Risk Management of Digital Information:

A File Format Investigation

by Gregory W. Lawrence

William R. Kehoe

Oya Y. Rieger

William H. Walters

Anne R. Kenney

June 2000

Council on Library and Information Resources
Washington, D.C.

ISBN 1-887334-78-5

Published by:

Council on Library and Information Resources
1755 Massachusetts Avenue, NW, Suite 500
Washington, DC 20036

Web site at http://www.clir.org

Additional copies are available for $20.00 from the address noted above. Orders must be prepaid, with checks made payable to the Council on Library and Information Resources.

∞ The paper in this publication meets the minimum requirements of the American National Standard for Information Sciences—Permanence of Paper for Printed Library Materials ANSI Z39.48-1984.

About the Authors

Gregory W. Lawrence is government information librarian at Cornell University's Albert R. Mann Library. He has participated in numerous research and development projects concerning the implementation of the electronic library, and he speaks and writes on these topics. He has received the American Library Association (ALA) Best of L.R.T.S. Award, the ALA Blackwell North American Scholarship Award, and the United States Department of Agriculture (USDA) Secretary's Honor Award. Mr. Lawrence is a past chair of the Preservation Committee, Depository Library Council.

William R. Kehoe is programmer / analyst specialist in the Information Technology Services Department at Cornell University's Albert R. Mann Library. He wrote his first assembly language "Hello, World" program in 1978, and has used C, Perl, VisualBasic, and Java to develop business applications and digital library delivery systems. He is currently the system architect for a service that will deliver interactive maps over the Web from numeric data supplied by the National Agricultural Statistics Service. For his participation in the USDA Economics and Statistics System, he received the USDA Secretary's Honor Award.

Oya Y. Rieger has been a librarian at Cornell University for eight years, where she has held positions as numeric files librarian, USDA Economics and Statistics System Project coordinator, and gateway manager for Cornell University Library's Web-based information system. She has participated in several development projects related to electronic libraries and user support services and has written and spoken frequently on these topics. In her current position as coordinator of the Digital Imaging and Preservation Research Unit, she manages a range of digital imaging and preservation research, demonstration, and training projects. She is the coeditor of *RLG DigiNews*. Ms. Rieger and Anne Kenney have written a new monograph, *Moving Theory into Practice: Digital Imaging for Libraries and Archives*, which was recently published by the Research Libraries Group (RLG).

William H. Walters was the social science bibliographer in the Albert R. Mann Library at Cornell University. Soon after the completion of this report, he accepted the position of collection development librarian at St. Lawrence University. A Ph.D. candidate in sociology (demography) at Brown University, Mr. Walters has conducted research in librarianship, demography, cartography, and economic sociology.

Anne R. Kenney is the associate director of the department of preservation and conservation and codirector of the Cornell Institute for Digital Collections. Since 1989, she has been involved in a continuing series of research and production projects centering on the use of digital imaging for preservation reformatting and enhanced access. She has written and spoken widely on the topic of digital imaging and been involved in several intensive digital training programs, both at Cornell University and on behalf of RLG. She is the coauthor of the award-winning publication *Digital Imaging for Libraries and Archives* (1996) and is coeditor with Oya Y. Rieger of *RLG DigiNews*. She and Ms. Rieger have written a new monograph, *Moving Theory into Practice: Digital Imaging for Libraries and Archives*, which was recently published by RLG. Ms. Kenney is a fellow and past president of the Society of American Archivists.

Contents

Preface

Given the right hardware and software, digital information is easy to create, copy, and disseminate; however, it is very hard to preserve. At present, it is impossible to guarantee the longevity and legibility of digital information for even one human generation.

The Council on Library and Information Resources (CLIR) has sponsored work on possible solutions to this problem. One such solution, the development of emulators, would enable access to information created with software and hardware that has become obsolete. The merits of emulation are widely debated, and the approach has yet to be developed for broad, practical use. A more viable strategy, many argue, is migration, which the CPA / RLG Task Force on Archiving of Digital Information defines 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."

This report does not argue the merits of emulation or migration for longevity; rather, it addresses the practical aspects of migration in an operating library. Migration is, in essence, a translation. With migration, as with all translations, some information is lost, no matter how skilled the interpreter. In migration, it is usually the context, rather than the data, that drops out or is improperly reconstructed in the new code. This can be crippling in dynamic formats, in relational databases, and even in simple spreadsheets. Nonetheless, given how much information already exists in digital form and the brevity of its projected life span, institutions must act now to move information forward. They cannot afford to wait for the optimal solution.

In 1998, CLIR asked the Cornell University Library to undertake a risk assessment of migrating a handful of common file formats. This report is the fruit of their investigation. It is intended to be a practical guide to assessing the risks associated with the migration of various formats and to making sound preservation decisions on the basis of that assessment. The authors start from the premise that migration is prone to generating errors, and they provide practical tools to quantify the risks. They organize migration into a sequence of discrete steps and offer assessment tools to manage each of those steps. The process is presented in a workbook that can guide digital preservation specialists in their day-to-day operations. The authors also present two case studies—one for image files and another for numeric files—that demonstrate their approach.

The goal of any risk assessment is to identify, as unambiguously as possible, the risk of loss over time and the measures that can be taken to mitigate that loss. This is what the tools are designed to do. The difficulty, of course, is determining when risk is acceptable and when it is not. The authors underscore the importance of experience and judgment in practicing the art of preservation.

Abby Smith
Director of Programs

Introduction

The steady growth of digital information as a component of major research collections has significant implications for college and research libraries. Many institutions, including Cornell University Library (CUL), have been creating or collecting digital information produced in a wide variety of standard and proprietary formats, including ASCII, common image formats, word processing, spreadsheet, and database documents. Each of these formats continues to evolve, becoming more complex as revised software versions add new features or functionality. It is not uncommon for software enhancements to "orphan," or leave unreadable, files generated by earlier versions. The threat to aging digital information has surpassed the danger of unstable media or obsolete hardware. The most pressing problems confronting managers of digital collections are data format and software obsolescence.

There is a tacit assumption that digital libraries will preserve the electronic information they create or the information that is entrusted to their care. To preserve this information, institutions must manage collections in a consistent and decisive manner. It is important to decide what should be preserved, in what priority, and with what techniques. Unfortunately, there is little guidance in this area. Leading organizations such as the National Archives and Records Administration have been cautious in adopting standards for document formats other than ASCII; specialized reports prepared by national committees have focused either on broad recommendations (Task Force on Archiving of Digital Information 1996) or on organizational and legal issues (Euhlir 1997). On the basis of its experience in managing electronic collections, the CUL chose to develop a method of "risk management" to replace "heroic rescue" as a means of preserving digital information. The concept of an information life cycle is emerging as a major theme in digital preservation, and as a model it provides some guidance on where risk-management efforts should be directed. In the abstract, a digital life cycle plans for the creation and stages of use of information and, ultimately, for whether the file will remain in a terminal, unchanging state or be transformed into another format for reuse. The choice of how or when to assess risk in the digital life cycle depends on circumstances, the state of the digital information, and the general preservation strategy adopted.

Currently, there are two radically different strategies for managing the later period of a digital life cycle: migration and emulation. *Preserving Digital Information* defines migration broadly, 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 Archiving of Digital Information 1996). A more specific definition would indicate that migration changes the structure of the original data file. With the exception of files that are simple data streams, most files contain two basic components: structural elements and data elements. A file format represents the arrangement of the structural and data elements in a unique and specific manner. In this context, migration is the process of rearranging the original sequence of structural and data elements (the source format) to conform to another configuration (the target format).

In practice, migration is prone to generating obvious and subtle errors. An obvious error occurs when the set of structural elements in the source format does not fully match the structural elements of the target format. For instance, in a spreadsheet file a structural element defines a cell containing a numeric value. If a comparable element is missing from the format specifications of the target format, data will be lost. A subtle error might occur if the data themselves do not convert properly. Floating point numbers (numbers with fractions) are found in many numeric files. Some formats might allow a floating-point number of 16 digits (e.g., 26.00126700998l9070) while others might allow only 8 digits (e.g., 26.00126701). For some applications, such as vector calculations in geographic information system (GIS) programs, small but significant errors could creep into calculations. In other situations, migration might preserve the content of the file but lose the internal relationships or context of the information. For example, a spreadsheet file migrated to ASCII may save the current values of all the cells but lose any formulas embedded within the cells that are used to create those values.

An alternative preservation approach, emulation, is concerned with preserving the original software environment. Emulators are programs that mimic computer hardware. Strategies 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 data files (Rothenberg 1999; 1995). Emulation has been practiced for many years, and there are several commercial and public domain emulators for a variety of hardware / operating system configurations. A good example is MS-DOS emulation in the Windows 95 / 98 / NT operating system.

Emulation as a strategy has some limitations. Emulation assumes future access to the following multiple data objects in a cluster or package:

- 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. Moreover, emulation is a patchwork effort, with contributions from commercial vendors and private individuals. There is no system for coordinating or maintaining these emulators, and maintaining obsolete emulators may prove to be as problematic as migrating obsolete file formats.

With two complex and very different strategies, it would be difficult to examine both options simultaneously. Our decision to select migration was partially based on the resources at our disposal. With locally developed and commercial off-the-shelf data migration software, migration could be tested, measured, and evaluated on the basis of certain common criteria from which we could design a suite of risk-assessment tools. File migration was also appealing because it could encompass the following different preservation scenarios:

- the routine refreshing of digital files;
- varying changes in digital formats when files are converted from one application to another;
- radical changes in digital formats, such as the conversion of numeric files from proprietary formats to ASCII; and
- the migration of derivative access copy systems; for instance, system software might convert Tagged Image File Format (TIFF), a master storage format for scanned images, into a Portable Document Format (PDF) derivative designed for easy use by the reader.

For the reasons described above, Cornell concentrated exclusively on developing aids to assess the safety of a migration strategy for its digital information.

Literature Search

We reviewed the literature for information concerning digital preservation, digital migration, risk assessment, and file formats.

Digital Preservation and Migration

An extensive survey of the library literature identified many papers that provided in-depth analyses of issues associated with different aspects of digital preservation. The Task Force on Archiving of Digital Information (1996) documents these issues most effectively, and they will not be repeated here. Most of the remaining literature discussed digital reformatting or file copying from one medium to another. We identified four papers that directly related to our project. The first is the work of John Bennett (1997). His study evaluates preservation requirements by genre, format, media, and platform and uses a rudimentary risk-assessment scoring system. Displayed in a two-dimensional matrix, these requirements effectively communicate the complexity and interdependence of digital materials. Haynes et al. (1997) reported on an in-depth investigation into the responsibilities associated with maintaining digital archives. This paper summarizes numerous interviews with focus groups and individuals and effectively communicates the range of opinions and expectations associated with different stakeholders. The third work is the Reference Model for an Open Archival Information System (OAIS) (CCSDS 1999). The report is remarkable for its breadth and depth. In the authors' words, the model they describe "provides a framework for the understanding and increased awareness of archival concepts needed for long-term digital information preservation and access, and for describing and comparing architectures and operations of existing and future archives." The last item is a report written by Ann Green, JoAnn Dionne, and Martin Dennis (1999). Their study describes a project at Yale to convert data from column binary to spread ASCII format. The nine-step data migration process is well documented, and the findings and recommendations clarify important preservation issues.

Risk Assessment

Our search of the library literature for information concerning risk assessment was not fruitful. We then examined the literature for computer science. In the last 50 years, computer science has witnessed numerous cycles of software development migration, and the literature contains many studies, case reports, and models. Several publications were very useful in developing our understanding of risk assessment of digital information. *Rapid Development* (McConnell 1996) is a monograph on the general problems associated with software development. In many respects, software development exhibits several of the same problems associated with basic digital preservation. Chapter 5 of McConnell's book, which concerns risk management, provides an excellent theoretical and practical introduction to controlling risk in software development. It is a good primer for risk studies in digital preservation. Van Scoy (1992) examines a similar topic in a study funded by the U.S. Department of Defense. His study identifies risk-management participants and their activities. A later study (Sisti and Joseph 1994), also for the Department of Defense, expands on the work of Van Scoy and offers a highly detailed software risk evaluation method. All three studies pay particular attention to the organizational issues in risk management.

