Disasters & Quiet Catastrophes: Preserving your Digital Assets

Danielle Mericle
Fiona Patrick
October 25, 2011

Cornell University
Library

Learning Goals

- Understand basics of digital preservation
- Be able to identify, assess, and strategize for high-risk material within your collections
- Understand basic copyright law when applied to preservation
- Implement an action plan for your institution

Course outline

9:30-10:30	Introduction & overview of challenges
11-12	Digital preservation basics, technical and institutional strategies
1-1:30	Copyright- what you need to know
1:30-2:30	Action Planning & Implementation strategies
2:45-3:30	Case Studies – Questions -

Part I
Introduction

What are digital resources?

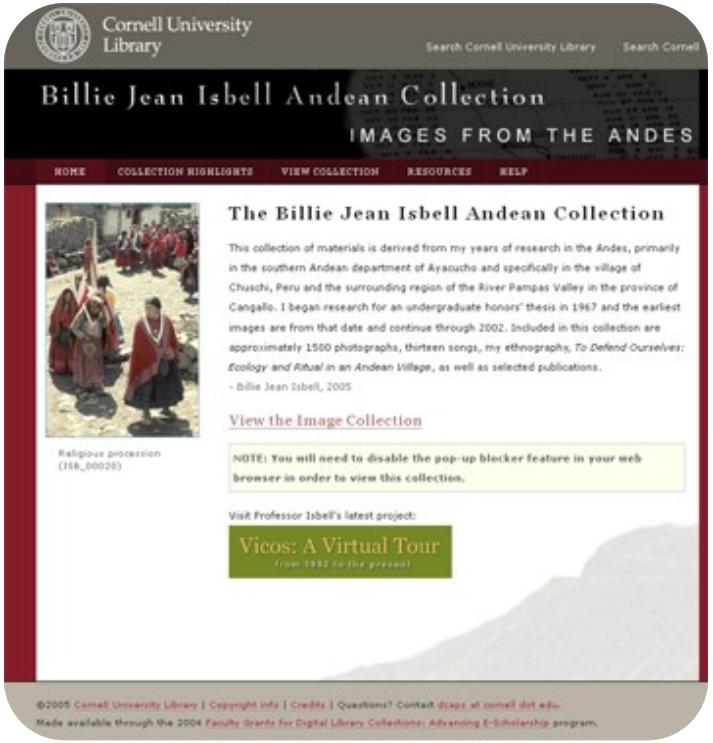
The following, existing singularly or in combination, could reasonably be described as a digital resource:

- An electronic text
- A series of digital images
- A database
- Multimedia and layered files
- A website

A digital resource is in *machine readable format*, a binary language that the computer can understand: 0000101001000001.

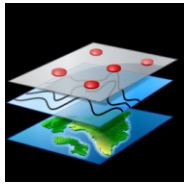
Important distinction

- Born-digital versus digitized content



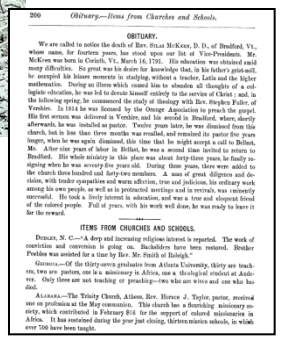
Born-digital resources

- Files that are created natively on electronic devises, such as computers, cell phones, digital cameras, and digital audio recorders
- Including but not limited to:
  - Electronic documents (.doc/ .pdf / .xls)
  - E-mail
  - Images, still and moving
  - Sound recordings
  - Websites
  - Interactive art


Digitized resources

- Analog objects that are transferred to a digital format through some conversion process.
  - Paper documents/printed books
  - Photographic materials like slides, prints, or glass-plate
  - 3D objects
  - Audio such as cassette tape and LPs
  - Film and other moving images



200 *Obituary.—Items from Churches and Schools.*

OBITUARY.

