Research Article

AND DATA SCIENCES
BULLETIN OF COMPUTER

From Framework to Metric: Developing a Quantitative Data Risk Score for Scientific Collections

Yan Hua Dong

Department of Computer Science, The University of York, UK

Abstract

Background: The qualitative data risk assessment matrix proposed by Mayernik et al. (2020) provides a crucial foundation for identifying threats to scientific data preservation. However, its qualitative nature limits its utility for cross-collection comparison and systematic resource allocation. **Objective:** This paper presents a methodology for transforming the qualitative risk matrix into a quantitative Data Risk Score (DRS), enabling objective prioritization of data preservation efforts. **Methods:** We employed a two-stage Delphi method with 30 international data stewardship experts to assign weights to the 21 risk factors and 10 categorization methods from the original framework. These weights were integrated into a scoring algorithm. **Results:** The resulting DRS was validated against three case studies: a modern genomic repository, a legacy social science archive, and a distributed ecological network. The score effectively discriminated risk levels between collections and provided a transparent basis for prioritization. **Conclusion:** The Data Risk Score operationalizes the conceptual risk framework, providing repositories, funders, and institutions with an actionable metric to guide preservation strategy and investment.

Keywords: data risk assessment, quantitative metric, data preservation, digital stewardship, prioritization, Delphi method

1. Introduction

Ongoing stewardship is required to maintain the utility and accessibility of scientific data collections [1–4]. Collections that are not actively curated, documented, and supported by stable infrastructure are vulnerable to a wide range of risks, including technological obsolescence, loss of key personnel, inadequate metadata, unstable funding, and shifting institutional priorities [5, 6]. As research communities increasingly rely on complex, distributed, and domain-specific data resources, the cost of data loss or degradation grows correspondingly: it can undermine reproducibility, restrict the ability to build upon prior work, and erode public trust in scientific evidence [7–9]. In this environment, data managers and funders must not only recognize that risks exist, but also reason systematically about how those risks accumulate across portfolios of collections with different sizes, formats, governance structures, and user communities [5, 10].

The data risk assessment matrix developed by Mayernik et al. provides an important foundation for such reasoning [3, 4]. Their framework enumerates an extensive set of risk factors spanning organizational, technical, and policy dimensions, and it offers a structured process for evaluating the

Corresponding author: Yan Hua Dong (yanhuadong@proton.me)

BCDS 2024, Vol. 5(3), 21-34. DOI: 10.71448/bcds2453-2

vulnerability of individual collections. Because it is qualitative by design, the matrix is flexible and accessible: it can be applied in multiple institutional contexts and adapted to local needs without requiring complex statistical modeling or detailed cost information [1, 2]. However, this qualitative nature also presents a significant limitation when organizations manage dozens or hundreds of collections and must make decisions at scale. As the original authors emphasize, their framework does not provide "out-of-the-box quantification measures or data risk prioritizations" of the kind implemented in more narrowly scoped initiatives, such as the USGS risk calculation project [11]. Without a quantitative synthesis, assessments remain descriptive and difficult to compare across collections, echoing broader concerns about the need for explicit, evidence-based risk metrics in digital preservation and data stewardship [10, 12].

For data managers and funders, the central challenge is therefore not merely *identifying* the presence of risk, but *prioritizing* among competing risks and interventions under constrained resources [5, 6]. A single institution may be responsible for many collections that differ markedly in size, domain, user base, and maturity. In such a setting, qualitative ratings alone cannot answer questions such as: Which of the collections in a portfolio is most vulnerable to loss or degradation if no action is taken? Which specific risk factor, if mitigated, would yield the greatest increase in overall data security per dollar spent? How should limited investments in infrastructure, staffing, or policy development be allocated to maximize preservation impact [9, 10]? Absent a defensible, quantitative basis for comparison, prioritization often relies on ad hoc judgments, historical precedent, or the relative influence of particular stakeholders, rather than on a transparent assessment of risk [1, 2].

This paper addresses this gap by developing a quantitative Data Risk Score (DRS) derived directly from the Mayernik et al. framework [3]. The core idea is to retain the rich, expert-informed structure of the qualitative matrix while introducing a systematic procedure for assigning numerical weights to its risk factors and aggregation categories, following established practices in metric and maturity-model design [11, 12]. In doing so, we seek to answer two guiding questions: first, whether expert-derived weights can be robustly assigned to the risk factors and categorization methods of the qualitative data risk matrix; and second, whether a resulting quantitative score provides a valid and useful mechanism for prioritizing data preservation efforts across diverse collections. The DRS is not intended to replace expert judgment, but to complement it by offering a reproducible, auditable summary measure that can be compared across time, across collections, and across institutions [5, 10].