While researching risk assessments, we were struck by the vast differences in basic definitions used by different disciplines. (For example, see Reinert, Bartell, and Biddinger [1994], Warren-Hicks and Moore [1995], McNamee [1996], Wilson and Crouch [1987], Starr [1969], and Lagadec [1982]). Numerous professions measure risk, and each assigns risks a unique vocabulary and context. The degree and type of risk associated with any data archive may be understood differently by administrators, operational staff members, and data users, depending upon their individual training and experience. The measurement of risk was equally problematic. One paper correlated risk level with the nonlinear relative probability of risk occurring (Kansala 1997). Another publication introduced an algebraic formula (McConnell 1996). In a third instance, a research group felt that cases where one could accurately assess the probability of a future event were rare because the information technology environment for software changes so rapidly. They preferred simple estimates, such as *high*, *medium*, and *low*, which they believed facilitated decision making (Williams, Walker, and Dorofee 1997). Risk-measurement scales, like risk definitions, are as distinctive as their developers.

File Format

File format information was located from format specification files available on the Internet and from descriptions of file formats appearing in several monographs. Specifications for TIFF and .wk1 files were located at the following Internet sites:

- Adobe Corporation (http://www.adobe.com)
- Lotus Corporation (http://www.lotus.com)
- Unofficial TIFF Home Page (http://home.earthlink.net/~ritter/tiff/).
- Wotsit's Format: the Programmer's Resource (http://www.wotsit.org).

Murray and vanRyper (1996) describe TIFF with numerous illustrations and a detailed narrative about TIFF structure. Brown and Shepherd (1995) provide an effective description of the low-level data stream organization of the TIFF format. Lotus Development Corporation (1986) has prepared the definitive work for Lotus 1-2-3 .wk1 files. More than just a reference about file structure, the work explains why Lotus moved away from simple ASCII representation of spreadsheet data and documents its early attempts to use a general file format for worksheet, database, word processing, and graphics activities. The Lotus book is the best source for information about the .wk1 format. Related .wks format information, released into the public domain in 1984 and found at File Transfer Protocol (FTP) sites, or published by Walden (1986), should be used cautiously.

Risk Assessment as a Migration Analysis Method

In its present state, migration as a digital preservation strategy can be characterized as an uncertain process generating uncertain outcomes. One way to minimize the risk associated with such uncertainty is to develop a risk-management scheme that deconstructs the migration process into steps that can be described and quantified. A risk assessment is simply a means of structuring the process of analyzing risk. If the risk-assessment methodology is well specified, different individuals, supplied with the same information about a digital file, should estimate similar risk values.

We believe that three major categories of risk must be measured when considering migration as a digital strategy:

- *Risks associated with the general collection*. These risks include the presence or absence of institutional support, funding, system hardware and software, and the staff to manage the archive. These are essential components of a digital archive, which the Task Force on Archiving of Digital Information (1996) describes as "deep infrastructure." The collection, and the stakeholders who use the collection, will be affected to some degree by a migration of data. Legal and policy issues associated with digital information will introduce additional risks.
- *Risks associated with the data file format*. These include the internal structural elements of the file that are subject to modification.
- *Risks associated with a file format conversion process*. The conversion software may or may not produce the intended result; conversion errors may be gross or subtle.

Analysis of these three categories can be illuminating. Table 1 presents information from the image file case study that illustrates the risks specific to image files in migration. The findings are based on research, discussions with digital preservation specialists, and our own experience.

As each risk category was explored, we recognized that we needed to develop different methods, or tools, to sample each situation and to help quantify risk probability and impact. Over the course of the project, we developed three assessment tools:

- A risk-assessment workbook for the general collection. The workbook provides a general review of risks associated with migration at the collection level.
- A reader software to examine specific files, or collections of files, for high-risk format elements.
- A test file for a .wk1 format of known structural and data elements to test, or exercise, conversion software.

Individually, these three tools provide useful information. Together, they offer a means to gauge the readiness of any archive to migrate information successfully from one format to another.

Risk Assessment of General Collections

In an ideal situation, risk assessments would be performed by a team of experts; each member would be a specialist in a specific area and would have general knowledge of digital preservation. In reality, access to expert advice is costly and not always timely. In place of a human adviser, a workbook can provide a systematic approach to assessing risks and problems. If the questions or exercises are sufficiently developed, the workbook can help the user not only identify potential risks but also measure risk in terms of impact.

RISK CATEGORY	EXAMPLES
Content fixity (bit configuration, including bit stream, form, and structure)	Bits/bit streams are corrupted by software bugs or mishandling of storage media, mechanical failure of devices, etc.
	File format is accompanied by new compression that alters the bit configuration.
	File header information does not migrate or is partially or incorrectly migrated.
	Image quality (e.g., resolution, dynamic range, color spaces) is affected by alterations to the bit configuration.
	New file format specifications change byte order.
Security	Format migration affects watermark, digital stamp, or other cryptographic techniques for “fixity.”
Context and integrity (the relationship and interaction with other related files or other elements of the digital environment, including hardware/software dependencies)	Because of different hardware and software dependencies, reading and processing the new file format require a new configuration.
	Linkages to other files (e.g., metadata files, scripts, derivatives such as marked-up or text versions or on-the-fly conversion programs) are altered during migration.
	New file format reduces the file size (because of file format organization or new compression) and causes denser storage and potential directory-structuring problems if one tries to consolidate files to use extra storage space.
	Media become more dense, affecting labels and file structuring. (This might also be caused by file organization protocols of the new storage medium or operating system.)
References (the ability to locate images definitively and reliably over time among other digital objects)	File extensions change because of file format upgrade and its effect on URLs.
	Migration activity is not well documented, causing provenance information to be incomplete or inaccurate (a potential problem for future migration activities).
Cost	Long-term costs associated with migration are unpredictable because each migration cycle may involve different procedures, depending on the nature of the migration (routine migration vs. paradigm shift).
	The value of the collection may be insufficiently determined, making it impossible to set priorities for migration.
	Costs may be unscalable unless there is a standard architecture (e.g., centralized storage, metadata standards, file format/compression standards) that encompasses the image collections so that the same migration strategy can be easily implemented for other similar collections.
Staffing	Staff turnover and lack of continuity in migration decisions can hurt long-term planning, especially if insufficient preservation metadata is captured and the migration path is not well documented.
	Decisions must be made whether to hire full-time, permanent staff or use temporary workers for rescue operations.
	Staff may have insufficient technical expertise.
	The unpredictability of migration cycles makes it difficult to plan for staffing requirements (e.g., skills, time, funding).
Functionality	Features introduced by the new file format may affect derivative creation, such as printing.
	If the master copy is also used for access, changes may cause decreased or increased functionality and require interface modifications (e.g., static vs. multiresolution image, inability of the Web to support the new format).
	Unique features that are not supported in other file formats may be lost (e.g., the progressive display functionality when Graphics Interchange Format [GIF] files are migrated to another format).
	The artifactual value (original use context) may be lost because of changes introduced during migration; as a result, the “experience” may not be preserved.
Legal	Copyright regulations may limit the use of new derivatives that can be created from the new format (e.g., the institution is allowed to provide images only at a certain resolution so as not to compete with the original).

Table 1. Risks associated with file-format-based migration for image collections

When used as a common method of analysis, a workbook should identify and describe problems in a concise, uniform, and easily understood manner that could be shared by administrators and archivists in a given setting.

For the risk-assessment workbook developed in this study, we prepared two risk-assessment scales: one to measure the probability a hazard would occur, and another to measure the impact of that occurrence. These scales were prepared for a risk-assessment case study of a numeric file collection, the test bed for much of our project. Admittedly, the scales lack scientific precision, and at the end one does not simply sum the results and decide to migrate on the basis of a single number. On the other hand, assessment scales can more precisely convey meaningful assessments of risk, and this can help set priorities in preparing for a migration project (Beatty 1999).

The complete workbook is presented in Appendix A.

Risk Assessment of File Formats

As noted earlier, file migration is the process of altering structural and data elements in one file format to conform to a new configuration in another format. In our project, we label the original format the “source” format and the new format the “target” format. Software programs that convert source formats into target formats are grouped into three general categories:

- Translation programs for a specific project written by a company, by the owner of the information, or by a third-party vendor. Data archives often write these programs at considerable cost. The CUL experience with locally developed software is described in the TIFF image file case study.
- 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. Programs such as DataJunction are written specifically for this purpose.
- A general-purpose commercial translation program. Conversions Plus by DataVis is a good example of this growing genre of software.

Each of these approaches to conversion has its benefits and liabilities. Many conversion programs developed by archives can incorporate extensive knowledge about the functions of the translation software, but require lengthy development cycles and are expensive to prepare. Off-the-shelf commercial programs provide little information about the translation process but offer many features at a low cost.

A format risk assessment has to explore two distinct areas of risk: the risk introduced by the conversion program and the magnitude of recurring risk inherent in a large collection. In addition, the features and usability of the conversion software should be considered as well as the impact on the metadata associated with the files.

Assessing Risk in Conversion Software

Assessing risk inherent in conversion programs can be accomplished by examining a file before and after migration. A test file can be passed through the conversion software, migrating from source to target format. If, following the format conversion, the fields and field values of the original source file are properly reproduced in the target file, the risks incurred in migration are significantly reduced. On the other hand, if the fields or their values are not properly converted, the risks of migration are significantly increased. If the field tags and values in the test file are known, data changes associated with file conversion can be independently verified.

In the numeric file case study, a test file for the Lotus 1-2-3 .wk1 format was created. With the use of public domain specifications and reference manuals published with the original application software, a large file was generated that exercised all the field tags and field values. A simple conversion test might determine how well a conversion program tests the following known values with those generated in a formula:

STATISTICAL@FUNCTIONS			
	correct result	computed result	expression
	495	495	@AVG(H293..DC293)
	100	100	@COUNT(H294..DC294)
	74	74	@COUNT(H295..DC295)
	994	994	@MAX(H296..DC296)
	28	28	@MIN(H297..DC297)
	297	297	@STD(H298..DC298)
	299	299	@STDS(H299..DC299)
	49,450	49,450	@SUM(H300..DC300)

Fig. 1. Sample test values for assessing conversion accuracy (Lotus 1-2-3 file)

In the example shown above, the “average” function (@AVG) operates on a range of cells (H293..DC293). The precomputed correct result (495) is compared with the computed result derived from the expression, and any differences between the two are recorded. In a similar manner, other complex formulas and functions can be compared before and after conversion.

It took us about three hours to compare our test files manually before and after conversion. Although this method is somewhat laborious, it is quite accurate for the formats we tested. Conversion of different structural elements and data elements is not always a matter of “hit or miss.” We were able to identify conversions that were almost, but not quite perfect. Testing these problematic conversions, we were able to develop a rough scale of conversion risk (1=minor risk, 5=high risk). Documentation for the test file can be found in Appendix B.

Assessing Recurring Risk Inherent in a Large Heterogeneous File Collection

Manual identification of risk associated with file structures is possible for a small number of files. For large digital collections that have thousands or millions of files that may contain one or more of these at-risk elements, manual methods are expensive and inefficient. One way to measure the collection for files that contain at-risk elements would be to prepare a file reader programmed to examine each file for these items. If one or more risk items are found, the program could be written to produce a report that identifies the file, its location in the collection, and the type and number of at-risk elements associated with that file. Good design would make the program flexible enough to read most, if not all, files with defined structural elements.

