Appendix A

Risk-Assessment Workbook

Contents

Prepared for the Council on Library and Information Resources by:
Gregory W. Lawrence
William R. Kehoe
Oya Y. Rieger
Anne R. Kenney
of Cornell University Library
Ithaca, New York

Introduction

A risk-assessment tool

From our perspective, a successful preservation strategy is created when one or more risk assessments are completed, analyzed, and interpreted by archivists and administrators, and culminate in a clear, well-understood action plan. A risk assessment is simply a means of structuring the process of analyzing your risks. If the risk-assessment methodology is well-specified, different individuals or organizations, supplied with the same information about a digital file, should estimate similar risk values.

This workbook is a risk-assessment tool. What this means and how it is to be used will become clearer as you work through the sections. The workbook will help you identify potential risks associated with migrating digital information, one of several options available for preserving digital information. In fact, our organizations routinely practice risk management. Traditional tasks, such as centrally housing materials, cataloging items to a position on a shelf, and binding loose items together, now have as their digital counterparts the creation of data centers, metadata, and data backups. Digital preservation is in its formative stage, and the use of risk-management procedures established today will seem exceptional until these procedures are integrated into standard practices.

Why a workbook?

In an ideal situation, the best risk assessments would be conducted by a team of experts, each a specialist in a particular area and with general knowledge of digital preservation. However, access to a single expert adviser is a luxury seldom available to archivists and data managers. In place of a human adviser, this workbook attempts to identify the information an expert might seek. When appropriate, the workbook provides definitions and brief issue summaries followed by questions and situation evaluations. It is hoped that this will provide a uniform method of organizing or structuring the assessment process so that all interested parties can be involved to their best advantage. Since digital collections differ appreciably in size, content, and format complexity, this workbook is general in focus. In proceeding through the workbook, you are encouraged to add, delete, or modify the questions to make it more useful to your situation.

Who should use this workbook?

There is a good chance that digital preservation will evolve into a distributed system. If so, it is likely to have the following characteristics:

- It will be hierarchical, with small, specialized organizations interacting with large national coordinating organizations.
- Preservation guidelines will flow from the top down.
- Materials for preservation will flow from the bottom up.
- Data processing and filtering will occur at all levels.

If this system emerges, digital preservation—specifically digital migration—will occur in many organizations, and ultimately embody the collective efforts of information specialists from many professions. Obviously, this is a broad audience for whom to prepare a workbook, especially a workbook on digital migration risk.

High on the list of those whose interest we hope to attract are the archivists, librarians, information managers, programmers, and administrators who oversee specialized digital collections. They will often make first contact with permanent digital materials and may perform the initial migration of these materials. Equally important, we hope to attract any data user who wants to understand the challenges of digital preservation. In general, we assume the reader has a good understanding of computers and software.

SECTION I MIGRATION—ISSUES AND OPTIONS

Definition of Migration

The Commission on Preservation and Access (CPA) and Research Libraries Group (RLG) Task Force on Archiving Digital Information defines digital migration as "the periodic transfer of digital materials from one hardware / software configuration to another, or from one generation of computer technology to a subsequent generation." (Task Force on Digital Archiving 1996:5).

The Task Force defined migration broadly, allowing room for the concept to evolve. Currently, migration can describe the following preservation scenarios:

- The routine refreshing of digital files. Until a few years ago, the transfer of files from one medium to another was central to the issue of migration. With the availability of more reliable storage media, this issue is less pressing than it once was.
- Changing digital formats when files are converted from one application to another. An example of this form of migration would be moving a document from a Macintosh to a Windows 98 operating system.
- Radically changing digital formats. An example is converting word processing files from proprietary formats to ASCII.
- Making derivative copies from digital master formats. Some digital preservation programs adopt a digital master file format not suited for general access and, from this master, generate a copy in a more suitable format. For instance, Tagged Image File Format (TIFF), a master storage format for scanned images, might be converted into a Portable Document Format (PDF) derivative for distribution and easy use.

Why Migrate?

There can be many reasons to migrate, many of which focus on file format. An unstructured or unformatted file is simply a stream of bytes. Software developers structure data files to allow their software to efficiently read or write data to the files. As software applications become more complex, the file formats specified also grow more complex. Ideally, there should be a consistent format of choice for any genre of information. In reality, as software evolves, new or revised formats are continuously displacing older, established formats. This makes the format of choice a moving target.

With this in mind, we would like to advance five possible reasons to migrate.

1. The format is obsolete or its market share is extremely low. The software company may have gone out of business or changed its business focus and stopped supporting the format. Third-party developers who follow market leaders may have abandoned the format. Finally, the format may not be flexible enough to support enhancements in supporting software.
2. The format is dependent on a specific hardware and operating system. If that environment is abandoned or superseded by another system, the only alternatives to migration are to sustain the technology at any cost or to depend upon software to emulate the technology.
3. The format is proprietary, and the vendor will not place the format information in the public domain.
4. Administration of the digital archive requires a simplification of formats. Large archives often have files created by different generations of the same application. Archives may pay unnecessary administrative, computing, and storage costs for maintaining copies of numerous versions of the same application.
5. Metadata requirements are increasing. There is a growing realization that current MARC records, code books and readme files, and file names are insufficient for managing large collections of data files. Embedding metadata may be practical and desirable in future versions of current software formats.

