Quantifying the technical debt of legacy cataloging and the case for automated quality control

Sadoqat Raimjonova
Scientific advisor: Rashid Turgunbayev
Kokand State University

Abstract: *The concept of technical debt, borrowed from software engineering, offers a powerful analytical lens through which to examine the accumulated costs of suboptimal cataloging decisions made over decades of library practice. While library literature has extensively addressed the challenges of database cleanup and retrospective conversion, these efforts have typically been framed as finite projects with clear endpoints rather than as manifestations of an ongoing liability that accrues interest over time. This article argues that legacy cataloging data constitutes a form of technical debt that imposes tangible costs on discovery, systems performance, staff productivity, and user satisfaction. Drawing upon case studies from academic and research libraries, the analysis demonstrates that traditional approaches to quality control, which rely heavily on manual review and expert judgment, are structurally inadequate to address the scale and complexity of contemporary metadata problems. The argument advances a compelling case for automated quality control as an essential component of sustainable metadata management, not as a replacement for professional expertise but as a force multiplier that enables catalogers to focus their attention on high-value interventions. Automated tools, including rule-based validation, machine learning applications, and entity resolution algorithms, offer the capacity to identify, prioritize, and remediate metadata deficiencies at a scale and speed that manual processes cannot match. The article concludes with a framework for implementing automated quality control programs that balance technological capability with professional judgment, institutional capacity with aspirational standards, and short-term remediation with long-term prevention. Ultimately, the case for automation rests not on the elimination of human expertise but on its strategic redeployment toward the most intellectually demanding and professionally rewarding aspects of metadata stewardship.*
Keywords: *technical debt, cataloging quality, automated quality control, metadata remediation, database cleanup, machine learning in libraries, cataloging workflows, metadata management, legacy data, quality assurance*

The language of finance has an uncomfortable habit of intruding upon professional domains that prefer to frame their activities in terms of service, stewardship, and intellectual contribution. Yet there are moments when economic metaphors prove so unexpectedly apt that they illuminate aspects of practice that more genteel vocabularies obscure. The concept of technical debt, which emerged from the software development community in the early nineteen nineties, describes the eventual consequences of choosing expedient solutions over more robust ones, with the understanding that such shortcuts incur future costs that must eventually be paid. When a software team opts to release a product with known bugs, incomplete documentation, or suboptimal architecture, they are incurring technical debt that will manifest as increased maintenance effort, reduced flexibility, and diminished performance in subsequent development cycles. This concept translates with remarkable fidelity to the domain of library cataloging, where decades of inconsistent practice, shifting standards, legacy system constraints, and pragmatic compromises have produced metadata that functions adequately for many purposes but carries significant hidden costs. The recognition of legacy cataloging as a form

of technical debt offers librarianship a framework for quantifying these costs, prioritizing remediation efforts, and making the case for investments in automated quality control that might otherwise appear difficult to justify.

Understanding the nature of cataloging technical debt requires a clear-eyed examination of how such debt accumulates in practice. The MARC format, for all its virtues of stability and interoperability, has never been applied with perfect consistency across the library community. Catalogers have made locally appropriate decisions about the depth of description, the choice of access points, the application of subject headings, and the interpretation of complex rules that have resulted in substantial variation even among records describing identical resources. These variations were often entirely reasonable given the circumstances of their creation, constrained by the tools available, the time allotted, the cataloger's level of experience, and the prevailing standards of the era. Yet each such decision, however justifiable at the moment of its making, contributed to a pool of metadata that would become increasingly difficult to manage as libraries migrated to new systems, adopted new discovery layers, and began to contemplate the transition to linked data environments. The debt accumulated not through malfeasance or negligence but through the ordinary operation of professional practice under less than ideal conditions, a pattern that mirrors the accumulation of technical debt in software development, where deadlines, resource constraints, and evolving requirements ensure that compromise is the rule rather than the exception.