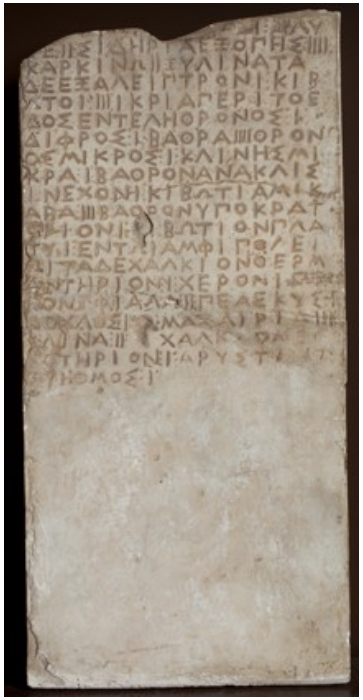
We are called to notice the death of Rev. Silas McKeen, D. D., of Bradford, Vt., whose name, for fourteen years, has stood upon our list of Vice-Presidents. Mr. McKeen was born in Corinth, Vt., March 16, 1791. His education was obtained amid many difficulties. So great was his desire for knowledge that, in his father's grist-mill, he exempted his leisure moments in studying, without a teacher, Latin and the higher mathematics. During an illness which caused him to abandon all thoughts of a collegiate education, he was led to devote himself entirely to the service of Christ; and, in the following spring, he commenced the study of theology with Rev. Stephen Fuller, of Vershire. In 1814 he was licensed by the Orange Association to preach the gospel. His first sermon was delivered in Vershire, and his second in Bradford, where, shortly afterwards, he was installed as pastor. Twelve years later, he was dismissed from this church, but in less than three months was recalled, and remained its pastor five years longer, when he was again dismissed, this time that he might accept a call to Bethel, Me. After nine years of labor in Bethel, he was a second time invited to return to Bradford. His whole ministry in this place was about forty-three years, he finally resigning when he was seventy-five years old. During these years, there were added to the church three hundred and forty-two members. A man of great diligence and devotion, with tender sympathies and warm affection, true and judicious, his ordinary work among his own people, as well as in protracted meetings and in revivals, was eminently successful. He took a lively interest in education, and was a true and eloquent friend of the colored people. Full of years, with his work well done, he was ready to leave it for the reward.

ITEMS FROM CHURCHES AND SCHOOLS.

Dudley, N. C.—A deep and increasing religious interest is reported. The work of conviction and conversion is going on. Backsliders have been restored. Brother Peebles was assisted for a time by Rev. Mr. Smith of Raleigh."

Georgia.—Of the thirty-seven graduates from Atlanta University, thirty are teachers, two are pastors, one is a missionary in Africa, one a theological student at Andover. Only three are not teaching or preaching—two who are wives and one who has died.

Alabama.—The Trinity Church, Athens, Rev. Horace J. Taylor, pastor, received one on profession at the May communion. This church has a flourishing missionary society, which contributed in February $14 for the support of colored missionaries in Africa. It has sustained during the year just closing, thirteen mission schools, in which over 700 have been taught.



Why distinguish between the two?

- People often forget about born-digital objects when considering preservation
  - Websites, Microsoft Office applications, digital photographs

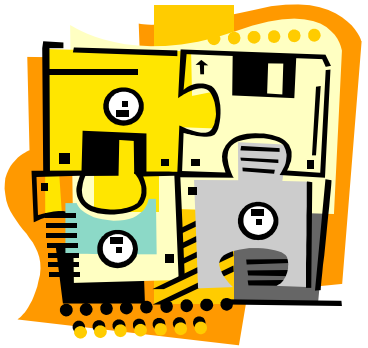
- Different types of digital objects *may* require different care
  - Established best formats for digital photographs and text
  - Digital audio and video formats are constantly improving/changing

- Digitization is not the same as preservation- digital files require extensive preservation efforts

Digital resources live on digital media

- Digital files stored on various digital storage media
  Three primary types:
  1. Disk: Magnetic hard-drives; CD/ DVD's; hard disk; floppies
  2. Tape: Mini-dv; Digi-beta; HD
  3. Solid state: Flash drives; memory sticks; smart media

- Further complicated by player longevity/compatibility
  - Computer operating system
  - Hardware availability/obsolescence
  - Software versioning/ obsolescence



Digital Preservation Risks

- File format or media obsolescence
- Media degradation or failure
- Insufficient metadata
- Insufficient control (eg copyright)
- Authenticity and provenance is unclear
- Multiple copies are not synchronized
- Natural disasters

Digital Preservation Challenges

- Insufficient funding/ staffing/ expertise /infrastructure
- Lack of preservation plan & policy
- No institutional commitment
- Lack of understanding of importance of preservation from key-stakeholders
- Many others…….

Part II
The Basics of Digital Preservation

What is Digital Preservation?

