docbyte

Beyond Bit Rot

A Risk Assessment Framework for Digital Preservation Amid Cyber Threats and AI

Wim Provoost



docbyte

www.docbyte.com

Agenda

1. Why preservation risk is broader than bits
2. The changing risk landscape
3. A practical assessment approach
4. Applying the approach in practice
5. Key takeaways

docbyte

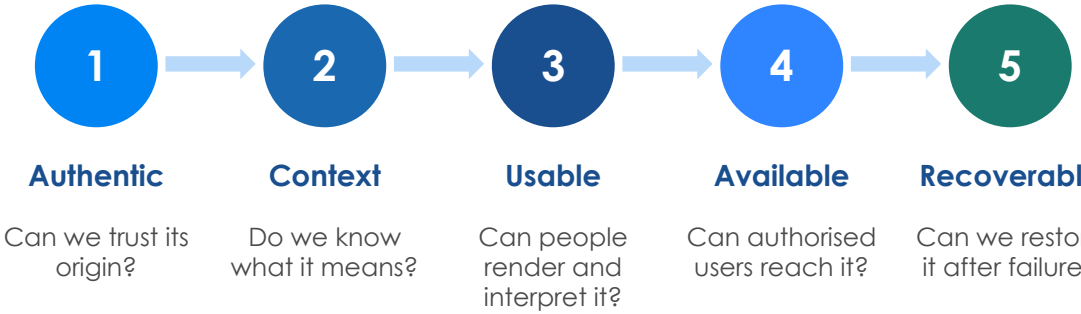
Part 1 · Why preservation risk is broader than bits

Evidence is more than a hash

docbyte

A record can be intact and still be lost

> **“The checksum still matches.”**
>
> But can anyone prove what it is, understand it, open it and restore it?



Preservation risk is anything that could stop records being trusted, understood or used.

docbyte

Four cases. None of them was bit rot.



2023

British Library, UK

Ransomware. Copies existed, but not the systems to restore and find them.

Lost: access

Source: British Library review (Mar 2024)



2024

National Museum of the Royal Navy, UK

A ransomware attack disrupted museum systems, including its collections system. Recovery took around 18 months.

Lost: recovery

Source: Museums Association (Dec 2024); DPC (2026)

Fina

2016

Fina, Croatia

About one million signed and timestamped records. Certificates expired; chains and revocation lists were not kept. Validation: "indeterminate."

Lost: authenticity
signatures can't be validated

Source: InterPARES Trust case study (2016)



2025

Auschwitz Memorial, Poland

AI-generated "photos" of Holocaust victims with invented stories spread online.

Lost: context

Source: Auschwitz Memorial via AFP (May 2025)

The bits mostly survived. Access, recovery, context or authenticity did not.

docbyte

Part 2 · The changing risk landscape

Preservation risk goes beyond bit rot

docbyte

What changed: from protecting bits to protecting dependencies

	Then: the classic view, from the 1990s	Now: after cloud, ransomware and generative AI
Systems and change	Records lived in one system for a long time	Records move between systems, clouds and suppliers every few years
People and daily work	A small team knew the collection by heart	Work is spread over teams and suppliers; ownership blurs
Cyber attacks	Archives were rarely attacked	Archives are ransomware targets (British Library 2023, Navy museum 2024)
AI tools	Records were made and described by people	Since 2022, AI tools classify, transcribe and generate content
Bits and formats	Main worry: media decay, bit rot, obsolete formats	Still real, and largely handled by fixity, copies and migration

✅We have mature controls for bit integrity.✅
⚠️The other dependencies need the same attention⚠️

docbyte

Four connected risk domains shape preservation

Systems and change

- Formats or software age
- Migration loses details
- Systems are retired
- A provider changes or exits

People and daily work

- Nobody is clearly responsible
- Steps are not written down
- Skills or time are missing
- Recovery has not been tried

Cyber attacks

- Ransomware or stolen passwords
- Records are deleted or changed
- Copies face the same attack
- Action logs are unavailable

AI tools

- Wrong labels hide records
- Changes are hard to explain
- Sensitive data might leak
- Updates change results

The domains add up: at the British Library, an attack on old systems cost months of access.

docbyte

Risk accumulates across the preservation lifecycle



Record made	Move into custody	Keep trustworthy	Find and use	Restore after failure
White House email, 2003–05 A new email system skipped automatic archiving. 22 million messages turned up on backup tapes years later. Source: NPR (Dec 2009)	**Windrush landing cards, 2010** Destroyed during an office move instead of being kept. People later needed them to prove when they arrived. Source: CNN (Apr 2018)	**BBC Domesday Project, 1986** Stored on laserdiscs for a BBC Master computer. By 2002 it was almost unreadable, and was rescued through emulation. Source: The Observer (Mar 2002)	**Ivar Aasen Centre, Norway** An encrypted catalogue of 11,000 works became inaccessible when the only people who knew the passwords died. The code was eventually cracked. Source: The National Archives (UK), (2017)	**Kyoto University, 2021** A faulty update to a backup script deleted 77 TB of research data. Four research groups lost files for good. Source: Tom's Hardware (Jan 2022)