These five reasons are summarized, with examples, in Table 1.

Should I Migrate?

This is the big question, and frankly, we are divided on how to answer it. Migration as a preservation strategy is risky. A major underlying assumption is that someone has sufficient knowledge both of an obsolete format and of its appropriate replacement to prepare a conversion program. For certain specialized, proprietary formats, the format specifications are not publicly available. Also, significant anecdotal evidence suggests that most formats are not fully interchangeable. Knowing what happens to a file or a collection of files inside that conversion program is a mystery to most data managers and archivists. Poorly planned or implemented migration projects may save the content of a file but accidentally lose certain fundamental features of the data that severely diminish its value.

An alternative strategy is emulation. Emulators are programs that mimic computer hardware. Projects adopting this approach store copies of the initial software and descriptions of how to emulate the initial hardware to run the software along with the digital files. Emulation assumes future access to multiple data objects: the data file to be preserved and reused, the application software that generated the

data file, the operating system in which the application functioned, and the hardware environment emulated in software using detailed information about the attributes of that hardware. This complex environment would most likely fail if one or more components were missing. Although emulation is a promising preservation strategy, we have not examined it in depth and make no attempt to evaluate emulation risk in this workbook. For a general overview of emulation as a preservation strategy, we refer you to Rothenberg (1999).

Many experienced digital archivists are fully aware of the current issues and options associated with migration. These professionals may simply require a thorough checklist to be sure they have not overlooked some high-risk activities. For this professional, the workbook can be modified to provide a comprehensive, compact checklist of migration steps.

Many information professionals have little training in digital preservation. These professionals have a steep learning curve to attain the expertise needed to make a sound, informed decision to migrate. A top-to-bottom analysis of their archive may help clarify their migration options. This workbook should prepare them to develop their own migration plan and checklist. These individuals should review the articles listed in the References on p. 43. The articles contain a wealth of information and explain many topics we do not include in this workbook.

PROBLEM	REASON	EXAMPLE
Format is obsolete	Developer is out of business Developer has stopped supporting the software Market share is declining Supporting programs have changed significantly Third-party support is lacking Paradigm has shifted	VisiCalc Borland Dbase (originally Ashton-Tate) WordPerfect Compression software changes for TIFF Common Ground Flat file to object database
Format depends on obsolete hardware or operating system	Files operate only if entire system is maintained	Commodore 64/128 Apple II
Format is proprietary	Vendor will not share format information, even if superseded	Xerox XDOC format
Administrative oversight is diffused	Files exist in related formats, different generations of same application	TIFF 4.0, 5.0, 6.0
Metadata management is complex	Use of embedded metadata increases with growth of metadata requirements	8.3 file name format (i.e., table1.wk1)

Table 1. Reasons for migration

SECTION II RISK ASSESSMENT AND MEASUREMENT

Introduction

Digital information is seeded with hazards. A common example of a digital hazard is a documentation file created in a word-processing application. Prepared on a Macintosh computer, this file will be imported into another application on an Intel-based computer. The chance or probability that you will not be able to read the file on the PC is considered your risk. If you are sure that you cannot read the file, your risk is 100 percent, and you have a problem. As you consider the hazards associated with the file and review software options available, you are performing a risk analysis. If during that analysis you prepared a list of risks in order of their importance, you have performed a risk assessment.

As mentioned earlier in this workbook, risk assessment is simply a means of structuring the process of analyzing your risks. The significance of risk estimates provided by the assessment should be easily understood and should contribute to a consistent and credible predictive process. With these thoughts in mind, we would like to make two points related to defining and measuring risk.

Defining Risk

Numerous professions measure and define risk with a unique vocabulary and context. To illustrate the difficulty in defining risk, consider the following definitions, drawn from the fields of environmental science, business, and computer science, respectively:

"The probability of a prescribed undesired effect. If the level of effect is treated as an integer variable, risk is the product of the probability and frequency of effect [e.g., (probability of an accident) x (the number of expected mortalities)]. Risks result from the existence of hazard and uncertainty about its expression." (Reinert, Bartell, and Biddinger 1994)

"Risk is a concept that auditors and managers use to express their concerns about the probable effects of an uncertain environment." (McNamee 1996)

"A risk is any variable on your project, which you may or may not have control over, that could take on a value within its normal distribution of possible values that either endangers or eliminates the possibility of project success." (Lister 1997)

We could provide more examples to illustrate our point. Clearly, the degree and types of risks associated with any migration activity may be understood differently by administrators, colleagues, and data users. This in itself may be a hidden, but significant, risk.

Measuring Risk

Measuring risk is as problematic as is defining risk. One paper we examined correlated risk level with the nonlinear relative probability of risk occurring. The author normalized risk levels to obtain a meaningful quantification (Kansala 1997). In another paper, a university research group indicated that cases where one can accurately assess the probability of a future event are rare because the information technology environment for software changes so rapidly. They preferred simple estimates, such as *high*, *medium*, and *low*, which facilitate decision making. The probability of risk is hard to quantify, and risk-measurement scales, like risk definitions, are highly contextual. (Williams, Walker, and Dorofee 1997).

