**How to predict long-term information integrity.***(Research Paper)*



**Abstract**

This article addresses the problem of formulating operational preservation policies that ensure bit-level information integrity over long periods, and in the presence of a diverse range of real-world technical, legal, organizational, and economic threats. We employ a systematic, quantitative prediction framework, combines formal modeling, discrete-event-based simulation, hierarchical modeling, and sensitivity analysis.

. Specifically, the framework formally defines an objective function for preservation that maps a set of preservation policies and a risk profile to a set of preservation costs, and an expected collection loss distribution. In this framework, a curator’s objection is to select optimal policies that minimize expected loss subject to their budget constraint. To estimate preservation loss under different policy conditions optimal policies we develop a statistical hierarchical risk model -- that integrated threats at four levels: the (abstracted) storage hardware; the physical environment; the curating institution; and the global (legal, economic and policy) environment. We then employ a general discrete event-based simulation framework to evaluate the results (loss and cost) of employing varying preservation strategies under specific parameterization of risks.

The framework offers flexibility for the modeling of a wide range of preservation policies and threats. Since this framework is Open Source, and easily deployed in a cloud computing environment -- it can be used to produce analysis based on independent estimates of scenario-specific costs, reliability, and risks. An analysis, based on hundreds of thousands of simulations using this framework, yields the following results.

Furthermore, by using a sensitivity analysis, to summarize the results of hundreds of thousands of simulations using this framework, we identify a number of robust and broadly applicable operational preservation policies. This analysis provides additional confirmation for the efficacy of some accepted preservation practices (such as diversifying replicas); provides novel insights into specific preservation tactics (e.g. systematic fixity checking is far more effective than random sampling), and provides evidence that contradicts receive wisdom (e.g. the widespread use of compression strongly reduces risks and costs, when optimally integrated with other strategies.).

# Significance

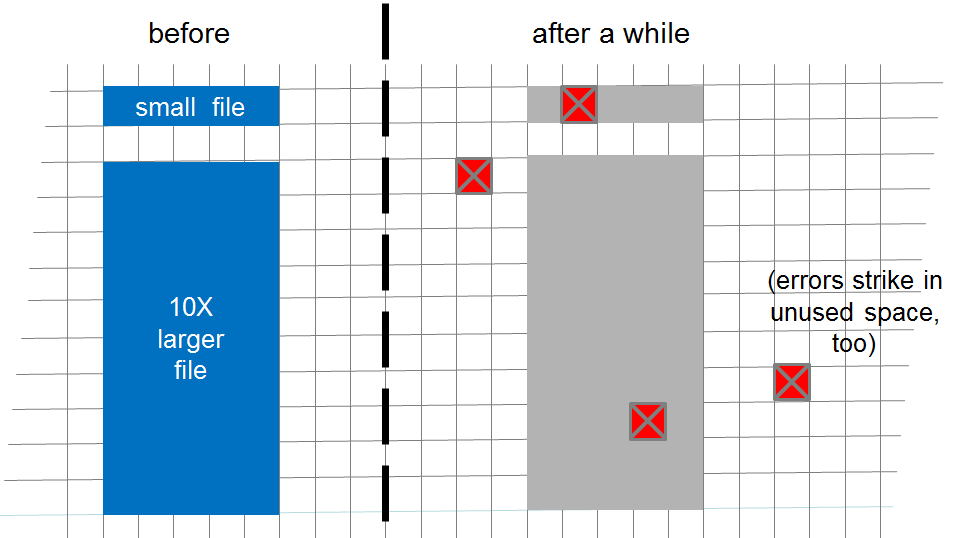
# Prior Work

# Methodology

Overall goals for methodology -- replicability, transparency, flexibility, quantitative, systematic, scalable

## A Formal Cost-Loss Characterization of Long-Term Preservation Objectives

## Documents, Colllections and Storage

* 1. 
  2. **Figure 1.** Conceptual model of sector-induced document failure.