The costs of this accumulated debt manifest across multiple dimensions of library operations, some of which are readily quantifiable while others remain stubbornly subjective yet no less consequential. The most immediate cost is borne by library users, who encounter inconsistent metadata in their discovery interactions and may fail to retrieve relevant materials, retrieve irrelevant materials, or abandon their searches in frustration when the system fails to respond to their queries as expected. Studies of catalog usability consistently demonstrate that poor metadata quality negatively affects recall and precision, with particular impacts on novice users who lack the search strategies necessary to compensate for inadequate indexing. These discovery failures represent a direct cost to the library's mission of connecting users with information, yet they are rarely tracked systematically or factored into assessments of cataloging productivity. The library that measures success by the number of records processed rather than by the quality of user outcomes is effectively ignoring the most significant consequence of its technical debt.

Staff productivity constitutes a second major cost dimension of cataloging technical debt, one that is more readily measurable but often underestimated in its cumulative impact. Catalogers and metadata specialists spend a considerable portion of their time not creating new records but correcting, enhancing, and reconciling existing ones. This maintenance work is essential to the ongoing functioning of the catalog, yet it represents a diversion of professional expertise from more intellectually demanding and strategically valuable activities. Each hour spent cleaning up inconsistent subject headings, resolving duplicate authority records, or correcting obsolete form subdivisions is an hour not spent on original cataloging, metadata consultation with research faculty, or experimentation with new data models. The interest on cataloging technical debt is paid in the currency of professional time, and the compounding effects are substantial, particularly in institutions where staffing levels have remained static while the volume and complexity of digital resources have grown exponentially. Libraries that have not systematically assessed the proportion of cataloging effort devoted to remediation are likely to be surprised by the findings, and even more surprised by the difficulty of reducing this proportion without strategic investments in automated tools.

The migration of library systems provides a particularly vivid demonstration of the costs associated with cataloging technical debt, as the transfer of data from one platform to another invariably exposes

inconsistencies and errors that were previously invisible or manageable within the native environment. Libraries that have undertaken major system migrations routinely report that the most labor-intensive phase of the project is not the technical installation of the new software but the preparation and cleanup of the data to be migrated. This cleanup work is often compressed into unrealistic timeframes, forcing staff to make rapid decisions about which problems to address immediately and which to defer for later correction, thereby incurring new technical debt even as they attempt to reduce old debt. The cycle is depressingly familiar to anyone who has participated in multiple migrations over a career, with each transition revealing patterns of data degradation that were created in previous transitions and never fully resolved. The financial costs of these migration-related cleanup efforts are substantial, often consuming a significant portion of project budgets and diverting resources from training, customization, and other activities that would enhance the value of the new system.

The emergence of linked data as the anticipated future of library metadata adds urgency to the quantification of cataloging technical debt, as the transition from MARC to semantic web technologies demands a level of data consistency and structural clarity that many legacy catalogs cannot provide. Linked data systems rely upon unambiguous entity identification, consistent relationship modeling, and clean data values that can be reliably mapped to external authorities. Records that contain conflicting data, ambiguous encoding, or non-standard vocabularies pose significant challenges for automated conversion processes and require extensive manual intervention to be made usable in a linked data context. The technical debt that was manageable within the forgiving environment of MARC becomes a serious impediment to progress when libraries attempt to participate in the global web of linked data. The costs of remediation in this context are not merely operational but strategic, as libraries that cannot produce clean linked data risk being excluded from the collaborative networks and discovery services that will increasingly define the library's role in the information landscape.

The argument for automated quality control rests upon a recognition that the scale and complexity of contemporary cataloging problems have outstripped the capacity of traditional manual approaches. The largest research libraries hold millions of bibliographic records, each containing dozens or hundreds of data elements, and the relationships among these elements are numerous and intricate. Human catalogers, even the most skilled and experienced, cannot possibly review every record in the collection, identify every inconsistency or error, and prioritize interventions in a way that optimizes overall quality improvement. Automated tools, however, can process entire databases in hours or days, applying rule-based validation criteria, statistical analysis, and pattern recognition algorithms to identify problems that would take human reviewers years to discover. The advantage of automation lies not in superior judgment but in superior coverage and speed, enabling libraries to know with confidence the extent and nature of their metadata problems rather than operating based on intuition and anecdotal evidence.