Workbook Risk Scales

For this workbook, we generated two migration risk-assessment scales: one to measure the *probability* that a hazard would occur; and another to measure the *impact* of that hazard, should it occur. These scales were prepared for a risk-assessment case study of a collection of numeric files, the test bed for much of our project. The scales are provided here and used throughout the workbook to illustrate how one measurement system was applied and evaluated. Admittedly, the proper use of any measurement process requires an understanding of the material under analysis. Also, the measurements lack scientific precision. At the end, you do not sum the results and decide to migrate on the basis of a single number. However, using assessment scales requires you to think in terms of probability and impact, and this can help you set priorities in identifying the steps for a migration project.

The *risk probability scale* has three related pieces of information: a label, a ranking value, and a description. The scale is not linear in that benchmarks for risk are skewed toward lower probabilities.

Risk Probability Scale

Label	Value	Description
Very High	**5**	A probability estimated between 26–99%
High	**4**	A probability estimated between 11–25%
Moderate	**3**	A probability estimated between 6–10%
Low	**2**	A probability estimated between 1–5%
Very Low	**1**	A probability estimated below 1%

The *impact scale*, shown below, also has three related information items: a label, a ranking value, and a description. Since we are focused on the migration of digital information, our impact focus is loss of data. Other impact scales for digital information could be generated.

Benchmarks for this scale are the difficulties associated with recreating corrupted or lost digital information. "Catastrophic loss" refers to a total loss of information that cannot be recreated from any other source—digital, print, or artifact. An example of a catastrophic loss would be the total loss of the sole archival TIFF image of a painting destroyed in a fire. "Serious loss" is the total loss of a digital file that could be recreated from other sources. In this situation, we are thinking of composite documents, not just the conversion of a single artifact. The least impact value would be applied to lost files that can be reconstructed from other digital documents.

Risk Impact Scale

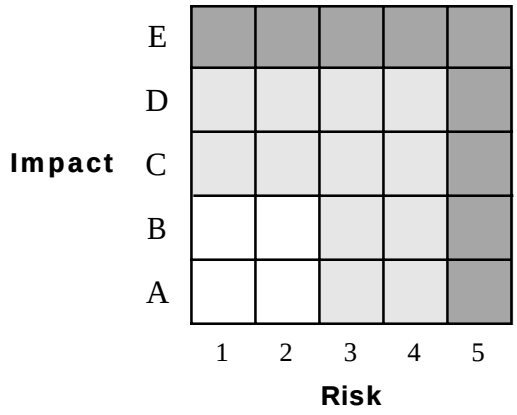
Label	Value	Description
Catastrophic	**E**	Complete, irreversible loss of data. Data cannot be drawn from other sources—print, artifact, or digital.
Very Serious	**D**	Partial, irreversible loss of data. Data cannot be drawn from other sources.
Serious	**C**	Complete loss of data. Data can be fully reconstructed from other sources.
Significant	**B**	Partial loss of data. Data can be fully reconstructed from other sources.
Minor	**A**	Complete or partial loss of data. Data can be copied from other data files.

Recording Risk Assessments

In our prototype scale, we recorded the risk probability value with the impact value as a single value. For example:

5E = Very high probability of occurrence with a catastrophic impact
3D = Moderate probability of occurrence with a very serious impact
2C = Low probability of occurrence with a serious impact
1B = Very low probability of occurrence with a significant impact
1A = Very low probability of occurrence with a minor impact

The combined values are easy to map in a two-dimensional decision matrix, using the probability and impact scales for the *x* and *y* axis, respectively. The grid provides a visual display of the overall state of risk that is described in the workbook.



The decision table yields the following outcomes:

1. If all assessment question responses fall within the white grid cells (1A-B, 2 A-B), the migration process is likely to pose low risk. With due caution, the migration can be carried out.
2. If assessment question responses fall within the gray shaded grid cells (1C-D, 2C-D, 3A-D, 4A-D), the migration process is likely to have high risk. Migration activity should be postponed until the risk probability of these items can be reduced.
3. If any assessment question responses fall within the dark gray grid cells (1E, 2 E, 3E, 4E, 5 A-E), migration of files is ruled out.

SECTION III SOURCE/TARGET FORMAT ASSESSMENT

Introduction

A common illusion used by magicians involves pushing a colored cloth into one end of a black box and removing a different-colored cloth from the other end. As spectators, we don't know what is going on in the black box, but to enjoy the illusion, we assume something in the box changes the color of the cloth.

This is a good analogy for file migration, where a program reads a file with one format and a new file with a different format appears.

In this instance, the "black box" is not magic, but a software application. These application programs include the following types:

- a translation program that is written by an archivist for a specific project.
- a commercial translation program written for a specific purpose. For example, some products extract data fields from numerous files with different formats and create a new data product with a different format.
- a general-purpose commercial translation program; for example, a program that translates files between PC and Macintosh file formats.

Each approach has its benefits and liabilities. Programs developed at an archive provide extensive knowledge about the functions of the translation software, but they have lengthy development cycles and are often expensive to prepare. Off-the-shelf, commercial programs provide little information about the translation process but provide many features at a low cost.