## A Hierarchical Statistical Model that Approximates Local, Institutional and Global Risks

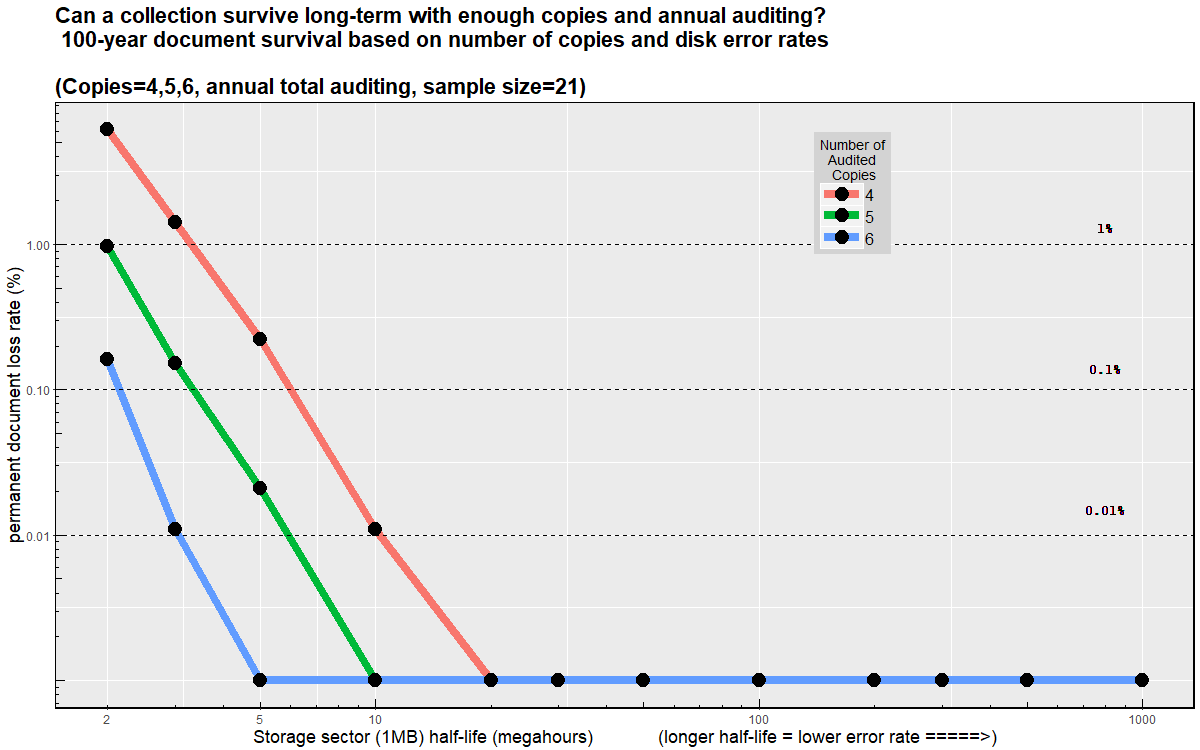
**Table 1.** A Hiearchical Typology of Preservation Threats

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Characteristics | | |  |
| Layer | Label | Role | Distribution |  |
| Hardware Storage | *Sector* | Causes single document loss | Poisson event |  |
| Local environment | *Glitch* | Increase rate of sector error | Poisson duration |  |
| Institution | *Server* | Causes loss of single replica | Exponential life |  |
| Institutional Environment | *Shock* | Causes loss of multiple replicas; increases the rate of server failure | Poisson event/duration |  |
|  |  |  |  |  |

## Estimating Loss through Discrete Event-Based Simulation

# Selected Results

## Selecting Preservation Policies that Are Robust to Storage Conditions



**Figure 2.** Analysis from discrete event simulations. Deploying five independent replicas and annual auditing is sufficient to maintain the integrity of collections across a wide range of storage level conditions.

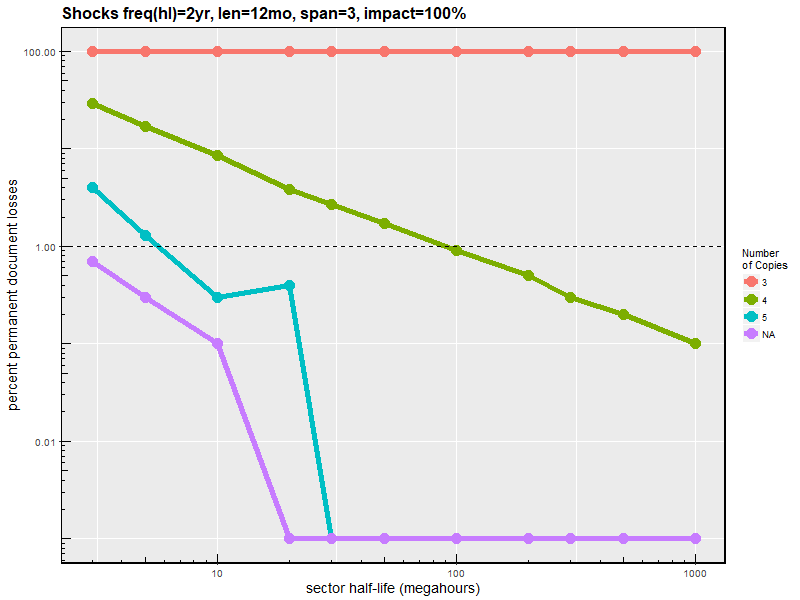
## Selecting Preservation Policies that Are Robust to Institutional Failures

Characterizing institutional threats.

**Table 2.** Modeling Scenarios

|  |  |  |  |
| --- | --- | --- | --- |
|  | Failure Frequency | |  |
| Error | Moderate | Severe |  |
| Sector | Media error | Hardware storage controller failure |  |
| Glitch | HVAC faults | Fire, flood, earthquake |  |
| Server | Curatorl Error | Cyber-attack  Bankruptcy |  |
| Shock | Recession | Regional War Government Suppression |  |
|  |  |  |  |

Even with infinitely-lived storage, institutional-life, and global shocks matter.



# Discussion

* Summary of our contribution
* Invitation to others to use code/parameterize for their risk and cost profile
* Summary Recommendations for Memory Institutions
  + Multiple copies. In most conditions 5-7. Systematic audition -- sample without replacement; higher-frequency -- x times yearly.
  + This is highly robust -- will survive a major recession or minor war
  + Use compression -- with a well-documeted oSS implemented
  + Audit vendors to ensure diversification across platforms -- avoid correlated legal/physical/technological exposure.
  + Regularly monitor measure failures.
* Community, Vendors, an d Research Recommendations
  + Support auditing primitives
  + Collect and share loss rates
  + Forget 11 nines … reveal replication strategy

# Acknowledgments

Any acknowledgements should be placed in a separate section immediately before the reference list.

# References