum collections. In this context it refers to the lack of ability to provide enough support to manage digital resources.

Metadata, including catalogue records, plays a vital role in maintaining access to digital materials. Information about storage media, file formats, object types are necessary to implement various processes for preservation and access.

In order to store and manipulate metadata for resource discovery and preservation management appropriate systems must be developed and populated. Then the digital resources also need to be made accessible and yet also kept secure from accidental and intentional corruption. This all requires significant commitment of planning and resources.

Table 10. Management Issues – Storage System Risk

Storage system issues	Details	Available?	Risk %
Storage system (SS)	Centralised, monitored method of storing digital materials	Yes	0
		No	40
Security (S)	Restricted access to prevent changes or loss	Yes	0
		No	20
Disaster control/ recovery (DC)	Suitable backup processes	Yes	0
		No	20
Financial commitment to storage (FS)	Budgeted funding for storage	Yes	0
		No	20

It is necessary to assess whether systems exist for the management of the digital resources and their metadata for resource discovery and preservation. This will most likely be in the form of metadata creation processes and existing catalogues.

Finance availability will be measured by budgetary or planning commitment to projects for current and future work to develop metadata and management systems.

Suitable methods of system security, backup and recovery need to be in place for digital materials and their metadata systems. This should be a standard feature of modern system design, yet should consciously be accounted for and not taken for granted, particularly when materials may be on hand held media and not stored on a central system.

These management risks can be analysed in two groups. One group are the storage system issues (see Table 10), and the other group are the management system issues (see Table 11). Each group contains the vital aspects required to address the risks in such a way that if any aspect is not addressed the risk for that group increases.

Table 11. Management Issues – Management System Risk

Management system issues	Details	Available?	Risk %
Resource discovery metadata (RDM)	Record that locate digital materials in the collections	Yes, Catalogue records including indicator it is digital	0
		Multiple objects per record, incomplete or accession records only	15
		No	30
Preservation metadata (PM)	Technical details on formats and system requirements, etc.	Yes	0
		No	30
Management processes (Man.)	Methods for tracking digital materials and technology changes	Yes	0
		No	20
Financial commitment to management (FM)	Budgeted funding for creating metadata and preservation	Yes	0
		No	20

Calculations

In order to meaningfully combine and compare the above risk factor values it is necessary to consider how risks will relate to each other. It only takes the occurrence of one risk at its most severe level to destroy the ability to access an item.

Each factor in a category of risk may work in isolation or escalate another risk, for example age will significantly increase the risk of technology obsolescence.

Factors that work together to effect a proportional increase such as media type and level of use should be multiplied, but factors that work additively in a linear increase should be added such as media type stability and environment. When each of these is multiplied to reach the proportional increase by age we achieve values for the estimated risk of media fragility and stability correspondingly, each of which should be compared only and not added.

To analyse risk for collections where a range of risk factor values may apply, as opposed to a single value, taking the average value is recommended for calculations.

Considering these arguments, the following seven equations are derived for the comparison of risks. Refer back to the relevant tables for the values to be calculated.

Media deterioration risk factor calculation

$$\textit{Media stability} = \textit{age (stability level + environment)} \quad (1)$$

$$\textit{Media fragility} = \textit{age (fragility level x use)} \quad (2)$$

Technology obsolescence risk factor calculation

$$\textit{Media obsolescence} = \textit{age (media formats)} \quad (3)$$

$$\textit{File obsolescence} = \textit{age (file formats)} \quad (4)$$

$$\textit{System req. obsolescence} = \textit{age (system requirements)} \quad (5)$$

Adverse Management risk factor calculation

$$\textit{Storage system risk} = SS + S + DC + FS \quad (6)$$

$$\textit{Management system risk} = RDM + PM + Man. + FM \quad (7)$$

It is necessary to remember that all the quoted risk values and equations are semi-quantitative estimations only and would benefit from continual discussion and revision.