A program was developed for the project that can read structured ASCII and binary files. Named Examiner, the program reads a file and detects the presence and frequency of specific file format elements. It does not read or evaluate the data value, although this feature could be implemented. The following example shows a few lines from a report generated during a scan of .wk1 files in the USDA Economics and Statistics System, hosted at Mann Library.

```
/usda/ftp/usda/data-sets/crops/94018/budget.wk1:
   Risk Level 5
   Tag 14: NUMBER: Floating point number—Qty: 584
-----
/usda/ftp/usda/data-sets/crops/94018/charactr.wk1:
   Risk Level 5
   There are no tags in this file at this level
-----
/usda/ftp/usda/data-sets/crops/94018/conf_int.wk1:
   Risk Level 5
   Tag 14: NUMBER: Floating point number—Qty: 59
```

In the output just listed, Examiner has examined a series of .wk1 files in a single subdirectory with the absolute path /usda.ftp/usda/data-sets/crops/94018. In two of the three files, it located a structural element, or Tag. The program writes to a report file the structural element number (14), the name of the structural element given in the format specifications (NUMBER:), a short description of the structural element (Floating-point number), and the total count of floating-point numbers discovered in that specific file (Qty:). The program also describes the risk level for the structural element. The risk level was determined during the initial source–target analysis described previously. The program can be set to report at-risk tags only if the risk value equals or exceeds a certain threshold.

One strong feature of the Examiner program is that it is nondestructive. It simply reads a file from beginning to end and declares what is found. Also, Examiner can be set to read a single file, all the files in a directory, or all the files on a drive. The program is reasonably efficient and scans approximately 10,000 .wk1 files per hour. Finally, Examiner is written in Java, a modern programming language designed to be easily compiled on different operating systems. The program has been fully tested in the Unix and Windows 95 / NT environments. General documentation for Examiner is described in Appendix C. The source code and full documentation are available on the Web site of the Council on Library and Information Resources.

Assessing Risk Associated with the File Conversion Process

Finally, there are risks associated with the features of different conversion software. The project examined two commercial off-the-shelf programs and quickly scanned the advertisements or published reviews of six others. In any mix of conversion programs available, each will provide some or all “core” functions as well as optional features. General performance benchmarks, which can be tailored for specific migration scenarios, provide some uniformity of measurement and highlight obvious defects. For example, we examined DataJunction as a general-purpose conversion program for spreadsheet and database formats. Conversion of .wk1 formats was trouble-free, except for one major flaw: DataJunction was difficult to program to work in batch mode. We did not recognize this flaw until the evaluation was nearly complete. Obviously, a project timetable could be seriously jeopardized by such a limitation. Although not an intended product of the project, we recorded software assessment questions that we should have asked at the start of the project. From these, we developed a short functionality assessment that is now available on the Web site of the Council on Library and Information Resources.

Identification of Metadata-Related Risk

We frequently think of disk files as the sole object of migration because, at first glance, the information they contain is what we have to move from one format to another. The individual files in a collection, however, are frequently useless without other information describing how the files are to be used or how they relate to one another. In other words, any group of files that constitute a cohesive unit can be considered a digital object, and what makes the digital object intelligible is metadata describing the contents and providing structure for the group. When such digital objects exist, the metadata, as well as the individual files containing the raw data, must be successfully migrated.

Metadata at the digital-object level can take various forms. For example, in the collection of TIFF images in one of our case studies, a file in a proprietary format, Raster Document Object (RDO), contains metadata that provides structure to the multiple TIFF files. The RDO file relates the page image stored in each TIFF file to the others that compose the document; in this case, the navigable and searchable digital object represents a paper document containing pages and chapters and other logical constructs. A second example, from our case study of a collection of numeric files in the .wk1 format, shows another way of structuring and describing digital objects. Each digital object—a set of related binary data files—has three metadata components: one that contains information about the structure of the object, one that describes the content of the object, and one that creates a link between the two. The structural metadata is contained in an HTML file whose links point to the individual files that constitute the digital object. The content metadata is in an English-language ASCII file. Its purpose is to provide searchable text so that the object can be located in a search across the larger collection of objects. The third component is a record in a database that creates a relationship between the content file and the structural file. In a successful migration to another data format, the structural metadata in the HTML file would have to be changed if the name or location of the individual files in the digital object were changed. The content description and the database record would not have to be touched.

Case Studies

The risk-assessment tools developed were tested on two digital collections at the Cornell University Library: the Ezra Cornell Papers and the USDA Economics and Statistics System. Each collection contains a dominant file format: TIFF or .wk1. The assessments of these two collections are presented in Appendixes D and E.

Findings and Recommendations

Migration Risk Can Be Quantified

Migration, or the conversion of data from one format to another, has measurable risk. The amount of risk will vary, sometimes significantly, given the context of the migration project. One form of risk depends on the nature of the source and target formats. We have shown that it is possible to compare formats in a number of ways and to identify the level of risk for different format attributes. The format analysis techniques and software may be technical, but the results can be described in general terms. Since basic file structure concepts are common to many file formats, experience with one format can be used to understand other formats.

We draw a similar conclusion concerning organizational, hardware, software, and metadata risks. Information delivery systems must sustain a certain level of organization simply to function. Consistent components of these systems can be evaluated; for example, personnel, funding, metadata, and rough but quantifiable measures of risk can be established for these subjects.

The greatest challenge is the interpretation of the risk, i.e., to determine when a risk is acceptable. Risk-assessment tools cannot replace experience and good judgment. The tools can be compared with navigation aids used on the high seas. Following five centuries of intensive effort to develop risk-reducing technologies, ships' helms are still manned, and collisions between ships at sea still occur.

In this study, we provide examples to illustrate the evaluation process. In practice, the risk-assessment tools are not fully developed. We recommend the further refinement of these tools to provide results that are more reliable. We must recognize, however, that this will take some time, during which we will lose some data.

Conversion Software

This study is unable to recommend a cost-effective, off-the-shelf commercial software program to implement a migration strategy. From our analysis, we believe that migration software should perform the following functions:

- Read the source file and analyze the differences between it and the target format.
- Identify and report the degree of risk if a mismatch occurs.
- Accurately convert the source file(s) to target specifications.
- Work on single files and large collections.
- Provide a record of its conversions for inclusion in the migration project documentation.

Neither of the two programs analyzed in this case study met all these criteria, although our results suggest that commercial conversion programs, with further development, have the potential to meet them. Considering the cost of writing conversion software for a wide range of file formats, we believe a commercially developed solution for migration software will ultimately be cheaper and more flexible than locally developed conversion software. We recommend further work with vendors, such as DataJunction and DataViz, to educate them about our needs and help them develop products that promote safer file migration.

Access to Format Data

The most difficult aspect of this project was the acquisition of complete and reliable file format specifications. Throughout the project, format-specific information was difficult to acquire from a single source. Ultimately, format information for this study was acquired from the following four general sources:

- software developers
- public FTP archives
- monographs
- Internet discussion lists

Developers of software applications who use a specific proprietary file format should be the best source for file format information. At the start of our search for Lotus .wk1 format information, this was not the case. Lotus, like other large software companies, treats file format information as a business product to sell to software developers. Lotus business products evolved, responding to revisions in 1-2-3 as well as to changes in the DOS / Windows operating system. With the introduction of Windows 3.1, developer interest in earlier DOS specifications disappeared. Since the specifications for the .wk1 format were integrated into the format specifications for later releases (i.e., .wk3, .wk4), the specifications and documentation for the earlier .wk1 format quietly disappeared. Lotus as a company also evolved, and key members of the early development staff—often the corporate memory in software companies—moved on to establish their own companies. In the last months of this project, we were able to contact an individual at Lotus who had been with the company since the mid-1980s. This individual helped us acquire a copy of *Lotus File Formats for 1-2-3, Symphony, and Jazz*. This work, authored by Lotus, is the only surviving documentation from the company for that period. Fortunately, it describes the .wk1 format in complete detail.

Throughout the year, Lotus staff repeatedly referred us to their FTP archive that contains 1-2-3 .wk1 format specifications. These specifications were indirectly certified by Walden (1986), who describes the specification in detail and provides a sample .wk1 file analyzed byte by byte. Unfortunately, these specifications are incomplete and describe the .wks file format, the format of 1-2-3 release 1A. We were surprised that Walden made such an oversight, but Wotsit's Format Web site (Oliver 1999) and the comp.apps.spreadsheets FAQ (1999) repeat the error. It is clear that neither the professionals nor the amateurs recognized the mistake.

TIFF specifications are accessible from two Internet locations. The official specifications for TIFF 6.0 are available from the Adobe developers' support site. Adobe's site does not list the specifications for TIFF 4.0 and 5.0. These can be located at the Unofficial TIFF Home Page. Our manual examination of the specifications showed them to be consistent with each other, but they are incomplete. For years, developers have been adding their own proprietary tags to the TIFF specification that they register with Adobe. Special tags do not appear in either the official or unofficial specifications. Several books have been written about the TIFF file format specifications and they survey many file formats. However, no single work presents a clear, comprehensive description of the TIFF file format specification or of information about proprietary tags.

We expect these difficulties to be repeated when other formats are explored. Conceptually, the solution is to adopt "open" format specifications, where complete, authoritative specifications are available for anyone to access and analyze. Our experience with TIFF and .wk1 suggests that with file formats, there are two specifications at work. One is the public document, which describes the basic or core elements of any format. The other is a private, nonstandard set of file elements, usually developed to extend the functionality of a file format. These private file elements provide the competitive edge for third-party software and rarely are openly circulated. Over time, new format elements are often integrated into format revisions. For example, TIFF grew from 37 tags in version 4 to 74 tags in version 6.0. New proprietary tags for TIFF version 6.0 are registered with Adobe, which does not make them public. It is uncertain whether all or some of these difficult-to-identify tags will be integrated into the anticipated TIFF version 7.0. We endorse the concept of open specifications and recommend that more thought be directed at coordinating access to both the relatively static, public domain specifications and the dynamic, nonstandard elements.

Public Access Archives of Format Information

If we measured the risk associated with public domain archives on the Internet, we would assess all these sites as high-risk operations. Sites such as Wotsit's represent the public service efforts of individuals. They lack any vision or plan to sustain the information. This limitation, combined with the unreliable nature of the information contained within these sites, make it unlikely that these sites will contribute meaningfully to digital preservation efforts. There is a pressing need to establish reliable, sustained repositories of file format specifications, documentation, and related software. We recommend the establishment of such depositories for format-specific materials related to migration as a preservation strategy. It is a concern, as well, for emulation programs and their documentation.

References

Beatty, J. Kelly. 1999. The Torino Scale: Gauging the Impact Threat. *Sky & Telescope* 98(4):32-3.

Bennett, John C. 1997. *A Framework of Data Types and Formats, and Issues Affecting the Long Term Preservation of Digital Material.* British Library Research and Innovation Report, No. 50. West Yorkshire, U.K.: British Library Research and Innovation Centre. Available from http://www.ukoln.ac.uk/services/elib/papers/supporting/#blric.

Brown, C. Wayne, and Barry J. Shepherd. 1995. *Graphic File Formats.* Greenwich, Conn.: Manning Press.

comp.apps.spreadsheets. 1999. comp.apps.spreadsheets FAQ. Available from http://www.faqs.org/faqs//spreadsheets/faq.

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.

Euhlir, Paul. 1997. Framework for the Preservation of and Public Access to USDA Digital Publications. Available from http://preserve.nal.usda.gov:8300/npp/frameprt.html.

Green, Ann, JoAnn Dionne, and Martin Dennis. 1999. *Preserving the Whole: A Two-Track Approach to Rescuing Social Science Data and Metadata*. Washington, D.C.: Digital Library Federation. Available from http://www.clir.org/pubs/reports/pub83/contents.html.

Haynes, David, et al. 1997. *Responsibility for Digital Archiving and Long Term Access to Digital Data*. JISC/NPO Studies on the Preservation of Electronic Materials. British Library Research and Innovation Report, No. 67. West Yorkshire, U.K.: British Library Research and Innovation Centre. Available from http://www.ukoln.ac.uk/services/elib/papers/supporting/#blric.