The primary contribution of this work is therefore a transparent, reproducible methodology for calculating a DRS that builds on, rather than substitutes for, existing qualitative practice. By formalizing how qualitative ratings are translated into quantitative scores, we provide a missing link between risk identification and resource allocation in scientific data stewardship [3, 4]. In the remainder of the paper, we describe the structure of the Mayernik et al. risk matrix, detail our expert elicitation and weighting procedure, present the mathematical formulation of the DRS, and illustrate its application to a portfolio of heterogeneous scientific data collections. We conclude by discussing how the DRS can support strategic decision-making, facilitate communication between data managers and funders, and guide future enhancements to risk assessment frameworks in the research data ecosystem [1, 9].

2. Background and Related Work

2.1. The Qualitative Foundation

The Mayernik et al. data risk matrix was designed as a pragmatic tool to help data stewards surface, structure, and discuss the many ways in which scientific data collections might be vulnerable over time [3, 5, 6]. The framework synthesizes 21 distinct risk factors, such as "Media deterioration," "Lack of documentation & metadata," "Loss of expertise," and "Unstable funding," along with 10 categorization methods that describe how each factor might be evaluated, including dimensions like "Severity of risk," "Likelihood of occurrence," and "Temporal immediacy." Taken together, these elements provide a rich vocabulary for articulating the socio-technical conditions that influence whether a collection remains usable and accessible, complementing broader digital preservation models such as OAIS and PREMIS [13, 14]. The matrix is deliberately modular: institutions can select the subset of factors most relevant to their context, adapt the wording to local terminology, and calibrate the meaning of qualitative levels such as High, Medium, and Low [2, 15].

This flexibility is a key strength for adoption across heterogeneous settings. A small institutional repository, a national data center, and a project-specific archive can all employ the same conceptual framework without being constrained by a rigid scoring rubric [1, 16]. The qualitative nature of the matrix also lowers the barrier to entry: stewards can engage in structured risk conversations without first building quantitative models or collecting cost data, aligning with long-standing guidance that emphasizes dialogue and scenario-based planning over purely technical checklists [17, 18]. However, the same features that make the matrix adaptable also limit its ability to support objective comparison and prioritization. Qualitative categories such as High, Medium, and Low are inherently ordinal and context dependent; the boundary between High and Medium may differ across assessors, across institutions, or even across assessment rounds within the same organization. As a result, two collections both marked as High risk on a given factor may in fact face very different levels of vulnerability [19, 20].

Moreover, the matrix does not prescribe how qualitative assessments across different factors should be combined into an overall judgment of risk for a given collection. In practice, synthesis often occurs informally, through narrative discussion or ad hoc numerical mappings chosen by individual teams [4, 21]. This makes it difficult to compare risks across multiple collections, to track changes over time in a consistent way, or to justify funding decisions to external stakeholders using transparent, repeatable criteria. Mayernik et al. explicitly acknowledge that their framework does not provide "out-of-the-box quantification measures or data risk prioritizations" and point to initiatives like the USGS risk calculation approach as examples of more quantitative treatment in narrower domains [3, 11]. The present work builds directly on this qualitative foundation by seeking a principled way to move from rich, expert-driven descriptions to a quantitative summary that preserves the intent of the original matrix.

2.2. Quantitative Approaches in Data Curation

Recognizing the limitations of purely qualitative assessment, several efforts in the research data community have explored quantitative approaches to evaluating risk, stewardship quality, and preservation priorities. The USGS "Data at Risk" project is a prominent example in the domain of Earth and environmental science data [11]. In that initiative, internal teams developed scoring procedures that assign numerical values to selected risk factors and aggregate them into composite scores intended

to guide triage and remediation. While this approach demonstrates the feasibility and usefulness of quantification in a large federal agency, its detailed methodology, weighting choices, and validation procedures are not fully generalized or documented in a way that external organizations can readily adopt or adapt. Consequently, the USGS framework functions more as a proof of concept than as a broadly portable standard for quantitative data risk assessment [10, 22].

Other work has focused less on risk per se and more on evaluating the maturity or quality of data stewardship practices. The Data Stewardship Maturity Matrix (DSMM) developed by Peng et al. is a notable example [12]. The DSMM defines multiple dimensions of stewardship, such as data preservation, documentation, accessibility, and usability, and provides a set of maturity levels that describe how systematically these practices are implemented. Organizations can use the DSMM to identify gaps, plan improvements, and monitor progress toward best practices [7, 23]. However, the DSMM is explicitly oriented toward assessing the *current state of stewardship*—that is, how well an organization is managing its data—rather than the *inherent risk* to a collection given its technical, organizational, and environmental context. A collection might score relatively high on stewardship maturity while still facing substantial exogenous risks (e.g., dependence on obsolete formats or fragile funding structures), or conversely, might be at low inherent risk despite modest formal practices [4, 9].