British Library Risk Analysis Case Study

The British Library, like many large libraries, had been collecting a variety of digital materials for many years. This material included various disks and tapes, progressing to online formats. The manuscript collections also received legacy digital material from punch cards to Apple computer hard drives and they were considering how to collect live or archived email correspondence.

As the introduction of legal deposit for digital materials loomed large (i.e. a copy of everything published must legally be given to the library) and digitisation programmes dramatically increased it became urgent that there was a more coordinated and consistent approach to archiving and managing digital materials. Methods for assessing the priority and urgency of material to be addressed were required.

Very little management information or technical metadata had been stored for individual items. Therefore generalisations across collections of similar materials were made.

Fourteen collections of digital material across the library were identified and analysed. This involved selecting the appropriate value for each risk factor described in this paper and calculating the risks for comparison using the seven equations listed above.

The results were graphed for an easy visual comparison and can be seen in Figures 1, 2 and 3.

Media Deterioration

Naturally media deterioration was most apparent as a threat to collections which contain older media, such as the Sound Archive, Manuscripts and the older existing purchased/donated collections.

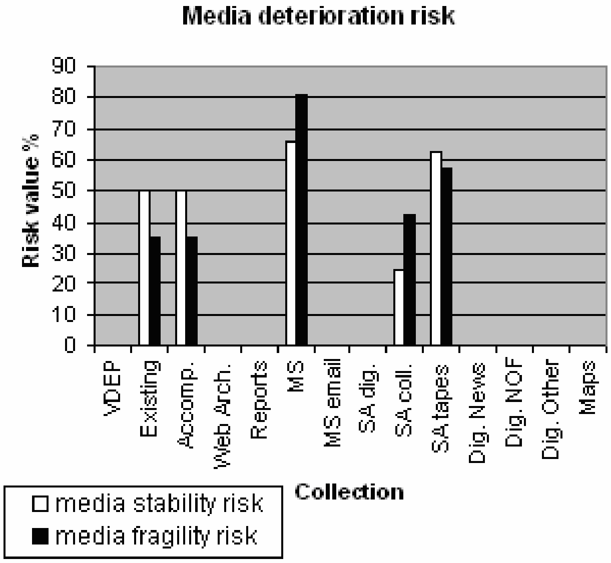

Figure 1. Media deterioration risk comparison

Technology Obsolescence

Again, technology obsolescence is most apparent as a threat to collections which contain older media, such as the Sound Archive, Manuscripts and older purchased/donated collections.

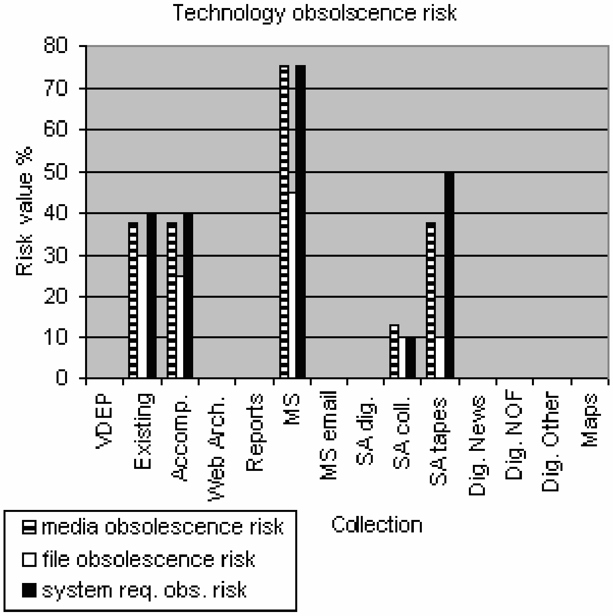

Figure 2. Technology obsolescence risk assessment comparison

Adverse Management

Analysis of adverse management risks showed they are currently the most threatening risks to all materials.

Here the Sound Archive materials that were in higher danger in the other categories demonstrated lower risk among the collections because there are better metadata and management processes in place for audio material. However they still require considerable work and a central storage and management system to properly mitigate these risks.