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

Lagadec, Patrick. 1982. *Major Technological Risk: An Assessment of Industrial Disaster*. Oxford, U.K.: Pergamon Press.

Lotus Development Corporation. 1986. *Lotus File Formats for 1-2-3, Symphony and Jazz: File Structure Descriptions for Developers*. Cambridge, Mass.: Lotus Books, and Reading, Mass.: Addison-Wesley Publishing Company, Inc.

McConnell, Steve. 1996. *Rapid Development: Taming Wild Software Schedules*. Redmond, Wash.: Microsoft Press.

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

Murray, James D., and William vanRyper. 1996. *Encyclopedia of Graphics File Formats*, second edition. Cambridge, Mass.: O'Reilly & Associates, Inc.

Oliver, Paul. 1999. Wotsit's Format: the Programmer's Resource. Available from http://www.wotsit.org/.

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.

Rothenberg, Jeff. 1995. Ensuring the Longevity of Digital Documents. *Scientific American* 272(1):42-7.

Sisti, Frank J. and Sujoe Joseph. 1994. Software Risk Evaluation Method. Version 1.0. Technical Report CMU/SEI-94-TR-19. ECS-TR-94-019. Pittsburgh, Penn.: Software Engineering Institute, Carnegie Mellon University.

Starr, Chauncey. 1969. Social Benefits versus Technological Risk: What is Our Society Willing to Pay for Safety? *Science* 165:1232-8.

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/.

Van Scoy, Roger L. 1992. Software Development Risk: Opportunity, Not Problem. Technical Report CMU/SEI-92-TR-30/ESC-TR-93-030. Pittsburgh, Penn.: Software Engineering Institute, Carnegie Mellon University. Available from http://www.sei.cmu.edu/publications/documents/92.reports/92.tr.030.html.

Warren-Hicks, William J., and Dwayne R. J. Moore. 1995. Uncertainty Analysis in Ecological Risk Assessment. Pensacola, Fla.: SETAC Press.

Walden, Jeff. 1986. *File Formats for Popular PC Software: A Programmer's Reference.* New York, N.Y.: John Wiley and Sons, Inc.

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

Wilson, Richard, and E. A. C. Crouch. 1987. Risk Assessment and Comparisons: An Introduction. *Science* 236:267-70.

Web sites noted in report:

Adobe developers' support site: http://partners.adobe.com/asn/developer/technotes.html.

Council on Library and Information Resources: www.clir.org.

The Unofficial TIFF Home Page: http://home.earthlink.net/~ritter/tiff/.

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	VisiCalc
	Developer has stopped supporting the software	Borland Dbase (originally Ashton-Tate)
	Market share is declining	WordPerfect
	Supporting programs have changed significantly	Compression software changes for TIFF
	Third-party support is lacking	Common Ground
	Paradigm has shifted	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

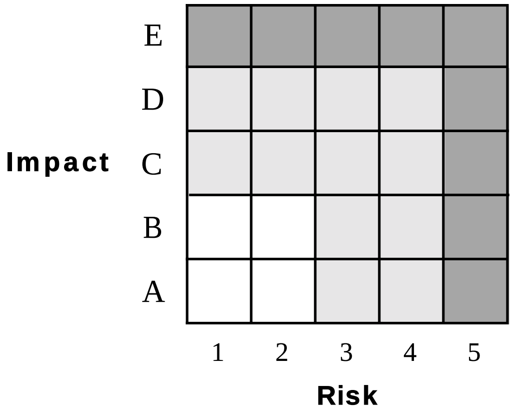
Label	Value	Description
Catastrophic	**E**	Complete, irreversible loss of data. Data cannot be drawn from other sources—print, artifact, or digital.
Very Serious	**D**	Partial, irreversible loss of data. Data cannot be drawn from other sources.
Serious	**C**	Complete loss of data. Data can be fully reconstructed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

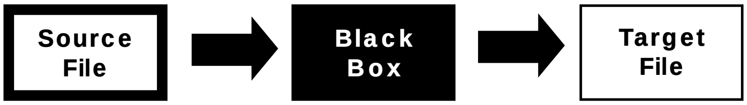
1. If all assessment question responses fall within the white grid cells (1A-B, 2 A-B), the migration process is likely to pose low risk. With due caution, the migration can be carried out.
2. If assessment question responses fall within the gray shaded grid cells (1C-D, 2C-D, 3A-D, 4A-D), the migration 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

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):

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?

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.

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

❒ Archive staff
❒ Organizational staff
❒ Trusted data users

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.

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 structure, 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/07 lagoze.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.

Appendix B

Documentation for Format Migration Test File Lotus 1-2-3, Release 2.2

1. Overview

A test file was created in Lotus 1-2-3, Release 2.2, as a tool for determining which file characteristics are maintained when spreadsheet data files are converted (migrated) from one file format to another. The test file can be used to assess the following:

- whether numeric values migrate properly, without loss of precision,
- whether characters and text cells migrate properly,
- whether worksheet characteristics such as column width are preserved after migration,
- whether cell formats (e.g., text, numeric, date, time) are maintained, and
- whether cell functions (@functions) can be successfully transferred from one file format to another.

The test file allows the assessment of all but one of the 92 cell functions (@functions) available in Lotus 1-2-3, Release 2.2. (The @CLEAN function is excluded because it relies on input from software other than Lotus 1-2-3.)

Before undertaking these test procedures, you should have at least a fundamental understanding of spreadsheet functions and an ability to check basic file characteristics such as column width without specific instructions.

2. Organization of the Test File

The test file, *Testfile.wk1*, is a single worksheet with 23 parts or test procedures, each of which corresponds to a particular set of file features or functions. The 23 parts are as follows:

layout
column width

maximum number of rows and columns
justification
maximum length of text in a single cell
defined cell ranges
minimum and maximum expressible values
preservation of significant digits
special formats (e.g., date, currency)
arithmetic operations
statistical @functions
financial @functions
calendar @functions
logical @functions
special @functions (excluding @cell [“filename”] and @cellpointer)
the @cellpointer function (excluding @cellpointer [“filename”])
the @cell (“filename”) and @cellpointer (“filename”) functions
text @functions
general math @functions (excluding @rand)
the @rand function
trigonometric @functions
database @functions
character set

Be careful not to insert or delete any columns or rows in the spreadsheet because several of the spreadsheet functions rely upon absolute cell references, i.e., the existence of particular values in particular cells.

3. Evaluation Procedures

The *source format* is the file format in which the original file (the test file) was created: .wk1 format (Lotus 1-2-3, Release 2.2). The source file, in this case, is the original test file.

The *target format* can be any spreadsheet format other than the source format. The target file is the file created by converting the test file (.wk1 format) into some other spreadsheet format.

To test the accuracy of migration from the source format to the target format, first convert the test file into the target format. This can be done with any spreadsheet program that accepts .wk1 files as input or with a stand-alone file conversion program such as DataViz. If you are using a spreadsheet program to convert the files, you will need to open (import) *Testfile.wk1* within the spreadsheet program and save it as a new file in the target format. If you are using a stand-alone conversion program, follow the instructions provided with the software.

All but one of the test procedures in *Testfile.wk1* can be completed without reference to the source file. That is, the target file alone, loaded into the target spreadsheet program, can be used to evaluate the effectiveness of the file migration process. (Unlike the other procedures, "character set" requires a visual inspection of source and target files.)

Layout, Formats, and Related Procedures

The first nine test procedures, from "layout" to "special formats," are used to determine whether particular worksheet and cell formats are preserved in the target file. In most cases, you can simply check whether the characteristics of the target file match the criteria stated in section 4 of this document, (e.g., whether column N is actually 34 characters wide, as stated, or whether the column widths established in the test file were lost during file format migration).

Arithmetic and @Function Procedures

The 13 procedures from "arithmetic operations" to "database @functions" evaluate particular arithmetic operations and cell functions. For each operation or function, *Testfile.wk1* presents

1. an expression: a string of text showing the syntax of the function statement as it appears in (2) of this list;
2. a computed result: an arithmetic or @function statement that corresponds to (1) and produces a particular numeric result;
3. a correct result: the expected result of the function; and
4. a set of numeric values, usually found in column H and subsequent columns, that are used as input for the function statement.

If the function works properly after file format migration, then the computed result (2) will match the correct result (3).

The transferability of any @function can, therefore, be evaluated through a simple comparison of the correct and computed results for that function. If the correct and computed values match, then the function is transferable from the source format to the target format. If the correct and computed results do *not* match, then the function did not migrate properly.

Additional evaluative procedures are required in just a few cases. These are noted in section 4 of this appendix.

Note: The testing procedure used in *Testfile.wk1* relies on the assumption that the numeric values used as input (4) and the characters used in the display of the correct result (3) both transfer without error from one file format to another. This assumption is reasonable, since only the standard alphanumeric characters have been used in cells that contain input values or correct results.

Procedure for Comparing Character Sets

The last procedure, "character set," requires a visual comparison of the source and target files. See section 4 for details.

4. Notes on Particular Evaluation Procedures

Layout

The header and margin settings shown in the Print—Printer—Options menu should match those described here. (See *Testfile.wk1* for details.)

Column Width

Ideally, the column widths will match those shown here. (See *Testfile.wk1* for details.)

It is possible that the target program will not use the same units to indicate column width as does the source format. (Lotus 1-2-3, Release 2.2, states widths in characters; other programs may not.) In that case, the *relative* width of each column should be the same as in the source file. That is, columns A and O should be 2/16 as wide as a standard column. Columns B and C should be 22/16 times as wide as a standard column, and Columns N and P should be 34/16 times as wide as a standard column.

Maximum Number of Rows and Columns

There should be at least 8,192 rows and 256 columns in the target file. (See cells A8192 and IV1.)

Justification

The word "left" should be left justified within the cell. The word "center" should be center justified. The word "right" should be right justified.

Maximum Length of Text in a Single Cell

The maximum length of text allowed in Lotus 1-2-3, Release 2.2, is a specified number of characters (240), not a specified width. The target file format should meet or exceed this length limit.

Specifically, these two cells should each start with "We the People" and end with "Liberty to ourselves and ou". (The last letter in "our" exceeds the length limit.)

Defined Cell Ranges

To test whether defined cell ranges are preserved in the target file, change one or more of the values in cells B44, C44, and D44. The sum, average, and count values should immediately change to reflect these modifications.

Minimum and Maximum Expressible Values

Expressible values (numeric values valid for display and computation) are shown in the source file as numbers, either in standard format or in scientific notation. Inexpressible values are denoted by a series of asterisks.

The maximum expressible value (column B) should be no less than 1.0E+99. That is, expressible values should appear in column B from row 59 through (at least) row 158.

The minimum expressible value (column E) should be no more than 1.0E-99. That is, expressible values should appear in column E from row 59 through (at least) row 158.

In the source file, all rows up to and including the marked rows (1.0E+99 and 1.0E-99) contain expressible values. Subsequent rows contain inexpressible values. If any rows below the marked rows contain expressible values in the target file, this indicates that the target file format allows the use of values more extreme than those allowed by the source file format.

Preservation of Significant Digits

The degree of precision available in the source file should be maintained in the target file. Specifically, the numeric values shown in cells N167 through N182 should match the values shown in cells P167 through P182. (These values are presented in columns N and P because columns A through M are not wide enough to display them in their entirety.)

Special Formats (e.g., Date, Currency)

The value in each formatted cell (the “formatted value” column) is 2846.3912. The formatted cells differ only in format, not in content.

If the target file supports the same formats as the source file, then each formatted value (column C) should look the same as the corresponding text cell (column B) does.

Arithmetic Operations

Each computed result (column C) should match the correct result shown in column B.

Statistical @Functions

Each computed result (column C) should match the correct result shown in column B.

Financial @Functions

Each computed result (column C) should match the correct result shown in column B.

Calendar @Functions

For all but the @NOW function, each computed result (column C) should match the correct result shown in column B.