A format risk assessment should be able to gauge the following three distinct areas of risk:

1. The risk created by the conversion program. This risk can be assessed by evaluating the state of known test files before and after the conversion process. Assume that you can generate a comprehensive test file or files that contain all the known attributes (features) of a specific format. The conversion software would process the test file(s) and create new files in a different format. Following the conversion, you would carefully examine the new file(s) to verify that all the attributes of the original file(s), and nothing else, were faithfully reproduced. Although this method is laborious, it was quite accurate for the formats we tested. If these results were independently verified elsewhere, a documented migration path would be available for use internationally.
2. Recurring risk inherent in a large, heterogeneous collection of data files. Assume that you have established the attributes at risk in a specific format. Also assume you have 10,000 files that may contain one or more of these at-risk attributes. One way to quantify the files that may contain these at-risk attributes would be to have a file reader examine each file and identify the file, its location and suspected attributes associated with that file.
3. Functionality of the conversion software. If several conversion programs are available, each will provide some or all core functions as well as optional features. General performance benchmarks that can be tailored for specific migration scenarios will provide some uniformity of measurement. An example of a rudimentary assessment for these features is provided in "Conversion Software Functionality Assessment," available at the project Web site (http://usda.mannlib.cornell.edu/reports/clir/CLIRConvSoftAssessment.pdf).

Conversion Software

The use of file conversion software has been a common practice for many years. Most conversion programs have been privately prepared and are very costly, or have been bundled into application software by developers for specific file formats. Recently, third-party vendors have begun to release inexpensive conversion software that can convert numerous file formats. It is important to analyze the cost, benefits, and risks associated with either locally developed or commercial off-the-shelf software.

3.a. Which form of conversion software do you expect your organization to implement for your archive?

- ❒ Locally developed
- ❒ Off-the-shelf commercial

3.b. If you answered "Off-the-shelf," have you been able to identify a software application to translate your data files?

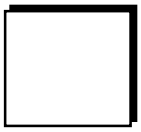
- ❒ Yes, for all project files
- ❒ Yes, for some project files
- ❒ No

Are there conversion software issues that remain unresolved for you? Could these issues create a risk for the files that might be converted? Can you assign a probability that these risks might occur? If damage or loss were to occur, how difficult would it be to recreate the data?

3.c. For each format identified for migration and using a locally developed or a commercial product, can the conversion software perform any or all of the following functions?

- ❒ Identify and select files that have the source format
- ❒ Process multiple files
- ❒ Identify and bypass files with potential conversion problems
- ❒ Generate processing or error reports, or both
- ❒ Provide online assistance

Risk-assessment value (1-5):
Impact-assessment value (A-E):
(Example: High Risk / Catastrophic = 5E)



Format

We are often concerned with the state of the file before and after conversion. The *Source* file format is the format that will be converted into a different format. The *Target* file format is the new file format present following conversion. Target formats tend to fit into one of the following three categories:

1. ASCII. The simplest representation of data, ASCII consists of a limited set of letters, numbers, and symbols. ASCII has been the archival format of choice for tabular numeric data and simple text files. ASCII cannot preserve images or many complex data structures.

2. Formats that conform to standards informally agreed upon by digital coalitions or archival organizations, or accepted by most data users. TIFF has not been formally adopted as the standard image format, but it has strong support among digital coalitions and archives.
3. Formats that are backward-compatible within applications. Lotus 1-2-3 .wk1-.wk4 formats are supported by Lotus Millennium.

Before deciding which category of target format to select, it is important to consider two questions. First, does the target format suit the purpose of the source file, for both the archive and the data users? Second, is the target format technically suitable for long-term access? The following questions about source / target formats can serve as a filter to identify appropriate formats for conversion.

3.d. Is the purpose of the proposed target format the same as the purpose of the source format?

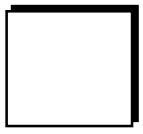
❒ Yes
❒ No

3.e. Is the target format a widely accepted standard, either de jure or de facto?

❒ Yes
❒ No

If you answered "No" to question 3.e., can you identify problems that might arise from using a nonstandard format? Can you assign a probability that they might occur? If these files are damaged or lost, how likely is it that you will be able to replace the lost data?

Risk-assessment value (1-5):
Impact-assessment value (A-E):



3.f. Do users have a readily available means of viewing or using the target format?

❒ Yes
❒ No

Some formats may be good choices for long-term preservation but are difficult for patrons to use. You may wish to consider whether a format that promotes low use presents a risk for the long-term preservation of that file.

3.g. Will conversion to the target format preserve the "functional experience" of the source?

❒ Yes
❒ No

Think of "functional experience" in this way: If the source file were created in a spreadsheet, would the target file format upload into a spreadsheet application and provide the same basic "look and feel"?

3.h. Is there organizational support for the format and related applications?

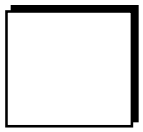
- ❒ Yes
- ❒ No

If you answered "No" to questions 3.h. or 3.i., will the lack of format support within your organization or by a developer create a measurable risk for the files in question? If so, could these files be recovered if they were damaged or lost? How?

3.i. Is there developer support for the format and related applications?