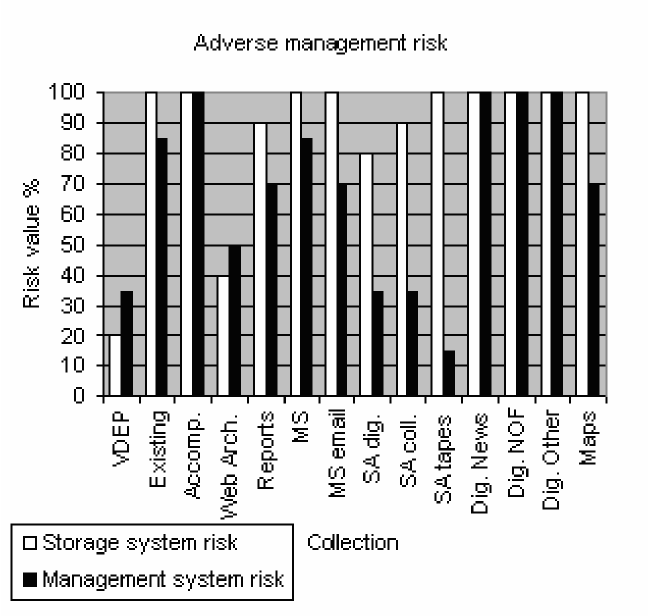

Figure 3. Adverse Management risk assessment comparison

Results and Recommendations

Results of the risk analysis showed the older collections were at high risk, particularly those in poor conditions and/or containing obsolete formats. i.e. the Sound Archive, Manuscripts, Existing purchased/donated collections and accompanying materials.

It was also apparent that implementation of metadata standards and appropriate cataloguing and a centralised storage system will mitigate large management risks for all materials.

Recommendations included regular reassessment of material, because as media and formats age they will have more prominent risk issues. This should include continued discussion and refinement of the analysis methods with other staff and institutions to achieve better quality results.

It was noted that to apply the results in a meaningful way to management decision making, other issues needed to be taken into account, such as the value of materials, the size of collections and the cost of risk mitigation.[5]

Conclusion

Risk analysis is a common and useful tool for the management of collections, however adjustment of traditional methods is needed for assessing digital materials due to their inherent weaknesses. The right method also needs to be chosen for the current point in the life cycle of material.

Preceding the implementation of thorough management practices for digital collections it is possible to assess the significant risks posed to digital materials at high level using the method described.

To apply the results of this risk analysis method to balanced decision making for management, value, size and risk mitigation costs should also be considered.

Acknowledgement

The method and results described here appear through the kind permission of The British Library where they were developed by Deborah Woodyard for the Digital Object Management Programme.

References

1. R. Waller, Conservation risk assessment: a strategy for managing resources for preventive conservation, IIC Preventive conservation preprints, pg.12-16 (1994)
2. A. Stanescu, Assessing the Durability of Formats in a Digital Preservation Environment: the INFORM Methodology, International conf. on Archiving Web Resources, http://www.nla.gov.au/webarchiving/StanescuAndreas.ppt (2004)
3. G.W.Lawrence, W.R.Kehoe, O.Y.Rieger, W.H.Walters, and A.R.Kenney, Risk Management of Digital Information: A File Format Investigation, Cornell University Library (2000).
4. National Media Lab, Data Storage Technology Assessment 2000: Part II: Storage media environmental durability and stability (2000).
5. D. Woodyard-Robinson, Risk Analysis and the Management of Digital Library Material, http://www.woodyard-robinson.com/pub/ (2005)

Biography

Deborah Woodyard is an international freelance digital preservation expert with a background in materials conservation and computing. Starting in digital preservation in 1996 at the National Library of Australia, her work included contribution to the NLA Preservation Metadata Guidelines and the PADI (Preserving Access to Digital Information) website. In 2001 Deborah became Digital Preservation Coordinator at the British Library in London. She was a representative on international projects such as PREMIS (PREservation Metadata Implementation Strategies), managed the initiation of the BL Web Archiving Programme and developed a risk analysis method for the management of digital materials before moving to New Zealand in 2004.