"A whole range of activities designed to extend the usable life of digital information and protect them from media failure, physical loss, and obsolescence."—Libraries & Archives Canada

"Digital preservation combines policies, strategies and actions that ensure access to information in digital formats over time"—ALA working definition

Preservation versus backup

- Back-up is the "periodic capture of information to guard against system or component failure or against accidental or deliberate corruption of the system or system metadata." —Trustworthy Repositories, Center for Research Libraries, March 2007

- ***Backing up is not the same as digital stewardship***

Goals of preservation activities

- Creation of digital objects that are, over time, -
  - Authentic
  - Renderable
  - Understandable
  - Viable
- Establishment of sustainable preservation system & accompanying institutional policies

Digital preservation strategies- *technical*

- Standardization
  - File formats / File naming
  - Minimum metadata requirements
- Basic metadata for preservation
- Copying & replication
- Refreshing
- Migration / Normalization / Localization
- Data recovery
- Equipment / technology
- Emulation
- Storage & backup

Standardization

Benchmarking strategies

- Use known, widely adopted file-formats vetted by community, preferably non-proprietary (http://www.digitalpreservation.gov/formats/)
- Establish consistent naming convention throughout collections
- Establish minimum metadata requirements
  - Technical
  - Administrative
  - Descriptive
  - Preservation

Preservation metadata

- Provenance: for authentication and a documented history of the file's contents
- Context: why the data was created, how it relates to other data
- Reference identifiers: ISBN, accession number, etc. to demonstrate the relationship between the digital file and any physical holding you have
- Technical: to describe the technology environment used to create the digital objects and suggest how the files might be read/used

Bitstream copying & replication

- Bitstream copying: the making of an exact duplicate of a digital object

- Replication: keeping many copies of the same digital object, preserving copies variously, with the hope that one of them will still be viable when it is needed

- *Combined copying, replication, and metadata form baseline preservation strategy*

Refreshing

- Refreshing is to copy digital information from one long-term storage medium to another of the same type with no change whatsoever in the bitstream.

- Example- Reburning CD's in collection every three years

Migration

- When a format is at risk for obsolescence, migration will convert data from one technology or format to another while preserving the essential characteristics of the data.

- Some inherent challenges to migrating data- always do thorough quality analysis of material (preferably in an automated fashion)

Normalization

- When a file is in a less than optimal format for preservation, a "normalized" version will be created in a preservation-worthy format, to be archived along with original

- Example- PDF broken into page-level TIFFS (note, some functionality lost, such as hyperlinks, etc)

Localization

- When a file contains links to other files, a localized version of file will be downloaded and maintained in the repository
- Example- a link to an XML file with a metadata schema definition
- Guarantees that references between files can always be resolved

Data recovery

- Data Recovery is rescuing content from damaged media or hardware
- Usually performed by commercial vendors prepared for broken CDs and other critical damage
- This is an emergency recovery strategy ONLY!

Technology preservation

- Technology Preservation is to preserve the historic technological environment-- equipment, software, operating systems

- Requires space / resource allocation, but is sometimes the only realistic solution given financial limitations

Emulation

- Emulation combines software and hardware to reproduce the essential characteristics of a different computer so that media designed for one environment can be used in another one.

- Can be quite challenging and not fully represent original object

Storage strategies

- Optical Media (CD's & DVD's)
- Magnetic tape
- External drive
- RAID (redundant disk arrays)
- Consortium (LOCCKS, Meta-archive)
- Commercial (Amazon, OCLC)

Storage strategies

- Always maintain at least two versions of master images in different geographic locations
- Distinguish your masters from your access files- severely limit access to master images
- When using localized media (disks, external drives), use common-sense handling- minimize dust, jarring, temperature fluctuations, magnets, UV, etc.

Other technical strategies

1) *Start early* in the project/program (when possible)
2) Create extensive documentation
3) Flexible design- have components separated out for easy migration, repurposing, so that each component can be updated, altered or removed without interfering with another part of the system.
4) Understand minimum functional requirements and cost/benefit of option to convert in future iterations

Digital preservation activities- *institutional*

- Develop preservation policy & plan
- Establish digital rights management, including access provisions
- Educate community
- Secure institutional commitment

A preservation policy:

- Clearly articulates roles & responsibilities among organization (who does what)
  - Maintains & migrates data (back-end)
  - Ensures ongoing access to data (front-end)
  - Monitors & ensures ongoing financial support for preservation
- Defines scope/length of preservation
  - Short/medium/long-term preservation models (may vary according to data-type)
  - Rate of refreshment
  - Redundancy
- Defines ultimate accountability for program

Types of preservation:

- Short-term
  - Access to digital materials either for a defined period of time while use is predicted but which does not extend beyond the foreseeable future and/or until it becomes inaccessible because of changes in technology.
- Medium-term
  - Continued access to digital materials beyond changes in technology for a defined period of time but not indefinitely.
- Long-term
  - Continued access to digital materials, or at least to the information contained in them, indefinitely.