The @NOW function shows the last date and time that any cell was entered or recalculated, not necessarily the *current* date and time.

If the target software recalculates all values continuously, then the @NOW function will display the current date and time. In most cases, however, it will be necessary to enter or recalculate a value (any value in the spreadsheet) in order to display the current date and time.

To recalculate all the values in the test file and thereby display the current date and time, press F9 in Lotus 1-2-3. (Other software programs may use a different key or combination of keys.)

Logical @Functions

Each computed result (column C) should match the correct result shown in column B.

Special @Functions: Excluding @CELL ("filename") and @CELLPOINTER

Each computed result (column C) should match the correct result shown in column B.

The @CELLPOINTER Function: Excluding @CELLPOINTER ("filename")

The @CELLPOINTER function shows the characteristics of the currently active cell (i.e., the cell with the cursor). If this function is working correctly, then the computed result in column C should match the correct result in column B when

1. the cursor is placed in column F of the same row as the function statement, and
2. the worksheet is recalculated. (In Lotus 1-2-3, press F9 to recalculate the worksheet. Other software programs may use a different key or combination of keys.)

To get the proper computed result in cell C437, for example, place the cursor on cell F437 and press F9.

After checking that the computed result in C437 matches the correct result in B437, follow this same procedure for each of the other rows in this section. That is, put the cursor on cell F438, press F9, and check that the computed result in C438 matches the correct result in B438. Then proceed to cell F439, press F9, and so on.

The @CELL("filename") AND @CELLPOINTER("filename") Functions

Cells B481 and B482 should each show the complete name of the target file, with the path from the hard drive to the file.

This test procedure does not require a comparison of correct and computed results.

Text @Functions

Each computed result (column C) should match the correct result shown in column B.

General Math @Functions: Excluding @RAND

Each computed result (column C) should match the correct result shown in column B.

The @RAND function

To test this function, recalculate the values repeatedly. (Press F9 to recalculate in Lotus 1-2-3. Other software programs may use a different key or combination of keys.)

The computed values should change each time but should always be approximately equal to 0.50 (the mean) and 0.29 (the standard deviation).

Individual deviations from these expected values are not a cause for concern as long as the values *usually* approximate 0.50 and 0.29 after each recalculation.

Trigonometric @Functions

Each computed result (column C) should match the correct result shown in column B.

Database @Functions

Each computed result (column C) should match the correct result shown in column B.

This section can be used to evaluate those database @functions available through the spreadsheet itself. Functions that rely upon input and output forms are not included in the test file, since those functions are generally used as tools for the construction of data files rather than as carriers of data.

Character Set

This section requires a visual comparison of the source and target files to ensure that each character in the source file is accurately represented in the target file.

The character set will show up properly in the target file only if
1. the target file uses the same character set as the source file, and
2. the @CHAR function works in the target file format.

The characters are listed here by LICS (Lotus International Character Set) number. Each character appears above the corresponding LICS code. It is important to realize that some target formats may support all the characters shown here without relying on the same character numbering scheme. In that case, the character set may migrate successfully even though the original and target files do not match precisely.

Appendix C

Documentation: Examiner and RiskEditor

What Is This Software for?

File migration and “black-box” converters

A major risk in migrating collections of files is the conversion software used to translate the files from the original format to our chosen target format. We start with a file whose content we hope to translate without corruption. We send it through a “black box” and hope that the integrity of the content will be preserved. We can presume success if we know that the conversion software faithfully maps every property of the source format to corresponding features in the target format (assuming, of course, that the target format has a feature set that is rich enough to store the properties and data of the source). For example, if the document format we are converting has a way to indicate bold text, and the target format can also indicate bold text, we want to know that the conversion software correctly maps bold to bold. More important, in most cases, data values, whether numeric, image, or text, should also move from one format to the other intact.

Two ways of evaluating the black box

If we can examine the mapping process and the data-moving techniques of the conversion software, we can evaluate the correctness of both functions. This examination must be repeated for every combination of source and target formats with which we are working, because each combination has a unique mapping. Moreover, to attempt this method, we must have access to the source code of the converter and possess the expertise to evaluate the code. Our experience in obtaining source codes from commercial software vendors has not been fruitful. Even if it were, the resources necessary for evaluating a specific mapping for every combination of source and target formats make this an impractical method for creating a general and expandable technique for assessing the risk involved in migrating collections.

Another method is to compare a converted file with the original file. If the result meets our standard of success, whatever that standard may be, we can say that the conversion software has performed adequately. However, we can make that statement solely about the particular source file we converted. The ideal file for the test would be one that tested all the features of the source format and tested data values at the minimum and maximum of every range possible. If that file were run through the converter, the resulting file could be compared at every point with the original.

Our approach

For our own collection of Lotus 1-2-3 files, we created a test file in the .wk1 format. With it, we can evaluate potential conversion software by running the software on the test file and then comparing the converted file with the test file. Visual inspection and comparison of all the properties and values are necessary to identify differences; this took about two hours. Proprietary software codes and knowledge of an uncertain number of format-to-format mappings are not needed for the visual inspection method. Another benefit of the test file is that it provides a baseline against which to evaluate and compare multiple conversion applications.

Regardless of the method used to evaluate the conversion software, if any of the properties or data values are not the same in the source and target files, then we know that the conversion software has introduced one or more points of risk. Thinking about the whole collection of files to be migrated, we will want to know whether some of the files in the collection have any at-risk properties. We can then decide whether to find another converter, to refrain from migrating those files, or to consider some or all of the loss acceptable.

We wrote the Examiner software application to test a collection of files for the presence of particular properties. Using the RiskEditor application, we indicate the properties that are at risk. If desired, we can order them by the degree of importance or impact. Then we run the Examiner application on a part or all of the collection. Examiner produces a report that lists which files contain the properties in question. With this knowledge, we can make an informed decision about the technical risks introduced by the conversion software.

The Examiner application is written in Java, and both its user documentation and technical documentation are available as HTML files. Examiner is designed to be extendable to any file format that indicates properties as numbered tags, including Lotus 1-2-3 and TIFF, the formats of our case-study collections. A requirement for running the application is a Java interpreter on the computer holding the collection. We wrote a command-line version of the program to be used on our Unix servers, but the program could be easily extended to have a graphical user interface.

Installation

Requirements

1. A JDK 1.1-compliant Java virtual machine installed on the same computer where your wk* files are stored.
2. Adequate Unix or Windows privileges to create a directory and bestow write permission to files within it.

Installing

Unix

1. Unzip and untar “`examiner.tar.gz`”.
2. Give the user permission to run the files in “`examiner/bin`”.
3. In the same directory, give the user permission to write to the files “defaultProperties”, “appProperties”, and any files ending with “.rsk”.
4. Add the /examiner/bin directory’s path to the CLASSPATH environment variable in the user’s profile, or edit the “examiner” and “riskEdit” scripts to point to the appropriate path. Comments in the scripts explain what must be done. You may want to put them in a directory in the user’s executable PATH.

Windows

1. Unzip “`examiner.zip`”.
2. The users should have permission to write to files by default. If that is not the case with a particular user, give the user permission to write to the files “defaultProperties”, “appProperties”, and any files ending with “.rsk”.
3. Add the “\examiner\bin” directory’s path to the CLASSPATH environment variable in the user’s profile, or edit “examiner.bat” and “riskEdit.bat” to point to the appropriate path. Comments in the batch files explain what must be done. Users may want to put them in a directory in their executable PATH.

Running Examiner and RiskEditor

If the environment variables CLASSPATH and PATH are set to include both the java files and the Examiner file, change to the directory with the Examiner class files, and type “`java Examiner`” or “`java RiskEditor`” on the command line. Then answer the prompts.

If the Unix scripts, examiner and riskEditor, or the DOS batch files examiner.bat and riskEditor.bat have been edited to include local directory information, type “`examiner`” or “`riskEditor`” on the command line. Then answer the prompts.

Using RiskEditor

For the Examiner program to selectively identify risk or impact associated with individual tags, the user must first assign a value to the risk / impact of the presence of a particular tag in the files. RiskEditor enables users to mark tags with a value between 1 (low) and 5 (high). After having converted a test file into another format and having compared the data and functions of the two files, the user knows what attributes have not been converted successfully. Some failures may be more important than others. By comparing the features to a list of the tags in the source format, users can identify the tags they want to look for in their collection.

Here is an example of a RiskEditor session, with comments.

```
What file type would you like to edit? [wk1, wks]
```

[Users are given a choice from among the file types for which there are .rsk files in the program’s working directory.]

```
Do you want to “browse” (move through the tags in sequence) or “specify” (edit specific tags)?
```

[The “browse” mode moves through the tags sequentially, while the “specify” mode simply asks for the number of a tag to be changed.]

```
Enter decimal number of tag to be changed, or “quit”: 14
```

```
Tag number: 14 Value: 5
```

[This is the “specify mode”. “Browse” mode shows only the tag number/ value line.]

```
Enter a new value (1-5), “ClearAll” to reset every
value to 1,“save” to save your changes, or “quit”
```

[“1” represents the lowest priority, “5” the highest—”save” always writes over the appropriate .rsk file—”quit” prompts you to save if you have changed some information.]

There is no need to assign values to all the tags. In practice, we have not had to mark more than three tags at one time. One of the tags was more important than the others because it represented a feature we felt we could not allow to be corrupted during migration; we marked it as a 5. The other two features represented risks we could accept; we gave each a risk / impact level of 4.

Using Examiner

If you have not assigned risk / impact levels to the tags you are interested in using RiskEditor, see the instructions for that application first. Then run the Examiner program on the collection of files you want to examine.

A sample session

Here is an example of a session, with comments. User input is in **bold.**

```
$ examiner  [this session was run from a Unix shell script]
Tue, Oct 12, 1999 02:45:39 PM   [start time]
Examiner....

Please enter the file type to be examined [wk1, wks]: wk1
```

[File types for which there are tag descriptions are in brackets—for example, .wk1, .wks.]

```
What is the starting directory? [/usr/local/Examiner] /usda/ftp/usda/data-sets
```

[The default is the directory from where the program is running.]

```
What is the minimum risk/impact level to be displayed? 5
```

[We are interested only in the highest level in this session.]

```
In which file should the report be stored?: /usda/testdir/wk1.run.5.report
```

[The user must have permission to write a file.]

```
Working... [Lots of dots deleted]...
```

[Dots march across the screen to indicate that something is happening—the program hasn't died.]

```
Number of files in the file hierarchy = 31268
Number of wk1 files examined = 8979
```

[These numbers appear in the report, not only on the screen.]

```
Tue, Oct 12, 1999 04:54:12 PM   [Time when the program finished its work]
```

[Elapsed time for this run: about 2 hours, 8 minutes.]

A sample report

Here is a heavily edited version of the report for this run—the original had almost 38,000 lines.

The converter we were evaluating does not create a file that displays floating-point numbers consistently. Since we were interested only in one tag, we marked it as level 5 in RiskEditor. All the other tags were set to 1.

```
Examiner
-----
/usda/ftp/usda/data-sets/crops/94018/budget.wk1:
    Risk Level 5
        Tag 14: NUMBER: Floating point number —
             Qty: 584
```

[There are 584 cells with floating point numbers.]

```
-----
/usda/ftp/usda/data-sets/crops/94018/charactr.wk1:
    Risk Level 5 There are no tags in this file at
       this level
```

[In this file, there are no floating-point numbers. We can trust the converter we are evaluating to convert this file successfully.]

```
-----
/usda/ftp/usda/data-sets/crops/94018/conf_int.wk1:
    Risk Level 5
        Tag 14: NUMBER: Floating point number —
             Qty: 59
-----
```

[...Deleted lines...]

```
ERROR:
/usda/ftp/usda/data-sets/crops/.district/.finderinfo/
parsline.wk1 not a supported file type
```

[This file was a text file with information about the wk1 files in a directory.]

```
-----
```

[...Deleted lines...]

```
-----
/usda/ftp/usda/data-sets/crops/.district/parsline.wk1:
    Risk Level 5 There are no tags in this file at
       this level
-----
/usda/ftp/usda/data-sets/livestock/89032/acheesu.wk1:
    Risk Level 5
        Tag 14: NUMBER: Floating point number — Qty: 182
-----
Number of files in the file hierarchy = 31268
Number of wk1 files examined = 8979
```

Summary of our approach

1. We created a spreadsheet that exercises all of the .wk1 file's attributes.
2. To test a file conversion application's capabilities, we converted the file from .wk1 to .xls.
3. We visually compared the files, point by point, to uncover any inconsistencies between the two versions.