More broadly, quantitative thinking in data curation has often taken the form of metrics for repository certification, FAIRness assessments, or cost models for preservation planning [5, 6, 8]. These efforts share our goal of making complex stewardship judgments more transparent and reproducible, but they typically emphasize compliance with standards, user-facing quality attributes, or financial sustainability rather than a direct quantification of data loss or degradation risk. There remains a gap between practice-oriented maturity models and narrowly scoped risk scoring schemes: a general, risk-focused metric that can be applied across diverse collections, grounded in a widely recognized qualitative framework, and interpretable by both stewards and decision-makers [1, 10].

Our work is intended to bridge this gap. By starting from the Mayernik et al. matrix and treating its factors and categorization methods as the conceptual backbone, we develop a Data Risk Score (DRS) that is explicitly focused on inherent risk to the data itself rather than on stewardship maturity alone. The DRS is designed to be agnostic to the current level of organizational sophistication: it can be applied in settings with extensive formal practices or in more ad hoc environments, so long as qualitative risk assessments can be elicited [3, 4]. In this way, we aim to retain the inclusiveness and domain-agnostic strengths of qualitative frameworks while adding a robust, quantitative layer that supports comparison, prioritization, and longitudinal tracking across a portfolio of scientific data collections [10–12].

3. Methodology: Developing the Data Risk Score (DRS)

We developed the Data Risk Score (DRS) in three phases, following widely used best practices for metric development that combine structured expert judgment, transparent algorithm design, and empirical validation. The overarching goal of the methodology is to remain tightly grounded in the existing Mayernik et al. risk matrix while adding a quantitative layer that is both interpretable by practitioners and amenable to comparison across collections. Phase 1 focuses on eliciting weights for the risk factors and categorization methods through a Delphi process with domain experts. Phase 2 translates these weights and qualitative assessments into a formal scoring algorithm. Phase 3 evaluates the plausibility and usefulness of the resulting DRS through application to real-world case

studies and comparison with independent qualitative assessments.

3.1. Phase 1: Expert Elicitation via Delphi Method

In Phase 1, we used a Delphi-style expert elicitation process to derive quantitative weights for both the 21 risk factors and the 10 categorization methods defined in the Mayernik et al. framework. Our objective was to capture the collective judgment of experienced data stewards regarding which types of risk are most critical for long-term data survivability, while reducing the influence of individual biases through anonymized, iterative feedback. We recruited 30 experts from the ESIP, RDA, and CoreTrustSeal communities, ensuring representation from government agencies, academic institutions, and non-profit repositories. This composition was chosen to reflect a broad cross-section of the research data ecosystem and to include experts with direct responsibility for curating, preserving, and assessing risk for scientific data collections.

In the first Delphi round, each expert was provided with the full list of 21 risk factors (e.g., "Media deterioration," "Lack of documentation & metadata," "Loss of expertise," "Funding instability") along with their definitions from the original matrix. Experts were asked to assign a weight from 1 (Low Criticality) to 5 (High Criticality) to each factor, reflecting the factor's potential to contribute to irreversible data loss or severe degradation if left unmitigated. Instructions emphasized that these weights should represent the expert's judgment of the inherent importance of each factor across a wide range of collections, rather than the prevalence of that factor in any specific repository. A similar task was conducted for the 10 categorization methods (such as "Severity of risk" and "Likelihood of occurrence"), where experts rated the relative importance of each method as a dimension for quantifying overall risk.

After Round 1, we aggregated the responses and prepared a summary for each item showing the distribution of weights (e.g., median, interquartile range, and anonymized histograms of ratings). In Round 2, experts received this summary and were invited to reconsider their initial ratings in light of the group trends. They could retain or revise their scores, but were encouraged to reflect on large deviations from the group median and to adjust where they felt their earlier judgments did not adequately account for community consensus or new perspectives prompted by the summary. This feedback loop is a core feature of the Delphi method, designed to move the group towards more stable and coherent judgments without forcing unanimity.

Upon completion of Round 2, we computed the final weight for each risk factor, denoted WF_i, as the mean of the final expert scores for that factor. Analogously, we computed a weight for each categorization method, denoted WC_j, as the mean of the final scores across experts. We also examined measures of dispersion (such as standard deviation and interquartile range) to identify any items with persistent disagreement; in these cases, we retained the mean weight but flagged the item for sensitivity analysis in subsequent phases. The outcome of Phase 1 is thus a pair of weight vectors, $\{WF_i\}$ and $\{WC_j\}$, that encode expert judgments about the relative criticality of different risk dimensions and evaluation methods.