Preservation policy considerations

- Align policy with goals and mission of institution
- Include all key stakeholders in policy creation
- Couple with other relevant policies, such as IP & IT standards (including benchmarking for minimum digitization requirements)
- Should be documented/dated/signed policy by all stakeholders

Preservation plan

- Relevant policies
- Selection- content types, scope, etc.
- Roles & responsibilities
- Institutional support/funding
- Format types, at appropriate benchmarks
- Metadata types, minimum requirements
- Migration, refreshment strategies
- Storage, including back-up / redundancy
- Copyright compliance
- Access and permissions, both for ingest and download

- ***All within an integrated technical architecture***

Digital Rights Management

- Acquire and maintain contractual and legal rights and responsibilities

- Includes, but not limited to:
  - Agreements with faculty regarding preservation and use of contributed items
  - Understanding of copyright law and fair use for preservation and education
  - Access provisions to content in preservation repository

Education

- Develop educational program for faculty, staff, and administrators to encourage widespread adoption of best practices and establish buy-in for importance of digital preservation
- Cover basic topics on creating preservation-worthy content, including scanning & metadata creation; copyright law; repository functionality
- Include overview of policies set by institution

Institutional commitment

- Recognition of value/ importance by key decision makers
- Financial support
- Plan or program for responsible digital stewardship
- Policies in place for preservation; selection criteria; collection development

Making your case…

"Framing the benefits from preservation in ways that emphasize outcome rather than process helps place the cost/benefit analysis underpinning digital preservation in its proper perspective."

(Sustaining Digital Resources, JISC, 2009)

Outcomes:

- Maintaining current access to digitized content
- Serving broad user base
- Institutional recognition
- Future possibilities for new use and innovation
- Reducing costs over the long-term

Preservation plan

- ✓ Develop relevant policies
- ✓ Determine selection- content types, scope, etc.
- ✓ Define roles & responsibilities
- ✓ Secure institutional support/funding
- ✓ Benchmark
  - ✓ Format types and settings
  - ✓ Metadata types, minimum requirements
- ✓ Migration, refreshment strategies
- ✓ Storage, including back-up / redundancy
- ✓ Copyright compliance, access and permissions, both for ingest and download
- ✓ Determine your solution for integrated technical architecture

Different strategies

- Baseline – inhouse
  - Minimum requirements
- Outsourced - OCLC
- Partnerships / consortiums
  - CLOCKSS
  - Meta-archive
- Full scale development- inhouse based on existing open-source software
  - DAITSS
  - Fedora

Baseline preservation requirements

- Standardized file formats & naming conventions
- Maintenance of "master" images separate from "access" images
- Plan for refreshing and/or migrating data
- Metadata capture and management
  - Technical- device; color management info; date; operator; format; inhibitors
  - Descriptive- any bib info; existing or new data
  - Administrative- project info; provenance; history
- Redundant storage
  - Preferred geographic separation and media variety

OCLC

- Outsourced digital archive- managed storage; monitoring & reports; integrated workflows
- Works with CONTENTdm; Fedora; D-Space
- Supports mandatory PREMIS information
- Can support in file-migration / normalization
- Somewhat expensive

LOCKSS/ CLOCKSS

- (Controlled)/ Lots of Copies Keep Stuff Safe
- LOCKSS can preserve local collections, including thesis, images, AND subscription content from participating publishers. CLOCKSS is primarily for publisher generated content. Both are geared towards newly generated content and born digital.
- Distributed preservation model- storage nodes exist throughout world
- Provides an OAIS-compliant, open source, peer-to-peer, decentralized digital preservation infrastructure. It is format-agnostic, preserving all formats and genres of web-published content, provided the content has an authoritative version. The intellectual content, which includes the historical context (the look and feel), is preserved. Content preserved by libraries in their LOCKSS Box becomes a part of their collection, and they have perpetual access to all of it.
- Annual fee