- ❒ Yes
- ❒ No

Risk-assessment value (1-5):
Impact-assessment value (A-E):



SECTION IV SYSTEM ASSESSMENT

Introduction

All computers operate on the same fundamental principles. You might think that the hardware and software of large networked systems would be quite different from that used on your desktop. However, both systems have the same component parts and fulfill the basic functions necessary to any computer system. As computers have evolved, numerous different hardware designs have been developed. In addition, many operating systems and computer applications have become available. The long-term preservation of a digital file is directly affected by the working environment, which is determined by the hardware configuration and operating system.

Hardware

A computer system is made up of several hardware components. The principal elements are as follows:

CPU (central processing unit), which does the actual computing. Different generations of computers are described by their CPU, which provides a rough indication of the currency or obsolescence of a specific system.

RAM (random access memory), the main memory in a computer.

Secondary storage devices, such as diskettes, hard drives, magnetic tape reels or cartridges, and optical disks.

Peripheral devices, also known as input/output (I/O) devices. These include the keyboard, mouse, monitor, printer, modem, and network card.

4.a. What is the general state of your system computer hardware?

- ❒ New
- ❒ Midlife
- ❒ End of lifetime

4.b. What is the status of your system CPU?

- ❒ Current generation
- ❒ Superseded by one generation
- ❒ Superseded by two or more generations

4.c. What is the status of your system memory?

- ❒ Optimal
- ❒ Adequate
- ❒ Needs upgrade

4.d. Do you plan to replace or upgrade your CPU?

- ❒ Yes
- ❒ No

4.e. What is the status of your system storage medium?

- ❒ New
- ❒ Midlife
- ❒ End of lifetime

4.f. Do you plan to replace or upgrade your storage medium?

- ❒ Yes
- ❒ No

These questions are intended to identify whether you need to plan a hardware change. If you migrate files to a new format, will they operate in the current hardware configuration? Equally important, does the general state of your computer hardware create a risk you can measure? Fairly reliable measurements can be formulated using product specifications. Also consider asking whether changes to the hardware configuration add new risk factors. If you have a hardware-related problem, how do you think it will affect the archive?

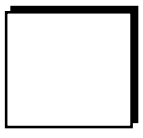
4.g. What is the current state of your system's peripheral devices?

- ❒ New
- ❒ Midlife
- ❒ End of lifetime

4.h. Do you plan to replace or upgrade any of your peripheral devices?

- ❒ Yes
- ❒ No

Risk-assessment value (1-5):
Impact-assessment value (A-E):



Operating System Software

An operating system (OS) is a set of control programs that manage the computer's resources and create a well-defined software environment for computer applications. Common examples of operating systems are the Macintosh, Windows, and UNIX systems. An OS has two levels of functionality. The first is the level seen by the user running applications and issuing system commands. The second is at

the system level, where primitive functions, such as reading from or writing to a file, occur. Data files that can be read by more than one OS are said to be more "portable" than those that are limited to a single OS.

4.i. Before migration, do you expect to change your computer operating system? If so, indicate the type of change.

- ❒ Return to previous version of same OS
- ❒ Minor upgrade
- ❒ Next-generation upgrade
- ❒ Switch OS

4.j. Before migration, do you expect to change your data organization. . .

1) information density on storage devices?

- ❒ Increase
- ❒ Decrease

These questions are more likely to be answered at data archives storing files on large servers. An OS change can have a big impact on system utilities and programs installed to support a specific format. If data files migrate to a new format, will the new OS programs support that format? If not, does this create a risk you can measure? Will this risk have an impact on the archive?

2) hierarchical organization of files?

- ❒ Yes
- ❒ No

3) proprietary file management system?

- ❒ Yes
- ❒ No

Risk-assessment value (1-5):
Impact-assessment value (A-E):

Data Compression

Data compression is a technique used to reduce the size of a file. The goal of compression is to represent a file, at some required quality level, in a more compact form. Compression operations seek to extract essential information from a file so the original data sequence can be accurately reconstructed. Nonessential information is discarded. *Lossless compression* preserves the exact data content of a file. *Lossy compression* preserves a specific level of data quality but does not preserve the absolute data content of the original. The compression ratio is measured by dividing the original data size by the compressed data size. The higher the ratio value, the smaller the compressed file has become. Compression is often done in preparation for file storage or transport. You may wish to analyze the data-compression risk for each format migrated.

4.k. Are the data in your collection compressed?

- ❒ Yes
- ❒ No

If yes, what percentage of the collection is compressed?

_____ %

If yes, is the current data-compression schema lossy?

- ❒ Yes
- ❒ No

4.l. Certain file formats specify a compression standard. If you migrate your files to a new format, have you reviewed the format specifications and will you continue to use the same compression method?

- ❒ Yes, without change
- ❒ Yes, but implementing latest revision
- ❒ No, will replace with another compression method
- ❒ No, will not compress files

After reviewing your data-compression practices, can you identify any risks that might occur during a file migration? If risks exist, can you assign a probability that you can measure? If a compression-related problem occurs during migration, will it have an impact on the archive?

