10.48047/jocaaa.2026.35.03.38

Automated Product Catalog Onboarding at Scale: A Confidence-Driven ASIN Creation Platform with Machine-Learning-Based Title Validation

Anil Mankali Masakal
Dropbox Inc, USA

Abstract

Large-scale e-commerce platforms face a persistent challenge in onboarding high-quality product data at scale. As marketplaces grow to hundreds of millions of listings, manual catalog review processes become increasingly costly, inconsistent, and slow, directly impacting seller onboarding speed, catalog quality, and customer discovery. This article presents a production-scale automation platform implemented within Amazon Retail Business Services (RBS) between 2012 and 2014 that transformed ASIN (Amazon Standard Identification Number) creation from a human-first workflow into a confidence-driven, automation-first system. The platform combined rule-based validation, statistical confidence scoring, and continuous audit feedback loops to safely automate a significant portion of ASIN creation workflows. In parallel, the work introduced a machine-learning-based approach to ASIN Title validation, reframing subjective human edits as a supervised classification problem. Together, these contributions demonstrate how large-scale catalog systems can achieve step-function gains in efficiency and quality while preserving governance and trust. The ideas presented anticipated a broader industry shift toward ML-assisted catalog onboarding now visible across major global marketplaces.

Keywords: Asin Creation, Catalog Onboarding, Confidence-Driven Automation, Machine Learning, Title Validation

1. Introduction

1.1 Motivation and Scope

Product catalogs are foundational infrastructure for digital commerce. Every search result, recommendation, pricing decision, and fulfillment action depends on the accuracy and consistency of underlying catalog data. For large marketplaces, catalog onboarding involves ingesting seller-submitted data, validating attributes such as titles, brands, variations, and dimensions, and ensuring compliance with platform standards before products become discoverable. The size of modern platforms makes it harder and harder to keep an eye on these processes by hand, leading to potential errors and inconsistencies in the catalog data that can negatively impact user experience and sales. Instead, they need to be automated with measurable quality signals [1][2].

1.2 Limitations of Manual Review

Historically, many marketplaces—including Amazon during the early 2010s—relied on large teams of human associates to manually review and correct catalog submissions. While effective at smaller scales, this approach introduces several structural limitations: manual review does not scale linearly with marketplace growth, quality decisions vary across reviewers, onboarding latency delays time-to-revenue, and operational costs grow proportionally with submission volume. These constraints become acute as seller ecosystems expand globally and product variation complexity increases. Addressing these challenges necessitated a fundamental rethinking of the processes involved in making and governing catalog quality decisions [1][2].

10.48047/jocaaa.2026.35.03.38

1.3 Contribution of This Work

This paper documents an automation platform developed within Amazon Retail Business Services that addressed these challenges by introducing confidence-based automation for ASIN creation and machine-learning-driven title quality validation. The work reframed catalog onboarding as an engineering and data problem rather than a purely operational one, enabling Amazon to scale catalog growth without proportional growth in human review capacity. The contributions described here are an important early example of using machine learning to automate operations in a crucial part of the commerce process.

2. Background and Industry Context

2.1 Catalog Onboarding as an Industry-Wide Challenge

All major e-commerce platforms must solve the same core problem: how to ingest heterogeneous, often noisy seller data and convert it into a clean, structured catalog suitable for search and discovery. Public documentation from large marketplaces highlights the complexity of listing requirements, particularly for variation products such as apparel with size and color dimensions. Even minor inconsistencies in titles or variation grouping can degrade search relevance and customer trust. The challenge is compounded by the diversity of seller data formats, languages, and product categories that platforms must accommodate simultaneously [3][4], which can lead to increased errors in product listings and further complicate the user experience for both sellers and buyers.