4. We examined the specifications for the .wk1 file format to identify the internal tags governing the at-risk attributes.
5. We used the RiskEditor program to configure the Examiner program, marking tags at risk.
6. We examined the collection of files with the Examiner program, which returned a report detailing the files containing the at-risk tags.

Appendix D

Case Study for Image File Format

1. Collection and Analysis of Source and Target File Format Related Information

Investigation Test Bed

To assess the risks associated with file format migration for digital image collections, the project team selected one of Cornell University Library's digital image collections as a test bed. The Ezra Cornell Papers consist of correspondence, financial and legal records, court proceedings, and other documents pertaining principally to the Cornell family, the telegraph industry, and the founding of Cornell University. The collection is composed of 30,000 images stored on small computer system interface (SCSI) disks. They are scanned as 600 dpi, 1-bit TIFF 5.0 ITU Group 4 images. Tag(ged) Image File Format (TIFF) is one of the most popular raster image file formats and is often the format of choice for master image files. It is platform-independent and supports 1- to 24-bit imaging using a variety of compression methods.

The Ezra Cornell materials were scanned in-house using a Xerox scanning system. This system organizes and stores the structuring information (e.g., page number, folder number) in a format called Raster Document Object (RDO), which is Xerox's adoption of the International Office Document Architecture (ODA) and Interchange Format.[1]

Goals of the File Format Migration Investigation

The goals of the file format migration investigation for image files were to:

[1] ODA, which became an ISO standard in 1988, has been developed to represent and allow the interchange of office documents. It contains facilities that allow both the structure and content of complex multimedia documents to be represented. Although ODA is an open standard, specifications for the RDO architecture are proprietary.

- identify the TIFF file format attributes at risk during migration,
- assess the need to move these TIFF 5.0 image files to the current version (6.0),
- evaluate the risks involved in converting TIFF 5.0 files to TIFF 6.0 files,
- investigate the status of upcoming revision to TIFF (7.0),
- assess the risks involved in skipping a generation (TIFF 6.0) and waiting for the release of TIFF 7.0, and
- assess risks and data loss associated with converting from RDO format to the open Cornell Digital Library (CDL) format.

Collection and Analysis of Source and Target File Format Related Information

To identify digital image format attributes at risk, the project staff collected and analyzed information on different versions of TIFF file format. The research process included the following:

- Conducting a literature search on digital archiving issues pertaining to digital image collections, with a specific focus on migration and the effects of file format choice in the migration chain.
- Investigating new digital preservation research and initiatives, such as ISO’s Open Archival Information System (OAIS) (International Organization for Standardization 1998), WGBH’s Universal Preservation Format (UPF) (Shepard and MacCarn 1999), and the Digital Rosetta Stone Model (Heminger and Robertson 1998), among others.
- Conducting a literature and projects survey to determine the extent of work performed on developing risk analysis based on image files.
- Reviewing risk-assessment tools developed for various purposes, focusing on the form and functionality of these tools and how they can be adapted for the purposes of this project.
- Exploring the dependencies that extend beyond basic image file format attributes, such as internal and external relationships between images and their accompanying metadata files (viewing images as “digital objects” and examining their metadata, associated scripts, programs, etc.).
- Identifying the attributes of digital images that are at risk during format migration, including the effects of migration on metadata, and various scripts and programs that support retrieval and management of the collection.
- Investigating the existing and emerging bitmap image file formats with a focus on their longevity and other archival attributes.
- Exploring vulnerabilities associated with file format migration and identifying risks associated with “migrating” or “not migrating” these files, with a focus on TIFF files.
- Analyzing the factors involved in decision making in migration projects, such as reformatting a collection of images from TIFF 4.0 to TIFF 5.0 format.

- Examining and comparing the TIFF file format specifications for Versions 4.0, 5.0, and 6.0.
- Exploring the future of TIFF as a file format, with a focus on the characteristics of the TIFF 7.0 format under development.
- Investigating the issues introduced by storing structuring metadata in Xerox RDO format.
- Identifying the risks involved in converting RDO files to the CDL format (http://www2.hunter.com/docs/rfc/rfc1691.html).

An outcome of this research process is summarized in Table 1, which categorizes the risks associated with file format-based migration.

Conclusions of the Source and Target File Format Analysis

Because most of the specifications are publicly available on the Adobe FTP site, the project staff was able to gather a substantial amount of information about the different versions of TIFF. TIFF was developed by Aldus and Microsoft, and the specification was owned by Aldus, which in turn merged with Adobe Systems, Incorporated. Consequently, Adobe now holds the copyright for the TIFF specifications. TIFF is a highly flexible and platform-independent file format. It is supported by numerous image-processing applications. A great strength of the TIFF file format is its file header option, which enables recording within the file itself of a wide variety of metadata (descriptive, administrative, and structural). The set of fields or "tags" in TIFF is extensive, making it the format of choice for most archival reformatting. However, a large number of TIFF fields are not defined by the standard. Therefore, while TIFF offers the advantage of being open and usable, there is the danger that different institutions will define these fields in different ways, leading to problems of compatibility. Another flexibility of TIFF that causes confusion is related to byte order. For example, the TIFF format permits both MSB ("Motorola") and LSB ("Intel") byte order data to be stored, with a header item indicating which order is used.

Tracking the TIFF 7.0 development turned out to be a challenging task. The project team's attempts to contact TIFF 7.0 developers, Adobe, and even TIFF listserv subscribers were fruitless. The TIFF 7.0 development group seems to be determined not to release any information regarding their work. Therefore, the project team was unable to make any comparisons between TIFF 7.0 and the earlier versions. After conducting an extensive evaluation and comparison of TIFF 5.0 and TIFF 6.0 specifications, the team ran several tests to compare the quality and utility of a subset of TIFF 5.0 images before and after conversion to TIFF 6.0. This exploration revealed no major differences between the versions. The project team concluded that there were no risks involved at this point in leaving the testbed images in TIFF 5.0 format. After reaching this conclusion, the team shifted its focus for the risk-assessment study for image files to an

RISK CATEGORY	EXAMPLES
Content fixity (bit configuration, including bit stream, form, and structure)	Bits/bit streams are corrupted by software bugs or mishandling of storage media, mechanical failure of devices, etc.
	File format is accompanied by new compression that alters the bit configuration.
	File header information does not migrate or is partially or incorrectly migrated.
	Image quality (e.g., resolution, dynamic range, color spaces) is affected by alterations to the bit configuration.
	New file format specifications change byte order.
Security	Format migration affects watermark, digital stamp, or other cryptographic techniques for “fixity.”
Context and integrity (the relationship and interaction with other related files or other elements of the digital environment, including hardware/software dependencies)	Because of different hardware and software dependencies, reading and processing the new file format require a new configuration.
	Linkages to other files (e.g., metadata files, scripts, derivatives such as marked-up or text versions or on-the-fly conversion programs) are altered during migration.
	New file format reduces the file size (because of file format organization or new compression) and causes denser storage and potential directory-structuring problems if one tries to consolidate files to use extra storage space.
	Media become more dense, affecting labels and file structuring. (This might also be caused by file organization protocols of the new storage medium or operating system.)
References (the ability to locate images definitively and reliably over time among other digital objects)	File extensions change because of file format upgrade and its effect on URLs.
	Migration activity is not well documented, causing provenance information to be incomplete or inaccurate (a potential problem for future migration activities).
Cost	Long-term costs associated with migration are unpredictable because each migration cycle may involve different procedures, depending on the nature of the migration (routine migration vs. paradigm shift).
	The value of the collection may be insufficiently determined, making it impossible to set priorities for migration.
	Costs may be unscalable unless there is a standard architecture (e.g., centralized storage, metadata standards, file format/compression standards) that encompasses the image collections so that the same migration strategy can be easily implemented for other similar collections.
Staffing	Staff turnover and lack of continuity in migration decisions can hurt long-term planning, especially if insufficient preservation metadata is captured and the migration path is not well documented.
	Decisions must be made whether to hire full-time, permanent staff or use temporary workers for rescue operations.
	Staff may have insufficient technical expertise.
	The unpredictability of migration cycles makes it difficult to plan for staffing requirements (e.g., skills, time, funding).
Functionality	Features introduced by the new file format may affect derivative creation, such as printing.
	If the master copy is also used for access, changes may cause decreased or increased functionality and require interface modifications (e.g., static vs. multiresolution image, inability of the Web to support the new format).
	Features that are not supported in other file formats may be lost (e.g., the progressive display functionality when Graphics Interchange Format [GIF] files are migrated to another format).
	The artifactual value (original use context) may be lost because of changes introduced during migration; as a result, the “experience” may not be preserved.
Legal	Copyright regulations may limit the use of new derivatives that can be created from the new format (e.g., the institution is allowed to provide images only at a certain resolution so as not to compete with the original).

Table 1. Risks associated with file-format-based migration for image collections

examination of storing structural metadata in the proprietary Xerox RDO format. The team will continue to monitor the development of TIFF 7.0.

Raster Document Object Files

An RDO file contains information about the structure of an image document as well as a file location pointer for each page image in that document. A single TIFF file represents each page in the document. The TIFF files each contain the digital data from the scanned page and a header that describes the characteristics of the image file. Because the Xerox Documents on Demand (XDOD) system is proprietary, the structure of image documents can be displayed only by using the appropriate Xerox software.

2. Selection and Evaluation of Conversion Software

Since a decision was made to maintain the files in TIFF 5.0 format, evaluation of the TIFF conversion software was unnecessary. There are several conversion programs on the market for converting TIFF files to various TIFF versions and other file formats (e.g., TIFF to GIF, TIFF to PNG). TIFF 5.0 to TIFF 6.0 conversion could be interpreted as an update rather than as a migration process.

In 1994, Cornell undertook a project to convert the proprietary RDO files to an open CDL format. The specifications for the CDL, which were released in August 1994 through a Request for Comments (#1691), defines an architecture for the storage and retrieval of Cornell University Library's image collection. Similar to RDO files, the CDL document structure provides direct access to the components of image collections (e.g., pages, sections, and chapters).

While the project team's main interest was exploring the export of files created on XDOD 3.0, its immediate concern was with the older RDOs, especially in light of the Y2K compliance issues (i.e., concern that the XDODs would no longer work unless an expensive upgrade were implemented).

The conversion from XDOD RDO to CDL format involved two steps. Cornell used a Xerox-supplied tool (XDOD Export Tool) to convert the RDO files into a series of ASCII metadata files. This tool is old and can run only in Windows 3.1, and its dissemination is authorized "only pursuant to a valid written license from Xerox." Second, through a locally developed PERL script, the ASCII metadata files were converted to the CDL format. These CDL-formatted structural metadata files are used for navigating through a document (http://moa.cit.cornell.edu/MOA/EZRA.html). The Cornell University Library information technology staff wrote the ASCII RDO-to-CDL program.

RDO-to-CDL conversion cannot be achieved through a single software tool since Xerox has not released any RDO specifications.

3. Development of Tools for Assessing the Source-To-Target Format Transfer

No specific software tool was developed to analyze the effects of migration from RDO to CDL format, because all files created using the XDOD scanning system possess identical information fields.