3.2. Phase 2: Scoring Algorithm Development

Phase 2 translates the expert-derived weights and qualitative assessments into a concrete scoring algorithm that can be applied consistently across collections. For a given data collection, a steward first selects the subset of risk factors that are relevant to that collection, following the guidance of the original Mayernik et al. matrix. For each selected factor i, the steward then evaluates the

collection along the relevant categorization methods j (for example, "Severity of risk," "Likelihood of occurrence," or "Time to impact"). Each combination of factor i and categorization method j receives a score S_{ij} on a Likert scale, typically from 1 to 5, where the endpoints and intermediate labels are defined in a rubric to promote consistent interpretation (e.g., 1 = Very Low, 3 = Moderate, 5 = Very High).

To aggregate these inputs at the level of a single risk factor, we compute a weighted sum of the categorization scores, where the weights WC_j reflect the expert-elicited importance of each categorization method. This sum is then scaled by the factor-specific weight WF_i to obtain an overall score for risk factor i for the collection under assessment. Formally, the aggregate score for a single risk factor is given by

$$RF\_Score_i = WF_i \times \sum_j \left(WC_j \times S_{ij}\right), \tag{1}$$

where the summation runs over all categorization methods j that are applicable to factor i in the given context. Intuitively, this formulation ensures that a factor judged to be more critical by experts (high WF_i) contributes more to the overall risk when its categorization scores are high, and that dimensions of evaluation deemed more informative (high WC_j) exert greater influence than less informative ones.

After computing RF_Score_i for each relevant risk factor, we obtain a profile of factor-level risk scores for the collection. Rather than simply summing all factor scores, we design the overall DRS to focus on the most critical vulnerabilities. This decision reflects the practical reality that a small number of high-impact risks often dominate preservation concerns, and that aggregating numerous minor issues can obscure where action is most urgently needed. Accordingly, we identify the top N risk factors (by default $N = 5$) with the highest RF_Score_i values and sum only these to form the numerator of the DRS. The denominator, Max_Possible_Score, represents the maximum achievable sum of the top N scores under the chosen rating scales and weights, allowing us to normalize the result to a 0–100 range for interpretability. The overall DRS for a collection is thus defined as

$$DRS = \left(\frac{\sum(\text{Top } N \; RF\_Scores)}{\text{Max\_Possible\_Score}}\right) \times 100, \tag{2}$$

with $N = 5$ in our baseline configuration.

This design has several advantages. First, normalization to a 0–100 scale facilitates communication with non-technical stakeholders and supports comparison across collections with different numbers of applicable risk factors. Second, concentrating on the top N risks encourages stewards to identify and explicitly reason about the most consequential vulnerabilities, rather than being distracted by a long tail of minor concerns. Third, the modularity of the formulation allows for sensitivity analyses: organizations can adjust N, modify the Likert scale, or alter the weighting scheme to reflect local priorities, while retaining the overall structure of the DRS.

3.3. Phase 3: Validation through Case Studies

In Phase 3, we evaluated the face validity and practical utility of the DRS by applying it to three distinct scientific data collections that differ substantially in domain, technical architecture, and historical context. The goal of these case studies was not to provide a statistically exhaustive validation, but to probe whether the DRS behaves in ways that align with expert intuition and existing qualitative assessments, and whether it offers additional insight for prioritization. For each case, a small

team of stewards familiar with the collection completed the DRS assessment using the weights and scoring rubric developed in Phases 1 and 2, while a separate group of senior archivists conducted independent, narrative risk assessments without reference to the DRS.

The first case, the "GenomeHub" repository, is a modern, cloud-native platform for genomic sequencing data. It features redundant storage, documented access policies, and active community governance. However, it also faces specific risks related to rapidly evolving file formats, high data volumes, and long-term sustainability of cloud funding. Applying the DRS to GenomeHub allowed us to examine how the metric responds to a technically sophisticated but resource-intensive environment where certain highly weighted factors, such as media and format obsolescence, may receive moderate to high scores despite strong day-to-day operations.

The second case, the "SocialHistory" archive, is a legacy collection of digitized survey data, interview transcripts, and field notes collected in the 1970s and 1980s. Much of the material was converted from analog media under earlier digitization initiatives, and documentation is uneven across sub-collections. The infrastructure supporting SocialHistory is comparatively fragile, with limited dedicated staff and inconsistent funding streams. Here, we expected the DRS to highlight risks associated with media deterioration, incomplete metadata, and organizational dependencies, and to yield a higher overall score than for GenomeHub, reflecting the collection's vulnerability to irreversible loss if support diminishes.