Risk-assessment value (1-5):
Impact-assessment value (A-E):

Security

A *secure* information system is one that maintains the integrity of the information stored in it. The system does not corrupt the data or allow accidental changes to it. Data corruption may be malicious or accidental, or it may be the result of careless handling or oversight. Wherever information is stored, it is important to verify the authenticity of data. Encryption, which entails attaching a code to a file, is a common method of managing data authentication.

Most computer malfunctions are caused by humans. Considering all the persons who have read/write access to data in your archive, and whether you have experienced data loss in the past, you might be able to assign a risk probability that such a loss can happen again and how difficult it would be to undo it. You may also want to examine the risks posed by user access to the data while a migration project was under way.

4.m. Who has read/write access to your data?

- ❒ Archive staff
- ❒ Organizational staff
- ❒ Trusted data users

Risk-assessment value (1-5):
Impact-assessment value (A-E):

4.n. Are your documents encrypted or watermarked?

- ❒ Yes
- ❒ No

If you encrypt your data, will this pose a problem for migration? (See Section III and think about conversion software.) Does encryption pose a risk you can measure? Will this risk affect migration of data? Would lost data be difficult to recover?

Risk-assessment value (1-5):
Impact-assessment value (A-E):

SECTION V METADATA

Introduction

Information is required to properly represent digital information held in archives, hence the need for metadata. Recent research seems to recommend at least three pieces of metadata: 1) a descriptive piece, which provides bibliographic information similar to that of a MARC record; 2) a history piece, which describes the life cycle changes applied to the data; and 3) a content piece, in which structural information (e.g., fields and field values) can be recorded. The history piece may be the most appropriate location to record information about how, what, and when migration was done.

For a good discussion about different forms of metadata records, consult Lagoze (1996), Consultative Committee for Space Data Systems (1999), and Dublin Core Metadata Initiative (1999).

5.a. Do you maintain documentation for the data in your archive?

- ❒ Yes
- ❒ No

If you answered "No," consider your files to be at high risk. Can you indicate why you do not maintain documentation for these files?

Notes:

5.b. Do you maintain publicly accessible documentation for this data collection?

- ❒ Yes
- ❒ No

5.c. If your documentation is in print format, do you plan to convert it into digital form?

- ❒ Yes
- ❒ No

5.d. What is the primary purpose of your metadata?

- ❒ System needs
- ❒ User needs

5.e. If you have metadata for both needs, which receives more attention from you or your staff?

- ❒ System needs
- ❒ User needs

5. f. Do you plan to revise the metadata during or after the data migration?

- ❒ Yes
- ❒ No

Can you estimate how many pieces of metadata you will have to revise? If so, what is that number?

_____

5.g. Are there content standards for both the source and target metadata, such as the Federal Geographic Data Committee (FGDC) Content Standard for Geospatial Metadata?

- ❒ Yes, for both
- ❒ Yes, for only the source or target metadata
- ❒ No

5.h. Is any part of your documentation in a proprietary format?

- ❒ Yes
- ❒ No

If your documentation is in a proprietary format, or if metadata are embedded in a file with a proprietary format, does this imply the documentation suffers the same risks as the data do?

5.i. Do the source or target metadata formats comply with or support standards for searching or resource discovery or both?

- ❒ Yes, for both
- ❒ Yes, for only the source or target metadata
- ❒ No

5.j. For either the source or target metadata, is there software to facilitate conversion to other metadata standards?

- ❒ Yes, for both
- ❒ Yes, for only the source or target metadata
- ❒ No

5.k. Is any part of your documentation embedded in the data file(s)?

- ❒ Yes
- ❒ No

If no, do you intend to embed metadata into files during migration processing?

- ❒ Yes
- ❒ No

5.l. For the purposes of migration, a historic record may be more important than a content record. Do you have, or can you create, a historic record for each file or file aggregation being migrated?

- ❒ Yes
- ❒ No

5.m. If you migrate or revise your documentation, will you need to modify system links or required programs?

- ❒ Yes
- ❒ No

5.n. In how many locations is archival data documentation stored?

- ❒ One location
- ❒ More than one location

If more than one location, do you have a plan to keep all locations up to date?

- ❒ Yes
- ❒ No

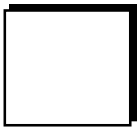
5.o. Do you plan to modify file names during migration?

- ❒ Yes
- ❒ No

5.p. Do you plan to modify system "scripts" or files dependent on file names or file paths?

- ❒ Yes
- ❒ No

Risk-assessment value (1-5):
Impact-assessment value (A-E):



SECTION VI ORGANIZATIONAL ASSESSMENT

Introduction

A digital migration project does not occur in a vacuum. Anyone planning such a project must consider many factors: the size and scope of the project, file content and structure, the project budget, the number of staff involved, and other variables. The successful completion of the project will depend upon the support it receives from the organization and the resources at its disposal. Attempts to preserve digital information may fail if they concentrate solely on a narrow set of technical issues and do not consider the broader managerial issues. Promising technologies cannot be applied without management's understanding and control. Unfortunately, each data collection will have a different management philosophy and struc-

ture, which will impose its own priorities on preservation issues and practices. With this in mind, in this workbook we narrow our examination of organizational risk to four key areas: presevation planning, budgets, staff development associated with program needs, and communication with data users.