Platform / Approach	Validation Method	Handling of Variation Products	Human Review Dependency	Quality Consistency
Traditional Rule-Based Systems	Static field validation	Manual grouping	High	Variable across reviewers
PIM-Based Platforms	Schema and attribute checks	Template-driven	High	Moderate
Amazon RBS (Pre-Automation)	Rule-based with manual review	Human-verified grouping	High	Inconsistent at scale
Amazon RBS (Auto ASIN Platform)	Confidence-driven automation	Automated validation logic	Low (exception-based)	Standardized and auditable
Modern ML-Assisted Platforms	ML-gated with rule overlay	Semi-automated	Moderate	High and reproducible

Table 1: Comparison of Catalog Onboarding Approaches Across Major E-Commerce Platforms [1, 3, 4, 5]

2.2 State of ML in E-Commerce Prior to 2014

During the 2010–2014 timeframe, most machine-learning investment in e-commerce focused on search ranking, recommendations, and demand forecasting. Catalog onboarding workflows, by contrast, remained largely rule-driven and human-intensive. Automation efforts typically focused on validating required fields rather than replacing human judgment in quality-critical steps, such as assessing the

10.48047/jocaaa.2026.35.03.38

relevance and accuracy of product descriptions and images. The application of ML to operational gating decisions—such as whether a product listing should be auto-approved or escalated for human review—was largely unexplored at production scale during this period [3][4].

2.3 Amazon Retail Business Services and ASIN Creation

Within Amazon, ASIN creation represents a core transactional workflow: every product must pass through this pipeline before it can be sold. Sellers submit product information through vendor- and seller-facing systems, after which RBS associates historically reviewed submissions, corrected errors, edited titles, and approved listings. The scale of this operation—supporting millions of sellers and hundreds of millions of customers—made ASIN creation an ideal candidate for systemic automation. The platform described in this paper emerged from the need to solve this problem at a scale that manual workflows could not sustainably support, particularly in automating the ASIN creation process to enhance efficiency and accuracy in product listings.

3. Problem Statement

3.1 Core Automation Challenge

This work tackled the central problem of safely reducing human involvement in ASIN creation without compromising catalog quality. To solve this, we needed to address three connected questions: how to ensure that automation can reliably make quality decisions that experienced human reviewers usually make, how to turn personal opinions, like whether a product title seems right, into clear and consistent measures, and how to make sure automation can handle varying amounts of submissions while still being controlled and able to be checked. Solving these problems required moving beyond static rules toward a system capable of reasoning about confidence and learning from historical human decisions [5][6].

3.2 Operationalizing Subjective Quality

A key challenge in automating catalog quality decisions is the inherently subjective nature of product data quality. For instance, keyword rules or schema validation cannot fully capture title correctness as a binary property. It involves stylistic conventions, category-specific norms, and contextual judgment that vary across product types and reviewers. Changing this subjective issue into something that can be learned meant looking at it in a statistical way, using past editing choices made by people as a standard for quality, and creating models that could apply those patterns to new submissions.

4. System Architecture: Auto ASIN Creation Platform

4.1 Automation-First, Confidence-Driven Design

The Auto ASIN Creation platform was designed as an event-driven workflow triggered by seller submissions. Instead of treating human review as mandatory, the system evaluated each submission using a combination of rule-based checks and confidence scoring. Submissions with a lot of confidence were automatically approved, and only submissions with a little confidence or very special cases were sent to human associates. This inversion of control—automation by default, humans by exception—represented a fundamental shift from prior approaches. Confidence thresholds were explicitly tunable, allowing the business to balance automation coverage against acceptable risk [5][6].

10.48047/jocaaa.2026.35.03.38

Characteristic	Human-First Workflow	Confidence-Driven Automation
Default processing path	Manual review for all submissions	Automated approval for high-confidence submissions
Human involvement	Mandatory for every submission	Exception-based for low-confidence cases
Scalability with volume	Linear growth in headcount required	Elastic scaling without proportional headcount growth
Decision consistency	Variable across individual reviewers	Standardized and reproducible
Quality governance	Dependent on reviewer expertise	Audit-sampled with threshold-based controls
Latency from submission to listing	Multi-day processing cycle	Near real-time for high-confidence submissions
Adaptability to errors	Manual process revision	Automated feedback loop with threshold adjustment

Table 2: Characteristics of Confidence-Driven vs. Human-First Workflow Design [5, 6, 11, 12]

4.2 Handling Complex Product Variations

Variation products, such as parent-child relationships for size and color, are among the most error-prone aspects of catalog onboarding, often leading to inconsistencies in product listings and customer dissatisfaction. Incorrect variation groupings lead to catalog fragmentation, where what should be a single unified product listing appears as multiple disconnected entries, degrading both search relevance and the customer experience. The platform incorporated specialized validation logic to ensure correct variation grouping and prevent these issues. This focus on variation handling was critical to achieving meaningful automation coverage across the diverse range of product categories on the marketplace [7][8].