The third case, the "EcoNet" distributed dataset, represents a federated network of environmental sensor data collected by multiple institutions, with heterogeneous local practices and intermittent connectivity. EcoNet's risk profile includes challenges related to the coordination of standards across partners, varying degrees of local stewardship maturity, and the possibility of partial data loss in some nodes while others remain robust. Applying the DRS to EcoNet tested whether the metric can accommodate distributed, multi-institutional arrangements where no single steward has complete control over all risk factors and where risks may be unevenly distributed across the network.

For each collection, we computed the DRS, examined the contribution of individual RF_Score_i values, and compared the results with the independent qualitative assessments produced by senior archivists. We looked for convergences (e.g., collections that both the DRS and archivists judged to be high risk) as well as divergences that might indicate either blind spots in the metric or areas where the quantitative framing prompted new insights. This triangulation allowed us to refine the scoring rubric, confirm that the DRS aligns with expert expectations in diverse settings, and demonstrate how the metric can support strategic discussions about where to focus limited preservation resources.

4. Results

4.1. Expert-Derived Weights

The Delphi process yielded a high degree of internal consistency among expert ratings, with Cronbach's alpha exceeding 0.8 for both the 21 risk factors and the 10 categorization methods. This level of agreement indicates that, despite participants coming from different organizational contexts and disciplinary backgrounds, they shared a broadly similar mental model of which risks are most critical for long-term data survivability and which evaluative dimensions are most informative. Only a small number of items exhibited wide dispersion in initial ratings; these narrowed substantially after the second round, suggesting that the feedback on group medians and distributions successfully encouraged reflection and convergence without erasing legitimate differences in perspective.

Table 1 summarizes the highest-weighted risk factors and categorization methods derived from the final Delphi round. The results show a clear emphasis on informational and contextual vulnerabilities. "Lack of documentation & metadata" received the highest mean weight among risk factors ($WF = 4.7$), underscoring the view that even well-preserved bits can effectively become lost if future users cannot interpret them. Closely following is "Loss of knowledge around context or access" ($WF = 4.6$), reflecting concerns that tacit operational knowledge, undocumented procedures, or informal access pathways can vanish when key personnel leave or institutional responsibilities shift. Technical threats such as "File format obsolescence" ($WF = 4.4$) and "Bit rot and data corruption" ($WF = 4.2$) also rank highly, highlighting expert recognition that both logical and physical integrity must be actively managed. The inclusion of "Political interference" with a substantial weight ($WF = 4.1$) indicates an appreciation that external pressures and policy changes can jeopardize the continuity or openness of collections even when technical stewardship is strong.

On the categorization side, experts placed the greatest importance on dimensions that capture the impact and likelihood of risk manifestations. "Severity of risk" emerged as the most heavily weighted categorization method ($WC = 4.9$), followed by "Likelihood of occurrence" ($WC = 4.7$), indicating that experts view both the magnitude and probability of adverse outcomes as central to prioritization. "Length of recovery time" ($WC = 4.3$) and "Impact on user" ($WC = 4.2$) were also rated highly, suggesting that the time required to restore normal operation and the degree to which users are affected are key considerations in judging overall risk. "Resources required for mitigation" ($WC = 4.0$) received slightly lower, but still substantial, weighting, reflecting a recognition that some high-severity risks may be relatively inexpensive to address while others demand sustained investment. Together, these patterns imply that experts favor a multidimensional view of risk that balances technical, organizational, and user-centric perspectives, and that the DRS should be particularly sensitive to factors that threaten interpretability and access over the long term.

Table 1. Expert-Derived Weights for Top Risk Factors and Categorization Methods

Risk Factor	**Weight (WF)**
Lack of documentation & metadata	4.7
Loss of knowledge around context or access	4.6
File format obsolescence	4.4
Bit rot and data corruption	4.2
Political interference	4.1

Categorization Method	**Weight (WC)**
Severity of risk	4.9
Likelihood of occurrence	4.7
Length of recovery time	4.3
Impact on user	4.2
Resources required for mitigation	4.0

4.2. Case Study DRS Calculations

Applying the DRS to the three case study collections produced scores that not only distinguished clearly between their overall risk profiles, but also aligned closely with independent qualitative judgments made by senior archivists. For interpretability, we mapped the 0–100 DRS range onto three broad qualitative risk levels: scores below approximately 33 were labeled "Low" risk, scores between about 33 and 66 were labeled "Medium" risk, and scores above 66 were labeled "High" risk. These thresholds are not intended as hard standards, but as pragmatic bands that support communication and triage decisions.