The development of automated quality control tools for library metadata has advanced considerably in recent years, with a range of approaches available to institutions at varying levels of technical sophistication. Rule-based validation remains the most widely accessible approach, with tools such as MARCEdit, OpenRefine, and various vendor-supplied utilities enabling libraries to check records against defined criteria and generate reports of non-conforming instances. These tools are relatively easy to implement and can produce immediate improvements in data consistency, particularly when applied to routine problems such as invalid subfield codes, obsolete fixed field values, and inconsistent punctuation. Their primary limitation is their dependence upon explicit rules that must be defined in advance, which means they cannot identify novel problems or adapt to evolving

standards without manual reprogramming. For many libraries, however, rule-based validation offers a low-risk entry point into automated quality control that builds institutional confidence and demonstrates tangible results.

Machine learning approaches represent a more advanced category of automated quality control, offering the capacity to identify patterns that are too subtle or complex for rule-based systems to detect. Supervised learning techniques, in which algorithms are trained on manually classified examples, have been applied successfully to tasks such as deduplication, subject heading assignment, and classification prediction. Unsupervised learning techniques, which identify patterns in data without pre-existing labels, have shown promise for anomaly detection, clustering of similar records, and discovery of implicit relationships within metadata. The application of machine learning to cataloging quality is still in its early stages, with many libraries lacking the data science expertise or computational infrastructure necessary for implementation. Nevertheless, the potential of these techniques is substantial, particularly as pretrained models and cloud-based services become more accessible to institutions that cannot develop their own solutions from scratch. The library community would benefit from greater collaboration in this area, sharing training data, algorithms, and best practices in ways that reduce the barriers to entry for less-resourced institutions.

Entity resolution algorithms constitute a third category of automated quality control that is particularly relevant to the linked data transition, as these tools enable libraries to identify and reconcile variant representations of the same entities across their catalogs. Name authority work, which has traditionally been a labor-intensive process requiring specialized expertise, can be substantially accelerated through algorithmic approaches that generate candidate matches for human review. The most sophisticated entity resolution systems incorporate features such as fuzzy matching, contextual disambiguation, and iterative refinement, reducing the number of false positives that require manual filtering. For libraries with substantial backlogs of uncataloged materials, entity resolution can also facilitate the automatic assignment of access points based on data extracted from the resources themselves or from external knowledge bases. The time savings achievable through these techniques are substantial, potentially reducing the effort required for authority work by orders of magnitude and freeing catalogers to focus on more complex and intellectually rewarding aspects of their responsibilities.

The implementation of automated quality control programs requires careful attention to the relationship between technological capability and professional judgment, as the goal is not to replace catalogers but to enable them to work more effectively. Automated tools excel at identifying problems that are well-defined and predictable, but they cannot substitute for human expertise in ambiguous cases, nor can they determine which problems are most important to address given local priorities and resources. The optimal workflow is therefore one in which automation generates a prioritized list of potential interventions that are then reviewed and acted upon by human catalogers, with the most straightforward cases handled automatically and the more complex cases delegated to professional judgment. This hybrid model preserves the essential role of cataloging expertise while dramatically increasing the scale and efficiency of quality control activities. Libraries that have implemented such workflows consistently report that their catalogers are initially skeptical but rapidly become enthusiastic as they realize how much time is saved and how much more satisfying their work becomes when they are relieved of repetitive tasks.

Sustaining quality improvement over the long term requires not only the implementation of automated tools but also the development of organizational processes that prevent new technical debt from accumulating. The most sophisticated quality control program will be undermined if the workflows that generate new records continue to produce the same problems that require remediation. Libraries

must therefore integrate automated quality control into their production workflows, validating new records at the point of creation or import and rejecting those that do not meet defined standards. This preventive approach is considerably more efficient than retrospective remediation, as it addresses problems before they can propagate across the catalog and creates incentives for catalogers to produce clean data from the start. The implementation of preventive controls does not eliminate the need for retrospective cleanup of legacy data, but it does ensure that the cleaned data remains clean, preventing the frustrating cycle of repeated remediation that has characterized so many cataloging improvement efforts.