Meta-archive

- Collaborative, community led initiative to provide low-cost, high-impact preservation services to help ensure the long-term accessibility of the digital assets of universities, libraries, museums, and other cultural heritage institutions
- Using a technical framework that is based on the LOCKSS (Lots of Copies Keep Stuff Safe) software, collections are ingested into a geographically distributed network where they are stored on secure file servers in multiple locations. These servers do not merely back up the materials, but rather provide a dynamic means of constantly checking each file and providing repairs whenever necessary.
- Requires programmer/ hardware commitment
- Annual fee

DAITSS, Fedora, and Planets

- Skeletal preservation repository application as open source software
- Dark Archive (DAITSS) / optional (Fedora)
- Supports full OAIS functional model & METS metadata with PREMIS compliancy
- Requires robust inhouse programming & hardware

Part III
Copyright for Preservation

Copyright & Preservation

- Institutions need to obtain the legal rights to preserve digital objects and make them accessible
- Complexity
  - Migration copies, archival copies, derivative versions, and changing over time
- Strategy
  - Procedures, protocols and documentation

Right to copy

- Digital preservation – copying occurs at some point
- The *exclusive right* to copy belongs to the author
- Therefore, digital preservation may impinge on these rights

- **If almost everything is copyrighted, and copyright author has extensive exclusive rights, how can any digital preservation occur without infringement?**

3 Exemptions

- 17 U.S.C. § 117. Limitations on exclusive rights: Computer programs

- **17 U.S.C. § 108. Limitations on exclusive rights: Reproduction by libraries and archives**

- 17 U.S.C. § 107. Limitations on exclusive rights: Fair use

17 U.S.C. § 108

- Ground Rules
  - Open to public
  - Not for direct/indirect commercial advantage
  - Any copies must carry a © notice
- Plus
  - Own a copy of the original
  - Solely for preservation
  - Original must be "damaged, defective, lost or stolen" OR existing format is obsolete
  - Reasonable investigation finds unused copy can't be obtained at a fair price
- Conclude – 3 copies allowed, if digital then access must be limited to the premises

17 U.S.C. § 107

- 4 Factors
  - Purpose of the use (transform or merely replicate)
  - Nature of originals (primarily creative or factual)
  - Amount duplicated
  - Effect on potential market or value of the original

- Fair use is a case-by-case basis, subject to judicial interpretation
- Digital preservation in a gray area of legality