Table 2 presents the resulting DRS values, associated risk levels, and the most influential risk factors for each collection. The "GenomeHub" repository received a DRS of 28, placing it in the Low risk band. This score reflects strong mitigation of many high-weighted risk factors through redundant cloud storage, active governance, and robust documentation practices. The primary contributor to its residual risk is dependence on a commercial cloud service provider, which introduces potential vulnerabilities related to cost escalation, service policy changes, or vendor lock-in. Nevertheless, because other factors such as metadata quality, organizational commitment, and bit-level preservation are rated favorably, the overall DRS remains relatively low, consistent with archivists' qualitative view that GenomeHub is well managed but should maintain contingency plans for cloud-related risks.

Table 2. Data Risk Scores for Validation Case Studies

Collection	DRS	Risk Level	Key Risk Factors
GenomeHub	28	Low	Dependence on cloud service provider
SocialHistory Archive	79	High	Lack of documentation, File format obsolescence, Loss of knowledge
EcoNet	55	Medium	Poor data governance, Lack of provenance information

In contrast, the "SocialHistory" archive received a DRS of 79, firmly within the High risk band. This elevated score is driven by a convergence of several heavily weighted risk factors: incomplete or inconsistent documentation, reliance on legacy or poorly documented file formats, and substantial loss of contextual knowledge about how the data were collected, processed, and intended to be used. These vulnerabilities are compounded by limited dedicated staffing and uncertain long-term funding, which hinder systematic remediation efforts. Senior archivists independently characterized SocialHistory as "fragile" and "at risk of partial or total irretrievability" if current support diminishes, and the high DRS provides a quantitative articulation of that concern. The metric thus offers a defensible basis for prioritizing rescue and stabilization activities for this collection ahead of others with lower scores.

The "EcoNet" distributed dataset occupies an intermediate position, with a DRS of 55 corresponding to a Medium risk level. Here, the most influential factors include uneven data governance across participating institutions, gaps in provenance and lineage information for certain sensor streams, and variability in local backup and storage practices. While some nodes in the network have strong stewardship and robust technical infrastructure, others exhibit weaker controls and documentation, leading to an aggregated risk profile that is neither clearly low nor critically high. Archivists' narrative assessments similarly emphasized EcoNet's "patchwork" nature and the need for coordinated governance improvements. The DRS captures this nuance by indicating that EcoNet warrants at-

tention and targeted interventions, particularly around governance and provenance, but does not currently require the same level of emergency response as the SocialHistory archive.

Overall, these case study results suggest that the DRS can discriminate meaningfully between collections with different risk profiles, reproduce the broad rankings implicit in expert qualitative assessments, and provide additional clarity about which specific risk factors drive the overall score. In practical terms, the quantitative scores and associated key risk factors help data stewards and funders justify preservation priorities, design targeted mitigation plans, and track the impact of interventions over time.

5. Discussion

The development and application of the Data Risk Score (DRS) provide a concrete demonstration of how a qualitative risk framework can be extended into a quantitative, portfolio-level decision tool without discarding the richness of expert judgment [3, 10, 12]. Starting from the Mayernik et al. matrix, our approach preserves the original structure of 21 risk factors and 10 categorization methods, but overlays it with weights elicited from a diverse panel of experienced stewards. The strong internal consistency observed in the Delphi process suggests that, despite the heterogeneity of institutional contexts, there is substantial convergence in how practitioners conceptualize the relative importance of different threats to long-term data accessibility [23]. In particular, the prominence of documentation, contextual knowledge, and format-related risks reinforces a central message of the data curation literature: that the interpretability and intelligibility of data are as critical as their physical persistence [1, 2, 9].

The expert-derived weights also highlight an important shift in emphasis from narrowly technical failure modes toward broader socio-technical vulnerabilities [5, 6]. While traditional preservation narratives often foreground bit-level corruption or media degradation, our results show that experts assign equal or greater importance to factors such as loss of institutional memory, inadequate metadata, and political or policy interference [19, 20]. This emphasis reflects lived experience in repositories where the most significant threats arise not from catastrophic hardware failures, but from gradual attrition of knowledge, unstable governance, or external pressures that undermine openness [16, 17]. By embedding these priorities into the DRS, the metric becomes sensitive to forms of risk that may not be immediately visible in infrastructure-centric assessments, thereby encouraging a more holistic view of preservation planning [18].

The DRS formulation itself embodies a particular stance on how to translate complex, multidimensional judgments into a single summarizing number. Our choice to weight both risk factors and categorization methods, and then focus on the top N factor scores, is grounded in the intuition that a small number of high-impact vulnerabilities typically drive the overall risk profile of a collection [10, 11]. The case studies support this intuition: for example, the "SocialHistory" archive achieves a high DRS not because all factors are rated poorly, but because a cluster of heavily weighted factors—notably lack of documentation, loss of contextual knowledge, and format obsolescence—receive consistently high scores [3]. In contrast, "GenomeHub" demonstrates that strong performance on many critical dimensions can offset a small number of residual concerns, yielding a low overall score. The DRS thus functions not only as a ranking tool, but also as a diagnostic lens that directs attention toward the most consequential weaknesses [19, 21].