4. Comparison and Analysis after Conversion to Source File Format

The comparison was done manually by comparing the structural metadata elements that were captured in RDO files with the CDL structure. The team compared the list of structural metadata elements captured during scanning with the CDL structuring requirements. All the structural elements mapped to the CDL structure, and there was no loss. Even if there had been a loss, the project team decided that it was much riskier (actually detrimental) to leave the structuring information in an unsupported proprietary format.

5. Releasing the Export Tool to Other Institutions

As part of this project, Cornell investigated the possibility of further developing the Export Tool and making it available to other institutions that have legacy collections in the proprietary Xerox RDO format. This investigation was spurred by two concerns. First, several institutions had requested access to the tool over the past few years, but only Yale University had secured permission from Xerox to use it. Second, in early summer 1999, Xerox informed Cornell that the XDOD 2.x scanning workstations would not be Y2K-compliant without an expensive upgrade. Because Cornell had begun to phase out use of the XDOD systems and had converted all RDO files to the CDL format, our concerns over the millennium focused on our sister institutions' collections.

We initially considered developing the Export Tool into more generic software for external use, but quickly concluded that this would be both expensive and time-consuming. Cornell did not receive any specifications from Xerox for the proprietary tool, and the software developer at Xerox indicated that he doubted that the company still had the tools and specifications to make the system work. We decided to focus on securing permission to release the current version of the Export Tool. A two-year effort to obtain a blanket permission from Xerox to make the tool broadly accessible had stalled, so we turned to documenting the extent of the problem, concluding that Xerox might be more amenable to a very limited release.

In late April 1999, Cornell posted the following announcement on 11 listservs.

Export Tool to Convert Xerox RDO Files to Open Digital Library Format
Has your institution created digital image files using the proprietary Xerox Documents on Demand software that generates Raster Document Objects (RDOs) to store structural metadata? Cornell University is seeking feedback from these institutions to determine what demand there would be for freeware to convert those RDOs for use in other metadata applications. Cornell has used the RDO2CDL export tool to migrate RDOs to ASCII metadata files that recreate the logical and physical structure format of the RDO (called CDL). If your institution is interested in utilizing such an Export Tool, please send contact information and a brief description of your needs to: Anne R. Kenney (ark3@cornell.edu).

By early June, surprisingly few responses were received. Universities with files created on XDOD 2.5 or older versions included Harvard, Penn State, the University of Tennessee–Knoxville, and Yale. Those responding with files created using XDOD 3.01 or DigiPath included the Hein Publishing Company, Illinois State Library, the National Document Center (Athens), Indiana University, the University of Toronto, and the National Oceanic and Atmospheric Administration (NOAA) Miami Regional Library (which was considering using the technology).

Inquiries to Xerox about releasing the tool to this group resulted in further clarification that the RDO Export Tool software would work only as configured on XDOD Version 2.x systems. The format of the RDO changed slightly from version 2 to version 3, and the Export Tool would not convert the structural data on version 3 or higher systems. The Hein Publishing Company had used the tool with version 3 files through a collaborative project with Cornell, but only page labels, not structuring information, were exported. William Anderson, the Xerox software engineer who created the tool, suggested that it would be possible to get the structure information out of the version 3 RDO files, but it would take a programmer with knowledge of the Office Document Architecture (of which RDO is a variant), fair knowledge of Unix tools, and a copy of the RDO Version 3 specification, which Xerox seemed unwilling or unable to make available publicly. Anderson suggested that, “If customers are looking to buy DigiPath today, and they need that facility, they should ask for it.” Xerox decided to grant access to this software only to XDOD 2.x customers who were not migrating to DigiPath.

From June to early September, efforts continued to reach legal agreement with Xerox over the release of the Export Tool software to XDOD 2.x users. Cornell received a copy of a proposed Software License Agreement on August 26, 1999. The agreement granted the institution a nonexclusive, perpetual, royalty-free license to use the software and the right to provide a sublicense only to those institutions that had reported using the XD Version 2.x systems, collectively referred to as "Authorized Educational Institutions" (AEI). Lee Cartmill, the chief financial officer at Cornell University Library, expressed concern about the indemnity clause in the agreement, which required Cornell to "defend, indemnify and hold Xerox harmless from and against any and all third party claims that arise from or relate to the Software and their respective use of the Software." Cornell attempted to have this clause modified. When Xerox remained adamant, Cartmill drafted a Software Sublicense Agreement that would require the AEIs to extend the indemnity and limitation of liability to Cornell University. As of this writing, the four institutions have been notified of these stipulations, and their legal advisers are reviewing copies of the agreements. It remains to be seen whether any or all of these institutions will agree to these license stipulations, but Cornell will not sign the agreement with Xerox unless they do so.

References

International Organization for Standardization. ISO Reference Model for an Open Archival Information System (OAIS). 1998. Available from http://ssdoo.gsfc.nasa.gov/nost/isoas/us/overview.html.

Heminger, Alan R., and Steven B. Robertson. 1998. Digital Rosetta Stone: A Conceptual Model for Maintaining Long-Term Access to Digital Documents. Available from http://crack.inesc.pt/events/ercim/delos6/papers/rosetta.doc.

Shepard, Thom, and Dave MacCarn. 1999. UPF: Universal Preservation Format. Available from http://info.wgbh.org/upf/.

Appendix E

Case Study for Lotus 1-2-3 .wk1 Format

1. Introduction

The United States Department of Agriculture (USDA) Economics and Statistics System is a heavily used collection of agricultural economic information that is evenly divided between time-sensitive economic reports and data series of numeric files in spreadsheet format. Some of the data series are 14 years old—ancient artifacts in the context of personal computers. Although dated, these files have historic and current value to a well-defined group of data users. The goals of this case study were to:

- evaluate file format migration as a strategy to maintain access to these numeric files,
- identify file format components at risk during migration,
- identify related risk attributes associated with migration, and
- evaluate data migration software.

The case study was conducted with several risk-assessment tools developed during this project. These include a specially prepared test file, file reader software, and a risk-assessment workbook.

2. Description of Archive and Data User Base

The USDA Economics and Statistics System is a joint venture between the Albert R. Mann Library of Cornell University and three USDA economic agencies: Economic Research Service (ERS), National Agricultural Statistics Service (NASS), and the World Agricultural Outlook Board (WAOB). These three agencies measure the production and health of domestic and international agricultural activities. ERS publications analyze current agriculture market activity and forecast future market conditions. Other ERS publications offer economic analyses in the areas of trade, production, rural development, farm inputs, and other economic topics. The NASS publishes estimates of production, stocks, inventories, deposition, utilization, prices of agricultural commodities, and such other items as labor and numbers of farms. The WAOB issues regular forecasts of United States and world supply-and-demand prospects for major agricultural commodities.

During the federal fiscal year 1999, more than 368,000 distinct hosts accessed the USDA Economics System. The system disseminates more than 500 MB in 7,000 file downloads daily. Many system users compare current and historic statistical series to detect trends. In a 5,000-user survey conducted between January and April 1999, 64 percent of respondents used the service for monitoring price trends, 47 percent for forecasting or obtaining market predictions, and 38 percent for research. Most of the more than 250 data series in the Economics System are published in a DOS binary format, usually Lotus 1-2-3 .wk1. The collection contains nearly 9,000 .wk1-formatted files stored online on SCSI disks, and offline on CDs and on floppy magnetic diskettes.

3. The 1-2-3 Format

Lotus 1-2-3 applications have undergone continuous revisions. Early versions of Lotus 1-2-3 created stand-alone .wk1 files. Functions and macro languages were associated with the application, not the data files. Special formatting instructions were saved in separate .fmt files. Since the Windows 3.1 release of Lotus, data and format files have been merged into a .wk4 format. Later releases have integrated scripting language and data objects into the files. The most recent release of 1-2-3, the Millennium edition, allows the user to embed hyperlinks within the spreadsheet, manipulate 1-2-3 files within Active Document containers such as Internet Explorer or Lotus Notes, embed ActiveX controls within documents, import real-time information into a spreadsheet, and so on. The current Lotus file is no longer just a spreadsheet; it is better described as an interactive data container.

As the format has evolved, all features have been maintained and supported through backward compatibility, and representatives at Lotus Corporation have underlined the company's continued support for all 1-2-3 file formats. Limited backward compatibility for 1-2-3 files is found in Excel. Microsoft provides extensive documentation for Excel and identifies the functions associated with 1-2-3 that do not convert properly to Excel. None of these problems appears to affect functions associated with .wk1 formatted files, but the extent of change is hard to measure fully.

Both companies keep information about some features of their software privileged. 1-2-3 files use a proprietary format closely controlled by Lotus (now IBM). The file specifications for release 1.0 were published in 1984 and revised for release 2.0 in 1985. Specifications for releases 3.0 to present have been controlled by agreements with business partners and software developers. No one has researched or documented the frequency and scope of these changes, or how well non-Lotus developers integrate these revisions into their products. An unfortunate side effect of proprietary restrictions on this information was the apparent loss of early 1-2-3 specifications within the Lotus Company. As the company revised its product, old specifications were fully integrated into the new releases, and the need for the old documentation disappeared. Further effort is required to assemble an authoritative set of specifications for all 1-2-3 and other major spreadsheet software releases.

Finally, early 1-2-3 files are bereft of descriptive data. Until Lotus 97, the Windows 95-compatible release, file names adhered to the 8.3 DOS format. Embedded descriptive data are often typed into cell A1. File names for USDA products are simplistic. Files labeled “table1.wk1” are quite common among the more than 250 data sets. None of the .wk1 files in our collection has any imbedded links to its respective documentation files. The main method of identifying a file is through its relative position in a hierarchical file structure.

4. Development of Tools for Assessing the Source-to-target Format Risk

File Migration and Black-Box Converters

A major risk in migrating collections of files is the conversion software used to translate the files from the original format to the target format. A migration project begins with a file whose content should be translated, without corruption, to another format. The file is passed through conversion software, essentially a “black box,” with the intent that the integrity of the content be preserved in the new file. We can presume success if we know that the conversion software faithfully maps every property of the source format to corresponding features in the target format (assuming, of course, that the target format has a feature set that is rich enough to store the properties and data of the source). For example, if the source format we are converting has a way to indicate bold text, assuming that the target format can also indicate bold text, we want to know that the conversion software correctly maps the source bold attribute to the target bold attribute. More important, in most conceivable cases, data values—whether numeric, image, or text—should also move from one format to the other intact.

Two Ways of Evaluating the Black Box

If we can examine the mapping process and the data-moving techniques of the conversion software, we can evaluate the correctness of both. This method must be repeated for every combination of source and target formats with which we are working, because each combination has a unique mapping. Moreover, to attempt this method, we must have access to the source code of the converter and possess the expertise to evaluate that code. Our experience with approaching commercial software vendors for the code to their programs has not been fruitful. Even if it were, the cost of evaluating each combination of source and target formats in the program algorithm makes this method impractical for controlling risk involved in migrating collections.

Another method is to compare a converted file with the original file. If the result meets our standard of success, whatever that standard may be, we can say that the conversion software has performed adequately. However, this method is limited to the particular source and target file formats under consideration. The ideal file for the test would be one that exercised all the capabilities of the source format and contained every possible feature and data value at the minimum and maximum of every range possible. If that file were run through the converter, the resulting file could be compared at every point with the original. This is the approach we chose for our case study.

Developing Risk-Assessment Tools

For our own collection of Lotus 1-2-3 files, we created a test file in the .wk1 format. The features documented in the user manuals provided with the 1-2-3 release 2.2 software, along with the published file format specifications, determined the content of the test file. The test file was generated with 1-2-3 release 2.2 software, which, according to our estimates, was the software that generated the oldest .wk1 files in our collection. With the test file, we can evaluate potential conversion software by running the software on the test file and then comparing the converted file with the test file. Visual inspection and comparison of all the properties and values is necessary to identify differences; this process took about three hours in our example. Proprietary software code and knowledge of an uncertain number of format-to-format mappings in the program are unnecessary for visual inspection. Another benefit of the test file is that it gives us a baseline against which we can evaluate and compare multiple conversion applications.