1. Preservation Plans

Heroic and ad hoc responses to preservation crises consistently fail to mobilize organizational resources in a comprehensive, meaningful way. Recurrent problems, regardless of the cause, appear wasteful and may diminish support for preservation. In contrast, preservation plans provide guidelines for accepted policies and practices, identify essential resources available for preservation activities, and, ultimately, better conserve information. Migration as a strategy will succeed only if it is consciously integrated with other preservation practices. With that said, there is something about preservation plans that fail to motivate an organization. In some situations, drafting a preservation plan is a paper exercise that, once completed, is filed and forgotten. In others, the plan lacks a strong advocate to secure organizational support and funding. Depending upon the circumstances, a precise and easily implemented plan may be superior to an authoritative manifesto.

2. Preservation Program Budgets

Budgets, like planning, direct digital preservation efforts. Funds for certain preservation activities, such as a migration project, simply may not be available. Or, following a catastrophe, funds that are allocated for preservation activities are insufficient to deal with a large data loss. It is difficult to alter budgets for situations that occur unexpectedly or at random. In many cases, institutions cannot redirect funds to purchase emergency services or replace worn-out equipment. Also, spending priorities and service contracts may emphasize one technology at the expense of others. Preservation budgets will be a source of risk in organizations where preservation is a minor activity in overall operations, or where it is not regarded as an essential activity.

3. Preservation Staff

A migration project requires the skills of many professionals within your organization, some of whom are not under your supervision. Digital information may be well understood by some staff members. For others, it may be something new and different. To achieve the goal of low-risk management of digital information, staff members must become competent technical and managerial problem solvers. Time and training are necessary to integrate these individuals into a motivated, self-directing team.

4. User Community

Finally, the organization must understand how a migration project will affect its user community. The stronger the user community's

interest in preservation, the greater the likelihood preservation choices will be successful and beneficial. The community of users is more likely to support preservation efforts if they participate in important decisions. Where there is no strong user interest in preserving certain information, the data managers may need to review whether it is worth committing resources for its migration.

Planning

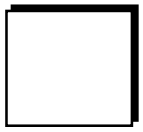
Digital preservation begins with planning. The purpose of planning is to identify significant risks and establish solutions that minimize or eliminate those risks.

6.a. Does your organization have a digital preservation plan?

- ❒ Yes
- ❒ No

If you answered "No" to question 6.a., does *not* having a digital preservation plan create a risk you can measure? Will this risk have an impact on the archive?

Risk-assessment value (1-5):
Impact-assessment value (A-E):



Can the organization's administration use the plan to understand how a format migration strategy fits into the operations of the archive?

Notes:

6.b. If you answered "Yes" to 6.a., has the plan been thoroughly reviewed by the organization's management?

- ❒ Yes
- ❒ No

Someone suggested we simply ask, "Is there a preservation plan, and if so, where is it?" To the point, but maybe missing the point. If a preservation management plan is not a useful, often-referenced document, does that suggest something is lacking? Most likely, there will need to be a revision if format migration is implemented.

6.c. If you answered "Yes" to 6.a., is the plan

- ❒ Readily available to archive staff?
- ❒ Readily available to the organizational management?
- ❒ Readily available to archive stakeholders?
- ❒ Regularly reviewed?

Financial

In this section, several questions are asked about the value of the data archive and the costs to maintain it. At first glance, the information requested may seem difficult to quantify. Try to answer the questions, even if you must guess the first time. After several attempts at working on this section, these estimates will become more

refined and will provide useful figures for discussion and documentation. If you are considering more than one migration project, you may wish to apply this section to each individual project.

6.d. In some respects, money spent on digital preservation efforts is an investment an organization makes to ensure continuing access to the information. In this sense, the value of the data, or the cost of not having the data, should increase with time. At this time, can you estimate the monetary value of the data in the archive?

❒ Yes
❒ No

If you answered "No" to question 6.d., is there a problem measuring the value of the data in the archive? (Some archives will be unable to assign a monetary value to their holdings. Another measure would be the cost of substituting another data product.)

Notes:

If you answered "Yes" to 6.d., what is the estimated value of the archive?

$_____

How did you calculate this value?

Notes:

6.e. Do you have an annual budget for digital preservation activities?

❒ Yes
❒ No

Digital preservation can include, but is not limited to, migration, emulation, refreshing, scanning, metadata creation, and related activities.

If you do not have a regular budget for digital preservation work, or if the budget is demonstrably insufficient, does this create a risk that you can measure? Are there problems that could be resolved with extra funds? If you have problems that persist, what impact would they have on the data archive?

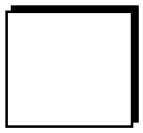
If you answered "Yes" to 6.e., what is your budget?

$_____

6.f. Is your budget sufficient for routine digital preservation activities?

- ❒ Yes
- ❒ No

Risk-assessment value (1-5):
Impact-assessment value (A-E):



A large organization may have several migration projects under consideration. Questions 6.g. and 6.h. can be applied to each project separately.

6.g. Can your current budget fund a migration project?

- ❒ Yes
- ❒ No
- ❒ Uncertain

If yes, enter the amount you can allocate to this purpose.

$_____

6.h. In your estimation, will these funds be

- ❒ Sufficient?
- ❒ Insufficient?