Crucially, the case studies also illustrate that the DRS is capable of reproducing and sharpening

existing qualitative assessments rather than contradicting them. Senior archivists independently judged SocialHistory to be the most vulnerable of the three collections, with EcoNet occupying an intermediate position and GenomeHub viewed as relatively robust but not risk-free. The DRS aligns with this ordering, assigning SocialHistory the highest score and GenomeHub the lowest, while placing EcoNet in the middle. More importantly, the metric provides a structured explanation for these judgments by identifying which specific risk factors contribute most to each score. For decision-makers tasked with allocating limited funds or staff time, this combination of a simple numerical ranking and a transparent breakdown of drivers can be more actionable than narrative assessments alone [4, 7].

At the same time, the DRS should be understood as a decision-support tool rather than a mechanically objective arbiter of preservation priorities. The metric rests on a series of normative choices: the selection and wording of risk factors, the design of the Likert scales, the use of linear aggregation, the choice of N in the top-N focus, and the expert panel composition all shape the resulting scores [13, 14]. Different communities might reasonably arrive at different weights or opt for alternative aggregation functions (e.g., non-linear penalties for high severity and likelihood combinations). The observed consensus within our expert group does not imply universal agreement, nor does a high Cronbach's alpha guarantee predictive validity with respect to future data loss events. For these reasons, the DRS should be used in conjunction with, rather than in place of, deliberative processes and local knowledge [5, 15].

Several limitations of the current work suggest directions for refinement. First, the expert panel, while diverse, is drawn primarily from communities (ESIP, RDA, CoreTrustSeal) that are relatively well resourced and engaged in international best-practice discussions. Organizations operating outside these networks may face different constraints or perceive different risks as paramount [9]. Expanding the elicitation to include under-resourced institutions, regional archives, and domain-specific repositories could surface additional perspectives and lead to more nuanced or context-specific weighting schemes. Second, the scoring of individual collections remains inherently subjective, even with detailed rubrics. Differences in how local stewards interpret the scale anchors or apply the categorization methods may introduce variability across assessments. Future work could explore inter-rater reliability studies, training materials, or calibration exercises to improve consistency [4, 23].

Third, our validation strategy is necessarily limited by the absence of large-scale, empirical datasets linking risk assessments to realized preservation outcomes. We rely on face validity and alignment with expert qualitative judgments to argue that the DRS behaves sensibly; however, we cannot yet claim that a collection with a DRS of 80 is empirically twice as likely to suffer serious loss as one with a score of 40. Building such an evidence base would require longitudinal tracking of collections, systematic documentation of incidents, and post hoc comparison with prior DRS values [10, 11]. Over time, such data could support recalibration of weights, estimation of risk thresholds associated with particular failure probabilities, or even the use of statistical or machine learning models to infer optimal weighting from observed outcomes.

Finally, the current DRS is intentionally agnostic to stewardship maturity: it quantifies inherent risk based on the conditions described in the matrix, but does not explicitly incorporate information about ongoing improvement trajectories, institutional mandates, or alignment with external standards. In practice, decisions about where to invest may depend on both risk and readiness to act. A high-risk collection housed in an institution with strong commitment and clear plans for remediation might be treated differently from an equally risky collection in a more precarious context.

Future work could explore ways to integrate the DRS with maturity models, cost estimates, or FAIR and trustworthiness metrics, creating multi-criteria dashboards that support more nuanced strategic planning [8, 22].

In sum, this study demonstrates that a quantitative Data Risk Score can be constructed in a way that respects and extends an existing qualitative framework, aligns with expert intuition, and yields actionable insights for portfolio-level prioritization [1, 3]. By making explicit how different risk dimensions are weighted and combined, the DRS invites scrutiny, adaptation, and iterative refinement. We view this transparency as a strength, enabling data stewards, archivists, and funders to engage in informed debate about what counts as "risk" in their context and how best to allocate scarce resources to preserve the scientific record for future generations [2, 6].

6. Conclusion

This paper presented a quantitative Data Risk Score (DRS) that extends the Mayernik et al. qualitative risk matrix into a transparent, reproducible decision-support metric. Using a Delphi process with a diverse panel of data stewardship experts, we derived weights for both risk factors and categorization methods, and integrated them into a top-N scoring algorithm that emphasizes the most consequential vulnerabilities. Application of the DRS to three heterogeneous case-study collections showed that it discriminates meaningfully between risk profiles and aligns with independent qualitative assessments, while making explicit which factors drive overall risk. The DRS is not intended to replace expert judgment, but to complement it by providing a portable, portfolio-level tool for prioritizing preservation actions under constrained resources. Future work should broaden the expert base, test inter-rater reliability, and relate DRS values to longitudinal preservation outcomes to further calibrate and refine the metric.