Whichever method is used to evaluate the conversion software, if any of the properties or data values are not the same in the source and target files, then we know that the conversion software has introduced one or more points of risk. Thinking about the whole collection of files to be migrated, we will want to know whether some of the files in the collection have any at-risk properties that will combine with the conversion software to create problems in the converted file(s). We can then decide whether to find another converter, to refrain from migrating those files, or, perhaps, to consider some or all of the loss acceptable.

We wrote the Examiner software application to test a collection of files for the presence of particular properties. The list of properties consists of the structural element tags that define the .wk1 format. Following the results of the test file comparisons with different conversion software, we were able to identify the tags that would not properly translate into the target format. Using a companion configuration utility, RiskEditor, we provided the Examiner program with a list of the properties that are at risk, optionally ordering them by the degree of importance or impact. Then, as a test, the Examiner program was run against parts of the collection. After further evaluation, Examiner was run against the entire collection, a hierarchy of 30,000 files. This requires a little more than two hours to complete. The program produces a report showing which files contain the properties in question. Examiner provides a quantitative assessment of the risks that could be introduced by the conversion software.

The Examiner application is written in Java, and both its user documentation and technical documentation are available as HTML files. It has been designed to be extendable to any file format that indicates properties as numbered tags, including Lotus 1-2-3 and TIFF, the formats of our case-study collections. A requirement for running the application is a Java interpreter on the computer holding the collection. We wrote a command-line version of the program to be used on our Unix servers, but the program could be easily extended to have a graphical user interface.

5. Selection and Evaluation of Conversion Software

The migration software examined for this project was a commercially available, off-the-shelf (COTS) product. Locally developed conversion software was avoided for two reasons: development costs and immediate obsolescence. Software development is labor intensive, and long-term maintenance is expensive. The costs to develop single, one-project programs cannot be justified for mainstreamed software formats. We were interested in examining the alternatives that commercial software developers might offer. We examined two products, DataJunction and Conversions Plus, using the following criteria:

- source and target formats,
- accuracy of conversion,
- file decompression,
- batch processing, and
- error reporting.

We evaluated other features, but these five criteria describe the critical features we thought essential when evaluating software for a migration process.

DataJunction

A product of Data Junction Corporation, DataJunction 7.0 is conversion software that appears capable of integrating and transforming data among hundreds of applications and structured data formats in both UNIX and Windows 95/98/NT environments. We counted 150 source and 155 target format options. Image-file formats—GIF, JPEG, PDF, TIFF—are not included in the list of supported file formats. For our case study, we narrowed file format options to Lotus 123 r2 and Excel 97.

We found DataJunction works as specified. In single-file tests, DataJunction quickly and accurately converted Lotus .wk1 files into Excel .xls files. Tests were not conducted to determine whether DataJunction could convert .wk1 to ASCII. Setting up DataJunction, however, was somewhat difficult. DataJunction is designed to work with large legacy database files, from which it extracts data and converts it into the target format of choice, using a complex set of rules. The graphical user interface is not intuitive to use and was mastered only after detailed review of the online documentation and considerable trial and error. We did not investigate the possibility of designing transferable conversion templates. DataJunction was very difficult to set up for batch-mode processing, a major problem if a migration project must process more than a few files. In addition, DataJunction does not have the ability to decompress archived or zipped data files. DataJunction can be configured to provide different error messages, including fatal and general errors, warnings, and information messages. DataJunction is handicapped by its batch-job restrictions. An upgrade in this feature would make this program suitable for conversion of a variety of standard file formats.

Conversions Plus 4.5

Conversions Plus 4.5 (CP 4.5) is a product of DataViz Corporation. It is a stand-alone program that complements several other software products in a suite of tools designed to read and/or write to a variety of file formats. We counted 74 source and 110 target formats available in four general file categories: word processing, spreadsheets, database files, and image files. For our case study, we narrowed file format options to Lotus 123 (.wk1) and Excel 97.

File conversion in Conversions Plus is implemented by pairs of file readers and writers. Each reader is written to read and decode a specific file format. The file reader identifies file format components and stores them within the program in a standard way. From this data template, each file writer program can extract each specific information object and restructure the data into the new file format. In addition, CP 4.5 can detect and uncompress files using the following compression algorithms: gzip, zip, tar, and Z. Conversions Plus can work on single files or in batch mode. CP 4.5 works in the Windows 95/98/NT environment.

We found that Conversions Plus provided accurate translations from .wk1 to .xls formats with the exception of a subset of floating-point numbers. We used our standard test file in a conversion test and discovered CP 4.5 read and displayed the source file properly, but would embed an incorrect display code for certain floating-point numbers in the .xls target file. Comparison tests with Excel indicated display problems with fractions that were represented with exponential notation. All other basic format attributes—text strings, integers, formula, and equations—converted properly.

The graphical user interface is intuitive and easy to use. One selects the file(s) or directory (directories) by clicking on them. Setting up target format choices or directories for converted files is easily done using pull-down menus. CP 4.5 works less smoothly when a directory contains numerous files of mixed format, a common situation on many servers. Batch preferences are limited to a single file format type for each of the four general categories of file type. In these situations, we would anticipate significant user oversight of the conversion operation. Conversion and error statistics are displayed at the end of a batch job, and the information can be written to a log file.

Comparison and Analysis After Conversion to Source File Format

During our tests of DataJunction and Conversions Plus, we manually compared the standard test file and other sample files in two states: before and after conversion. The comparison was conducted on two Windows NT workstations using two monitors of similar size and features. The comparison entailed a line-by-line examination of structural and data elements in each file. Conversion errors occurred in Conversions Plus only on data that contained floating-point numbers. We modified the Examiner program to identify .wk1 files in our archive that contained a structural element for floating-point numbers (tag 0Eh) and ran it against our collection. The program indicated that 8,619 files, or 96 percent of the collection, contained floating-point numbers. Because this is a significant portion of the collection, conversion using Conversions Plus was not attempted.

6. Migration Risk Analysis for 1-2-3 .wk1 Files

We examined three migration options for 1-2-3 files: backward compatibility, and file migration to Excel .xls and to ASCII characters.

Backward Compatibility

Backward compatibility of 1-2-3 files provides a baseline for comparisons with Excel and ASCII. Data captured in older 1-2-3 files are still readable in more recent 1-2-3 software. This strong backward compatibility support indicates that old file formats are superseded, but not obsolete; older files are not “orphaned” by major revisions in Lotus software. In addition, earlier application software can operate either in DOS operating systems or in the DOS emulator in Windows 95/98/NT. Although 1-2-3 has a reduced share of the spreadsheet market, Lotus is still providing strong support for this product. Evaluating these factors, it appears that avoiding migration and relying on the backward compatibility to sustain the .wk1 format can be considered a low-risk option.

Migration to Excel .xls Format

Excel is currently the market leader for spreadsheet software in both Windows and Macintosh operating systems and it has established a large corporate and private user base. If we can make predictions on the basis of the documented history of 1-2-3, Excel .xls should be a heavily used format for the next 10 years. We also believe that if Excel is superseded by another spreadsheet program, Microsoft will provide reliable migration software from Excel to the new target format. Examination of our standard test file and a random sampling of files from the archive indicate that the latest version, Excel for Windows 97, provides an accurate conversion of 1-2-3 .wk1 files. Four major components of a spreadsheet file—text strings, integers, floating-point numbers, and embedded formulas and functions—were properly converted; they retained content, context, and a reasonable reproduction of the “functional experience.” Our evaluation indicates that migration of .wk1 to .xls format is a low-risk option. Unfortunately, Excel itself is a poor choice for conversion software because it cannot perform batch conversions of files.

Migration to ASCII

ASCII is the format of choice for large numeric file archives. ASCII files are easily scanned manually and can be imported into most software programs. ASCII is perceived as a low-maintenance format. On the basis of three decades of experience, most digital archivists predict ASCII will still be a common file format in 50 years. Given the proper circumstances, a single migration to ASCII should be more cost-effective than repeated migrations through other evolving file formats. At its most fundamental level, migration of 1-2-3 files to ASCII converts the content of spreadsheet cells to values located in a matrix of *x–y* coordinates. The actual values of embedded functions, equations, or pointers to other cells will be retained, but the functions, equations, or pointers in those cells will be lost. Long text strings, essentially embedded metadata, are truncated at different lengths, depending on the conversion software. The formula, functions, pointers, and text strings could be recorded in an external conversion record, but for large collections, this might be impractical.

Our assessment suggests that ASCII is a low-risk preservation option for the .wk1 format and could be adopted if dependence on backward compatibility or conversion to Excel .xls format were impossible.

7. Metadata Risk

There are six sources of metadata to coordinate:

1. MARC record. Each of the 250+ data sets is cataloged and has a full MARC record in the campus online catalog.
2. Gateway record. Each data set also has a MARC-like record for the Cornell University Library gateway to online resources.
3. USDA Economics and Statistics System Database. Each data set also has a detailed record in a searchable database designed specifically for USDA System users.
4. README documentation. Each data set has a separate ASCII text README file. The README documentation is a mix of descriptive, content, and administrative metadata.
5. File hierarchy. Placement and selection of individual 1-2-3 files depend on descriptive data associated with directory names.
6. File names. 1-2-3 files use the DOS 8.3 naming convention. A common name for a .wk1 file in the collection is table01.wk1. The file extension (.wk1 or .xls) is a necessary feature for the file to be read by a spreadsheet application.

Migration of the 1-2-3 files to another format would require modification of two of these six metadata sources: the file name and the README files.

- The file name modification would change the file name extension.
  For example: table01.wk1 → table01.xls
  The conversion software should implement the file name modification. A possible risk could be introduced if the file name were manually changed without processing by the conversion software.
- README documentation would be modified to reflect the new file name extension. For instance, the following table example was extracted from a sample USDA data set, Feed Yearbook:
  Table12.wk1 12. Farm programs and participation, 1975–1998
  Table13.wk1 13. Average prices received by farmers, United States, by month, and loan rate, 1975 / 76–1998 / 99
  The file name extension in the documentation would be converted from
  Table12.wk1 → Table12.xls
  Table13.wk1 → Table13.xls

If this conversion is required, the file extension conversion can be modified in a word processor using a find / replace feature. Using powerful search-and-replace functions on the documentation introduces a low-level risk. Modification of documentation will need thorough review before the migration process is considered complete. In addition, new administrative information describing the migration process will need to be added to the record. This would include, but not be restricted to:

- date of migration or modification,
- data set involved,
- description of conversion process,
- identification of conversion software,
- identified risks and actions taken to manage them, and
- person supervising the migration or modification.

Because of the limited modifications required to file names and documentation and the lack of system scripts, the migration risk introduced to metadata is low.

8. Summary

From these findings, we believe we can draw the following conclusions:

1. 1-2-3 format-specific migration risks associated with specific file conversion software can be identified and described. The use of a standard test file allows side-by-side comparison of file structure and content before and after conversion. Errors or significant changes to the information can be isolated and examined for their persistence or effect, or for both.
2. Some measures of risk can be quantified. The Examiner program can measure the number and frequency of problematic file tags within a collection. Although the measurement is crude, it provides a general measure of risk for a collection, a step that has been lacking in digital file management. Further refinements to the program could provide greater resolution of problem file attributes.
3. Migration paths for the .wk1 format can be mapped. We can describe the relationship between the source and target formats and the conversion software with reasonable certainty. For example:

Source	**Black Box**	**Target**	**Risk**
.wk1	Excel	.xls	Low
.wk1	DataJunction	.xls	Low
.wk1	Conversion Plus	.xls	High

4. Cost-effective COTS migration software was not identified. The two programs analyzed in this case study were considered the best candidates for a COTS migration program. Both were found deficient in basic, but significant, tasks. Minor modifications to the programs would alleviate our reservations. We intend to bring our results to the attention of these developers.