Part IV
Action Planning & Implementation

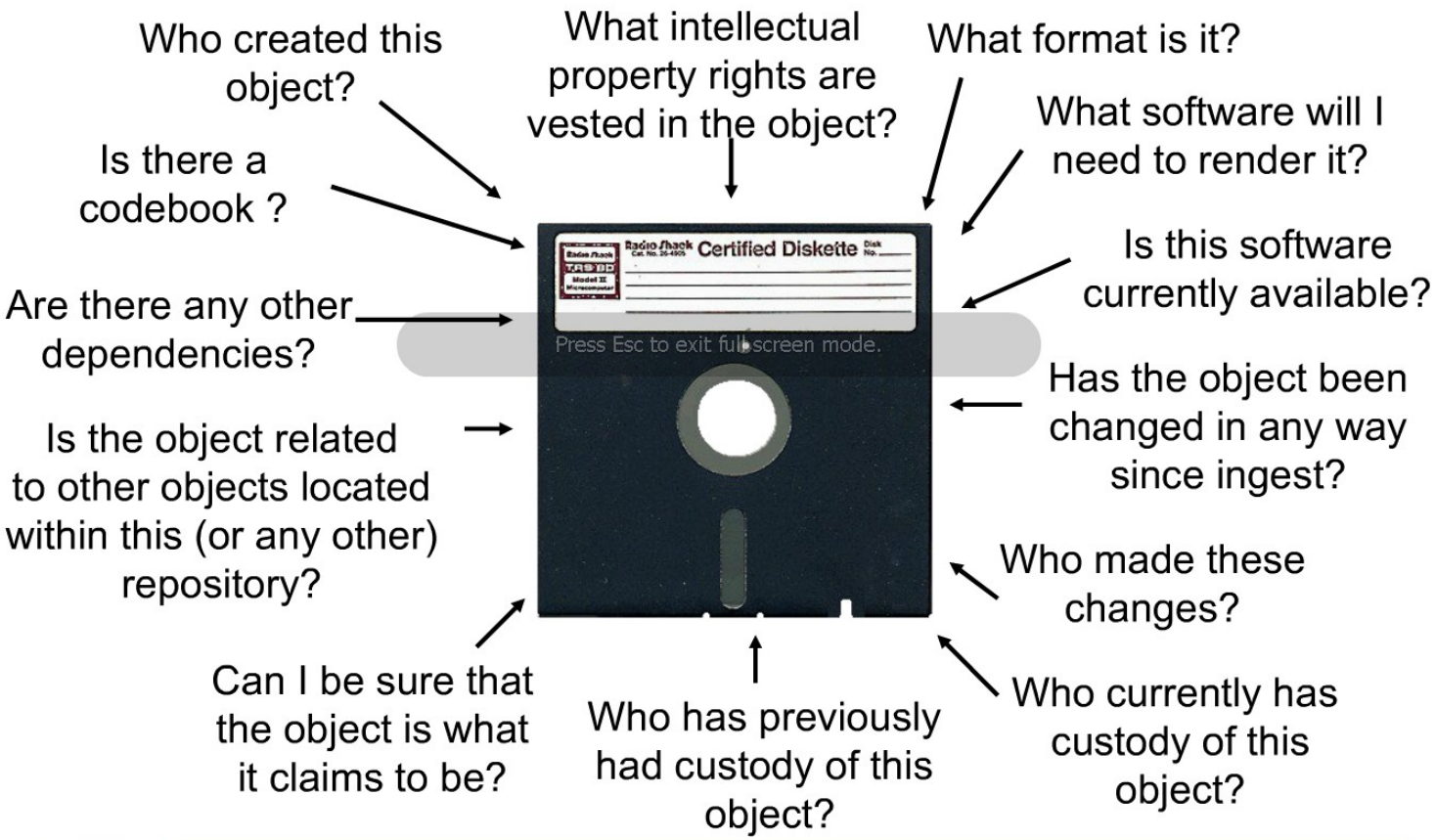
Strategy – Analysis
Risk Management

Strategy - Analysis

- Know what you have
- Know how to handle your digital materials
- Know what is coming ahead
- Know how to keep it safe
- Identify your needs
- Identify your resources
- Document your decisions

Strategy – Inventory/Survey

- Format Data
- Content Data
- Storage
- Access Data
- Technical Resources
- Growth
- Needs


Who created this object?
What intellectual property rights are vested in the object?
What format is it?
What software will I need to render it?
Is there a codebook ?
Is this software currently available?
Are there any other dependencies?
Has the object been changed in any way since ingest?
Is the object related to other objects located within this (or any other) repository?
Who made these changes?
Can I be sure that the object is what it claims to be?
Who has previously had custody of this object?
Who currently has custody of this object?
Radio Shack Certified Diskette
Press Esc to exit full screen mode.

Collection Analysis

- Size of existing collection: number of items and storage requirements
- Format of objects
- Relationships between objects
- Anticipated growth rate- existing and new collections
- Existing metadata
- Copyright/ access restrictions
- Search functionality for objects
- Vulnerabilities

Policy analysis

- Who coordinates & approves preservation policies? Copyright compliance?
- Who creates and manages content?
- What content will be retained and for how long?
- Who can submit to the repository?
- Who can download or view content?
- What metadata standards are used?
- Who owns the content?
- How is copyright and IP controlled?

Risk Management

- Digital preservation is not just about
  - Data
  - Access
  - Risk
- Digital preservation is about
  - People and opportunity
    - People and technology change
      - An ongoing process

Source: William Kilbride, Digital Preservation Coalition

Risk Management - Simple

- Identify a risk
- Consequences of that risk
- Likelihood
- Impact
- Score = Likelihood * Impact
- Frequency of occurring
- Check strategy
- Responsibility
- Response strategy

Digital Preservation Risks

- File format or media obsolescence
- Media degradation or failure
- Insufficient metadata
- Insufficient control (eg copyright)
- Authenticity and provenance is unclear
- Multiple copies are not synchronized
- Many others…….

DigitalPreservationCoalition

Starting with risks ....