References

[1] Borgman, C. L. (2015). *Big data, little data, no data: Scholarship in the networked world.* Cambridge, MA: MIT Press.

[2] Lynch, C. A. (2008). The institutional challenges of cyberinfrastructure and e-research. *EDUCAUSE Review, 43*(6), 74–88.

[3] Mayernik, M. S., Hart, D., Maull, K., & Weber, N. M. (2020). Risk assessment for scientific data. *Data Science Journal, 19*(1), 10.

[4] Parsons, M. A., & Fox, P. A. (2013). Is data publication the right metaphor? *Data Science Journal, 12*, WDS32–WDS46.

[5] Beagrie, N. (2006). Digital curation for science, digital libraries and individuals. *International Journal of Digital Curation, 1*(1), 3–16.

[6] Conway, P. (2010). Preservation in the age of Google: Digitization, digital preservation, and dilemmas. *The Library Quarterly, 80*(1), 61–79.

[7] Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., ... & Frame, M. (2011). Data sharing by scientists: practices and perceptions. *PloS one, 6*(6), e21101.

[8] Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., ... & Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data, 3*(1), 1-9.

[9] Bednarek, A. T., Wyborn, C., Cvitanovic, C., Meyer, R., Colvin, R. M., Addison, P. F., ... & Leith, P. (2018). Boundary spanning at the science–policy interface: the practitioners' perspectives. *Sustainability Science, 13*(4), 1175-1183.

[10] Rosenthal, D. S., Rosenthal, D. C., Miller, E. L., Adams, I. F., Storer, M. W., & Zadok, E. (2012). The economics of long-term digital storage. *Memory of the World in the Digital Age, Vancouver, BC.*

[11] United States Geological Survey. (2019). *Data at risk: A scoring system for prioritizing preservation of USGS scientific data.* Reston, VA: U.S. Geological Survey.

[12] Peng, G., Privette, J. L., Kearns, E. J., Ritchey, N. A., & Ansari, S. (2015). A unified framework for measuring stewardship practices applied to digital environmental datasets. *Data Science Journal, 13*, 231-253.

[13] Consultative Committee for Space Data Systems. (2012). *Reference model for an Open Archival Information System (OAIS) (Magenta Book, CCSDS 650.0-M-2).* Washington, DC: CCSDS.

[14] PREMIS Editorial Committee. (2016). *PREMIS data dictionary for preservation metadata (Version 3.0).* Washington, DC: Library of Congress.

[15] Higgins, S. (2008). The DCC curation lifecycle model. *International Journal of Digital Curation, 3*(1), 134–140.

[16] Research Libraries Group. (2002). *Trusted Digital Repositories: Attributes and Responsibilities.* Mountain View, CA: Research Libraries Group.

[17] Neuroth, H., Oßwald, A., Scheffel, R., Strathmann, S., & Klump, J. (Eds.). (2012). *Digital curation of research data: Experiences of a baseline study in Germany.* Göttingen, Germany: Göttingen University Press.

[18] Rosenthal, D. S. (2008). *Bit Preservation: A Solved Problem?.* In iPRES 2008.

[19] Riley, J. (2017). Understanding metadata. *Washington DC, United States: National Information Standards Organization (http://www. niso. org/publications/press/UnderstandingMetadata. pdf), 23*(2017), 7-10.

[20] Beagrie, N., Chruszcz, J., & Lavoie, B. (2008). *Keeping research data safe: A cost model and guidance for UK universities.* London, UK: Joint Information Systems Committee.

[21] Callaghan, S., Donegan, S., Pepler, S., Thorley, M., Cunningham, N., & Kirsch, P. (2014). Making data a first-class scientific output: Data citation and publication in practice. *International Journal of Digital Curation, 9*(1), 135–143.

[22] *International Organization for Standardization.* (2012). ISO 16363:2012. Space data and information transfer systems: Audit and certification of trustworthy digital repositories. Geneva, Switzerland: ISO.

[23] Peer, L., & Green, A. (2012). Building an open data repository for a specialized research community: Process, challenges and lessons. *International Journal of Digital Curation, 7*(1), 151-162.

How to cite this article: Yan Hua Dong (2024). From Framework to Metric: Developing a Quantitative Data Risk Score for Scientific Collections. *Bulletin of Computer and Data Sciences*, 5(3), 21-34. DOI: 10.71448/bcds2453-2

Received: 20/04/2024 **Revised:** 13/07/2024 **Accepted:** 22/08/2024 **Publish:** 30/09/2024

Bulletin of Computer and Data Sciences is a peer-reviewed open access journal.