The business case for automated quality control ultimately rests upon a calculation of costs and benefits that libraries have been slow to perform with rigor. The costs of automation are relatively straightforward to estimate, encompassing software acquisition or development, staff training, implementation support, and ongoing maintenance. The benefits are more diffuse and difficult to quantify, including improved discovery outcomes, enhanced staff productivity, reduced migration costs, and greater readiness for linked data participation. Libraries that have attempted to perform these calculations have generally found that the return on investment for automated quality control is substantial, particularly when the full range of benefits is considered over a multi-year timeframe. The difficulty lies not in the arithmetic but in the institutional willingness to invest in quality improvement when there is no crisis demanding immediate action. The concept of technical debt provides a useful rhetorical frame for these discussions, as it makes visible the costs of inaction and frames quality improvement not as an unbudgeted expense but as a strategic investment that reduces future liabilities.

The future of cataloging quality control will likely involve increasingly sophisticated automation, with artificial intelligence techniques becoming more accessible and more capable over time. Libraries that have invested early in automated quality control will be better positioned to adopt these emerging techniques, having built the institutional infrastructure and staff expertise necessary for their effective deployment. Conversely, libraries that have deferred quality improvement may find themselves at a growing disadvantage, with their technical debt compounding as the demands of linked data and other advanced discovery services outstrip the capacity of their legacy data. The time to act is therefore now, not because the current situation is unsustainable in an absolute sense, but because the window of opportunity for relatively low-cost intervention is closing as the complexity of the library data ecosystem continues to increase. The catalogers of the future will still need professional judgment, intellectual curiosity, and commitment to service, but they will increasingly rely upon automated tools to amplify their effectiveness and focus their expertise where it matters most. The libraries that support their catalogers with these tools will be the ones that thrive in the linked data future, while those that persist with manual methods alone will find themselves unable to meet the demands of an increasingly data-intensive information environment.

References

1. To‘Ychiyeva, S. (2023). O ‘ZBEKISTONDA ELEKTRON KUTUBXONALAR VA XALQARO HAMKORLIKNING RIVOJLANISH SAMARALARI. Oriental Art and Culture, 4(6), 38-43.
2. Yuldasheva, S. N. (2022). Innovative Ways to Improve the Efficiency of Trade Policy in Uzbekistan. Galaxy International Interdisciplinary Research Journal, 10(5), 505-508.
3. Yuldasheva, S. (2023). The Experience of Learning to Read from Foreign Countries. European Journal of Innovation in Nonformal Education, 3(10), 42-46.
4. Yuldasheva, S. (2025). BIBLIOGRAFIYA FANNI O ‘QITISH METODIKASINING MAQSAD VA VAZIFALARI. World of Philology, 4(4), 37-47.

5. Туйчиева, Д. (2022). О ЛИТЕРАТУРНОЙ ЭТИКЕ АМИРА ТЕМУРА. Oriental Art and Culture, 3(4), 810-816.
6. To‘Ychiyeva, D. (2023). AXBOROT-KUTUBXONA MUASSASALARIDA REKLAMANING O ‘ZIGA XOS XUSUSIYATLARI VA ULARNING TURLARI. Oriental Art and Culture, 4(1), 802-807.
7. Yo’Ldasheva, S. (2022). KITOBNING IJTIMOIY-PSIXOLOGIK XUSUSIYATLARI. Oriental Art and Culture, 3(4), 413-417.
8. Manzirova, M., & Tuychiyeva, S. (2023). KUTUBXONACHI KUTUBXONA-AXBOROT JARAYONINING YETAKCHI ISHTIROKCHISI SIFATIDA. Oriental Art and Culture, 4(1), 62-65.
9. No‘Monovna, T. Y. S. (2022). AXBOROT-KURUBXONA MARKAZI XODIMLARI FAOLIYATIDA NUTQ MADANIYATI VA NOTIQLIK SAN’ATINING AHAMIYATI. Oriental Art and Culture, 3(1), 589-597.
10. Dildoraxon, T. Y. (2023). BIBLIOTERAPIYA KITOBLAR VOSITASIDA DAVOLASH USULI. Oriental Art and Culture, 4(6), 90-94.