4.3 Audit and Feedback Loops

To maintain trust and governance, the system integrated continuous audit sampling. A statistically significant subset of auto-created ASINs was routed to manual review, and error rates were tracked over time. If quality metrics exceeded predefined thresholds, automation rules and confidence levels were adjusted accordingly. This closed-loop control system made sure that quality stayed high while automation could safely grow over time. The feedback system also acted as an early alert, allowing quick action to be taken if quality dropped due to changes in seller actions or the types of products being sold.

5. Machine-Learning-Based ASIN Title Validation

5.1 Motivation for Title Automation

Product titles are a primary signal used in search and discovery, directly influencing customer click-through rates and conversion. In the past, associates spent a lot of time changing system-generated titles to make sure they met style and quality standards. These changes were based on personal opinion and specific categories and were hard to codify with rules alone. These variations often led to differences in title quality across different product categories, which hurt the overall customer experience and sales. The volume of title review work represented a substantial share of total associate effort, making it a high-

10.48047/jocaaa.2026.35.03.38

value target for automation. Reducing unnecessary title edits without sacrificing quality required a principled, data-driven approach rather than an expansion of manual heuristics [9][10].

5.2 Supervised Classification Formulation

The title validation problem was reframed as a supervised binary classification task: given a system-generated title, predict whether it requires human editing. Training data was derived from historical workflows by comparing system-generated titles with final associate-edited titles. Cases where the title was changed were labeled as requiring review; cases where it was accepted unchanged were labeled as acceptable. This approach turned a vague and personal task into a clear problem that a machine learning model could work on, using past human choices as the reliable information for the model to learn useful patterns.

5.3 Feature Engineering and Model Performance

The model incorporated domain-aware text normalization, attribute and numeric tagging, and n-gram linguistic features extracted from title text. Experiments showed that using only metadata features was not as effective as using both metadata and language features, indicating that language signals were important for understanding quality details that structured attributes alone couldn't capture, like small differences in title wording and context that influence how well a title performs. The final model achieved a level of accuracy and completeness that made it ready for real-world applications, enabling automatic approval of high-quality titles without manual checks, while still allowing for human review. The work was selected for presentation at the Amazon Machine Learning Conference in 2013, reflecting internal recognition of its technical novelty and operational impact [9][10].

Feature Category	Description	Role in Classification	Applicable to Metadata-Only Baseline
Domain-aware text normalization	Standardization of raw title text for consistent representation	Reduces surface-level noise before classification	Partial
Attribute tagging	Identification and marking of structured product attributes within title text	Anchors classification to known product properties	Yes
Numeric and unit tagging	Detection and normalization of measurements, quantities, and units	Captures formatting inconsistencies in dimension and size expressions	Yes
Unigram linguistic features	Single-word frequency and presence signals from title text	Captures vocabulary-level quality signals	No
N-gram linguistic features	Multi-word phrase patterns extracted from title text	Captures contextual and stylistic quality patterns	No

10.48047/jocaaa.2026.35.03.38

Category-specific lexicons	Domain vocabulary drawn from product category conventions	Enables category-aware quality judgment	Partial

Table 3: Feature Categories Used in ML-Based Title Validation [9, 10, 13, 14]

6. Evaluation and Impact

6.1 Automation Coverage and Latency

The combined platform automated a substantial share of ASIN creation workflows across a large population of catalog associates, producing significant reductions in total manual effort. ASIN creation latency was reduced from days to near real-time for high-confidence submissions, directly accelerating seller onboarding and time-to-sale for new products. The shift from a human-first to an automation-first model allowed the same associate capacity to handle a meaningfully larger volume of submissions, effectively decoupling catalog growth from headcount growth [11][12].