Preservation depends at every step on people, processes and technology, and all of them change and introduce risks.

docbyte

Part 3 · A practical assessment approach

Start from what you care for, not from the threat

docbyte

A simple set of questions for assessing preservation risk

1	**What are we caring for?**	Name the records, collection or service, and what must remain usable.
2	**What could happen?**	Consider changes, mistakes, supplier problems, cyber incidents or AI use.
3	**What would be affected?**	Could trust, context, understanding, access or recovery suffer?
4	**How likely and how serious would it be?**	Rate the event and its consequences using shared terms.
5	**What will we do, and who will check?**	Name an action, an owner and a way to test whether it worked.

Keep talking with the people who create, manage and use the records. Review the answers when things change.

docbyte

Question 5 for systems: Tolerate, Invest, Migrate or Eliminate

	Low business value	High business value
Good technical fit	**Tolerate** Keep it running as it is, but watch what your records depend on	**Invest** Upgrade or extend it: an upgrade can change data and formats, so test before and after
Poor technical fit	**Eliminate** Retire it, but archive the records with their context first	**Migrate** Move it, and test exports with future users before cut-over

Invest, migrate and eliminate are all preservation moments.

docbyte

Based on the Gartner TIME model for application portfolios

Questions 2 and 5 for AI tools in our own workflows

Source
Original record

→

AI step
Classify or extract

→

Decision
Human or rule

→

Archive
Keep record + explanation

Wrong label
Records may be missed or sent to the wrong place

Unclear changes
Staff may not know why the output changed

Data exposure
Sensitive records may leave approved systems

Tool update
Results may change after a service is updated

Question 5: keep the original, the tool output, the tool and version, and any staff decision.

docbyte

Review your plan when the archive changes

Systems	Retirement, upgrade, move or new connection
Suppliers	Service changes, contract ends or an incident occurs
Attacks	Ransomware, account theft or a failed restore
AI tools	A new tool, update or different use
Tests	A restore fails or a record cannot be found or read

Name a person to follow each issue, a next step and a date to check again.

docbyte

Part 4 · Applying the approach in practice

Three scenarios, the same five questions

docbyte

Scenario 1: The British Library

1	What are we caring for?	Digital collections, their descriptions, and the catalogue and services researchers use to find them
2	What could happen?	Ransomware hits old systems that are hard to rebuild or replace
3	What would be affected?	Access and recovery: copies survive, but nobody can find or use them for months
4	How likely and how serious?	Likely: heritage institutions are targets and the estate was complex. Serious: research stops
5	What will we do, and who will check?	TIME: migrate the legacy systems. Keep recovery copies apart; test a full restore of catalogue and search yearly

docbyte

Scenario 2: A photo archive

1	What are we caring for?	Photo collections and their descriptions, as we publish them online
2	What could happen?	AI fakes imitate or "enhance" our records; our own AI tools mislabel or alter them
3	What would be affected?	Authenticity and context: fakes are believed, real records are doubted
4	How likely and how serious?	Likely: it is already happening. Serious: trust in the archive as evidence
5	What will we do, and who will check?	Keep originals with provenance, publish with content credentials (C2PA), label, log and review AI output

docbyte

Scenario 3: FINA's expired signatures

1	What are we caring for?	About one million retirement-fund records, signed with qualified electronic signatures and timestamped
2	What could happen?	Certificates expire; certificate chains and revocation lists are not kept; validation evidence stays in FINA's system
3	What would be affected?	Trust: the signatures can no longer be independently validated. Validation returns "indeterminate"
4	How likely and how serious?	Likely: certificates always expire. Serious: the proof behind a million records, might be challenged in court
5	What will we do, and who will check?	Keep chains, revocation data and validation reports with the record at ingest; renew the evidence; test a sample yearly