RISK	Likelihood	Impact	Score	Frequency	Owner	Response
Media will degrade	5	5	25	ongoing		Technology watch Good media storage Refreshment Routine checks of media Keep copies on different media ...
Media will fail outright	3	5	15	ongoing		Technology watch Careful procurement Warranties and liabilities Multiple copies Develop migration plan ...
Media obsolesence	5	5	25	ongoing		Technology watch Refreshment Routine checks of media Multiple media ...

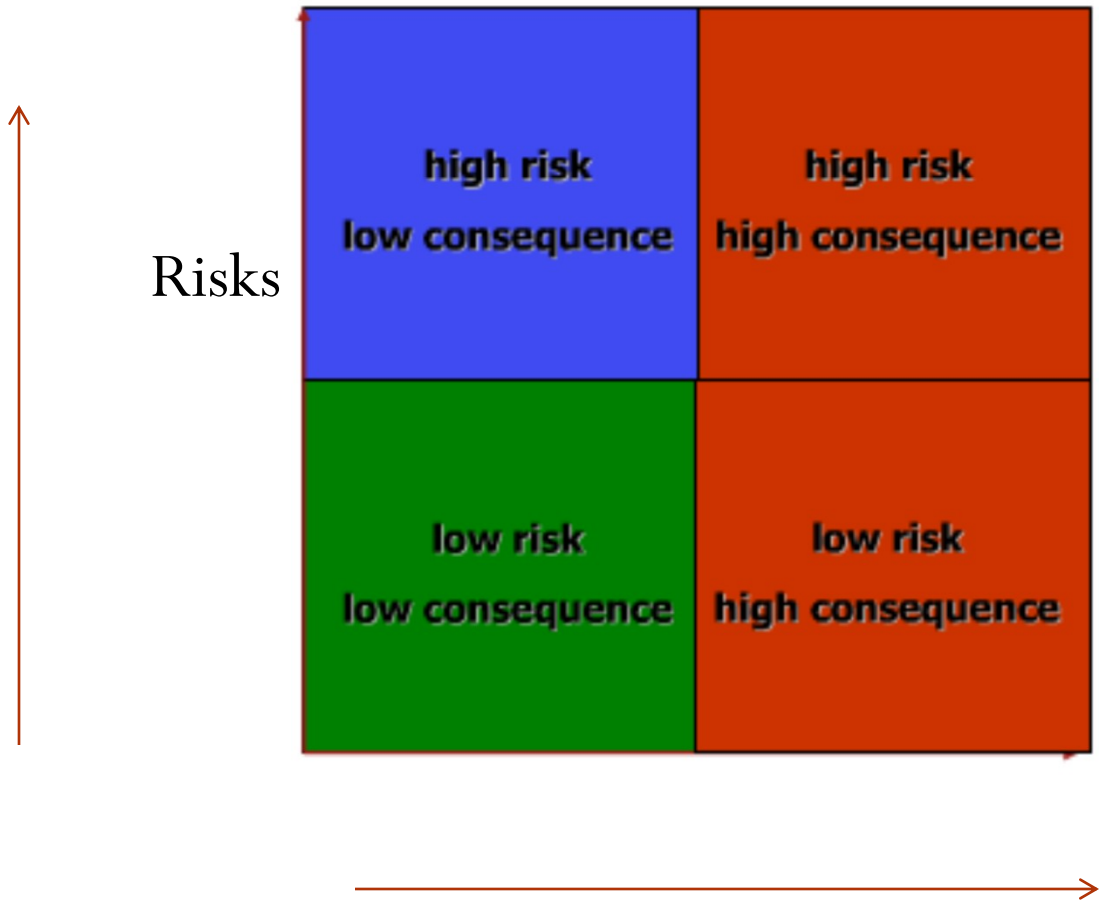
DigitalPreservationCoalition

Post mitigation risks ....

RISK	Likelihood	Impact	Score	Frequency	Owner	Actions to maintain mitigation
Media will degrade	3	1	3	ongoing		Technology watch Good media storage Refreshment Routine checks of media Keep copies on different media ...
Sudden media failure	1	4	4	ongoing		Technology watch Careful procurement Warranties and liabilities Multiple copies Develop migration plan ...
Media obsolesence	1	3	3	ongoing		Technology watch Refreshment Routine checks of media Multiple media ...

O ‘g

Measuring Risks & Consequences



Additional Workshop Resources

Digitization Projects – Cornell University Library public wiki

https://confluence.cornell.edu/display/digitres/Digitization+Resources

Tiny Link: https://confluence.cornell.edu/x/NRgUBw