Personnel

Rarely does an organization have enough staff to meet the responsibilities of current programs as well as emerging projects. This problem is aggravated by the fact that new technologies demand rapidly evolving skills.

6.i. How large is the preservation staff?

_____ Full-time employees
_____ Part-time employees (FTE)

6.j. In your estimation, is the number of staff:

- ❒ More than sufficient?
- ❒ Sufficient?
- ❒ Insufficient?

6.k. Have you identified all the skills required to maintain a data archive, including those required to conduct a file migration project?

- ❒ Yes
- ❒ No

6.l. Can your organization provide staff who have the skills required to complete a file migration project?

- ❒ Yes
- ❒ No

6.m. Will a migration project draw staff away from other projects?

- ❒ Yes
- ❒ No

6.n. Can you estimate how long the migration project should take? If so, indicate the approximate time.

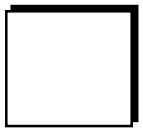
- ❒ Less than 3 months
- ❒ 3–12 months
- ❒ More than 12 months

6.o. Can you expect to have the same staff who begin the migration project complete the project?

- ❒ Yes
- ❒ No

A migration project will require a sustained period of analysis, planning, implementation, and evaluation. Downsizing has created lean organizations. It is quite likely the staff who begin the project may not be assigned to complete it. Are the current staff resources a potential risk to a migration project? Can you assign an approximate probability of a serious mistake occurring? Can you imagine the possible staff errors that would occur? Would these errors have a significant impact on the archive?

Risk-assessment value (1-5):
Impact-assessment value (A-E):

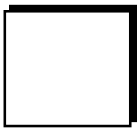


6.p. Does your organization need to contract or obtain outside assistance for

- ❒ Minor component(s) of the project?
- ❒ Major component(s) of the project?
- ❒ The complete project?

If you plan to contract part or all of a migration project, can you identify risks that might have an impact on the archive? Can you measure these risks?

Risk-assessment value (1-5):
Impact-assessment value (A-E):



Data Users

Ultimately, the data user is the primary reason to maintain the digital collection. Understanding the data users and their interests will help clarify the requirements for the system, improve the match between data structure and user needs, and improve the archive's overall usability.

6.q. The logical starting point for an examination of user characteristics is to determine the users' identity. A user community can comprise organizations, individuals, or both. For your data archive, do you have a well-defined constituency?

- ❒ Yes
- ❒ No

Migration of files to a new format will have a significant impact on the data user. If your data users are not involved in the decision to migrate and the planning that follows, will this create a risk you can measure? (How about a volume of protest?) Would user dissatisfaction have an adverse impact on the archive?

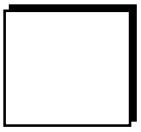
6.r. Data users may or may not be stakeholders in your archive. Stakeholders are interested individuals or groups who have a voice in the various aspects of the archive's implementation. Are data users stakeholders in the archive?

- ❒ Yes
- ❒ No

6.s. If you answered "Yes" to question 6.r., can you describe how your data users are involved in preservation decisions?

- ❒ Constituents heavily involved
- ❒ Constituents routinely consulted
- ❒ Constituents contacted only as needed

Risk assessment value (1-5):
Impact assessment value (A-E):



REFERENCES

Dublin Core Metadata Initiative. 1999. The Dublin Core: A Simple Content Description Model for Electronic Resources. Available from http://purl.org/DC/index.htm.

Consultative Committee for Space Data Systems. 1999. Reference Model for an Open Archival Information System, Red Book, Issue 1 (CCSDS 650.0-R-1). Available from http://wwwdev.ccsds.org/documents/pdf/CCSDS-650.0-R-1.pdf.

Kansala, Kari. 1997. Integrating Risk Assessment with Cost Estimation. *IEEE Software* May/June 1997:61-7.

Lagoze, Carl. 1996. The Warwick Framework: A Container Architecture for Diverse Sets of Metadata. *D-Lib Magazine*, July/August 1996. Available from http://www.dlib.org/dlib/july96/lagoze/07lagoze.html.

Lister, Tim. 1997. Risk Management is Project Management for Adults. *IEEE Software* May/June 1997:20,22.

McNamee, David. 1996. Assessing Risk Assessment. Available from http://www.mc2consulting.com/riskart2.htm.

Reinert, Kevin H., Steven M. Bartell, and Gregory R. Biddinger, eds. 1994. *Ecological Risk Assessment Decision-support System: A Conceptual Design*. Pensacola, Fla.: SETAC Press.

Rothenberg, Jeff. 1999. *Avoiding Technological Quicksand: Finding a Viable Technical Foundation for Digital Preservation*. Washington, D.C.: Council on Library and Information Resources. Available from http://www.clir.org/pubs/reports/rothenberg/contents.html.

Task Force on Archiving of Digital Information. 1996. *Preserving Digital Information. Report to the Commission on Preservation and Access and the Research Libraries Group*. Washington, D.C.: Commission on Preservation and Access. Available from http://www.rlg.org/ArchTF.

Williams, Ray C., Julie A. Walker, and Audrey J. Dorofee. 1997. Putting Risk Management into Practice. *IEEE Software* (May/June):75-82.