docbyte

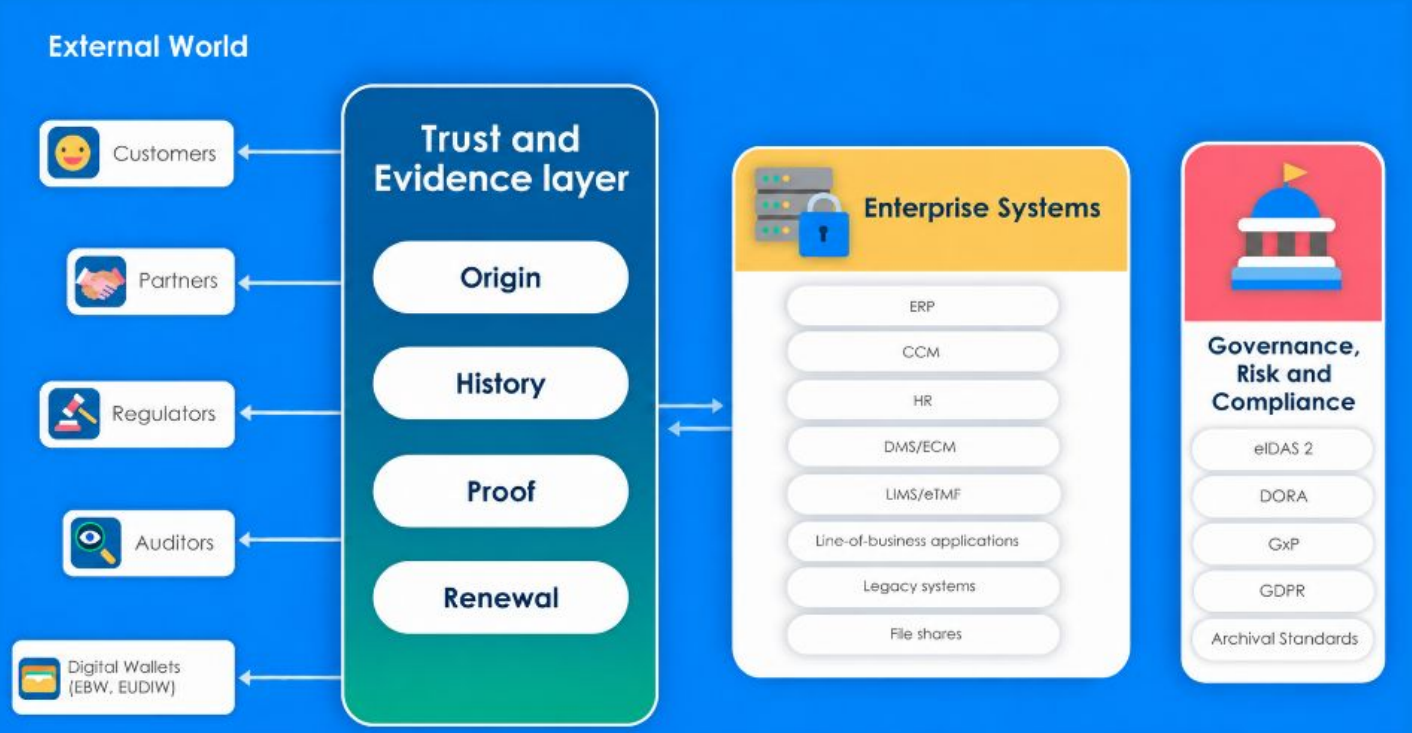
Arendelle Library: where the risks stand

Risk: event and consequence	Domain	Affects	Likelihood	Impact
Ransomware hits legacy systems: collections can't be restored for months	Cyber attacks	Access, recovery	High	High
Collections system is replaced: links between catalogue records are lost	Systems and change	Context, access	High	High
A restore fails because it was never tested end to end	People and daily work	Recovery	High	High
Key staff or a supplier leave: format and workflow knowledge is lost	People and daily work	Understanding	Medium	High
Signing certificates expire: signatures can no longer be validated	Systems and change	Trust	High	Medium
An AI tool adds wrong catalogue metadata, mixed with staff work	AI tools	Context, trust	Medium	Medium
AI fakes of collection items circulate as genuine	AI tools	Trust	High	High
Attackers steal and publish personal data	Cyber attacks	Privacy, reputation	High	High

Likelihood = how likely the event is. Impact = what it does to the records.
H = high · M = medium · L = low. Residual ratings, illustrative.

docbyte

A trust and evidence layer keeps the proof when systems change



docbyte

Arendelle Library: the same risks after treatment

Risk	Before (L / I)	Treatment	After (L / I)
Ransomware hits legacy systems: collections can't be restored for months	H H	Migrate legacy systems (TIME); packages restorable without them	M L
Collections system is replaced: links between catalogue records are lost	H H	Metadata and relationships kept in the package (METS, PREMIS)	H L
A restore fails because it was never tested end to end	H H	Full restore test every year, checked against fixity and evidence	L L
Key staff or a supplier leave: format and workflow knowledge is lost	M H	Context and procedures documented in the package and runbooks	M L
Signing certificates expire: signatures can no longer be validated	H M	Validation data kept with the record; evidence renewed (RFC 4998)	H L
An AI tool adds wrong catalogue metadata, mixed with staff work	M M	AI values marked and reviewed; tool and version logged	M L
AI fakes of collection items circulate as genuine	H H	Originals kept with provenance; published with C2PA credentials	H M
Attackers steal and publish personal data	H H	Not addressed by this: needs security controls	H H

Likelihood = how likely the event is. Impact = what it does to the records.
H = high · M = medium · L = low. Residual ratings, illustrative.

Evidence with the record lowers impact. Migrating and testing also lower likelihood.

docbyte

Part 5 · Key takeaways

Make it a habit, not a document

docbyte

Key takeaways

1. Evidence is more than a hash: a record can be intact and still be lost
2. The risks moved from bits to dependencies: systems, people, attackers and AI
3. Start from what you care for, and ask the same five questions every time
4. Every TIME decision except tolerate is a preservation moment
5. Review the answers to the 5 questions when things change.

docbyte