6.2 Quality and Consistency Outcomes

Automation reduced subjective variation across reviewers, producing cleaner titles, more consistent variation structures, and improved search discoverability. Continuous auditing ensured that quality metrics remained within acceptable bounds throughout the rollout and beyond. By standardizing the decision logic applied to each submission, the platform also reduced the dependency on individual reviewer expertise, making quality outcomes more consistent and reproducible across categories and geographies. These improvements had downstream effects on search ranking and customer satisfaction [11][12].

6.3 Economic Significance

Given the scale of Amazon's marketplace, even marginal efficiency gains in catalog onboarding translate into substantial operational savings when multiplied across millions of submissions. Faster onboarding also produces indirect revenue benefits by reducing time-to-sale for new products and enabling sellers to respond more quickly to market demand. The platform showed that investing in automation tools leads to increasing benefits as the number of submissions rises, making it a better choice than simply increasing the number of people reviewing submissions as the marketplace expands.

7. Related Work

7.1 Rule-Based Catalog Validation and PIM Systems

Early large-scale e-commerce platforms relied primarily on rule-based validation systems to enforce catalog correctness. Commercial Product Information Management systems used schema validation, required field checks, and deterministic business rules to make sure that all the data was there. While effective for syntactic validation, these systems lacked the ability to reason about semantic quality or subjective correctness. Research into PIM system design emphasized the significance of structured data governance but failed to tackle the issue of automating quality assessments that necessitate contextual comprehension beyond the capabilities of static rules [13][14].

7.2 Machine Learning in E-Commerce Prior to 2014

The dominant applications of machine learning in e-commerce during the period covered by this work focused on customer-facing optimization: search ranking, recommendation systems, personalization, fraud detection, and demand forecasting. Academic literature and industry publications from this era largely framed ML as a revenue-optimization or relevance-improvement tool rather than an operational

10.48047/jocaaa.2026.35.03.38

quality-control mechanism. Machine learning in e-commerce primarily aims to enhance customer experiences, including search results, recommendations, personalization, fraud detection, and demand prediction. During this time, most studies and industry reports viewed ML as a way to boost revenue or relevance.

7.3 Catalog Moderation and Exception-Based Review

Content moderation systems in adjacent domains introduced early forms of exception-based review, where automated filters escalated uncertain cases to humans. The Auto ASIN Creation platform took this idea and improved it for structured commerce data by adding clear confidence scores, using statistical methods to check samples, and adjusting thresholds based on feedback. This initiative changed exception-based review to focus on a clear, manageable engineering issue and showed a new way to apply moderation techniques to improve product catalog quality. The use of ongoing feedback made this method different from the usual fixed moderation systems found in content platforms, as it enables immediate changes and enhancements based on how users interact and the data collected, ultimately leading to improved product catalog quality and user satisfaction.

7.4 ML-Assisted Title and Attribute Quality Validation

Prior approaches to title quality enforcement relied on static heuristics, keyword blacklists, and manual style guides. These methods could find clear mistakes but missed the subtle details that skilled reviewers usually consider, like different word choices or specific industry rules that affect title quality. Research into supervised classification for product title and attribute quality showed that teaching a model based on past human decisions allows it to learn and store valuable knowledge in a way that can be reused and scaled. This method allows for decision-making based on probabilities rather than strict rules, making it possible to automate quality assessments that were once thought too subjective for machines to handle.

7.5 Distinguishing Characteristics of This Work

The main differences of this work compared to previous efforts are that it views catalog onboarding as a problem that can be automated based on confidence instead of a set human process, uses machine learning for quality checks instead of just for ranking products for customers, and combines rules, machine learning, and ongoing feedback into one complete system. While related areas had examined parts of this method, the truly new development was bringing all these pieces together into one effective system that works at a large marketplace level and shows a safe and lasting decrease in human review in a crucial process. These characteristics position the work as an early and substantive example of ML-powered operational automation in large-scale e-commerce [13][14].

10.48047/jocaaa.2026.35.03.38

10.48047/jocaaa.2026.35.03.38

Work / Approach	Domain Focus	Automation of Quality Gating	ML Applied to Operational Workflow	Feedback Loop Integrated	Scale of Deployment
Rule-based PIM validation systems	Catalog data completeness	No	No	No	Enterprise-level
Content moderation platforms	Policy and abuse enforcement	Partial (rule-based escalation)	Partial	No	Large-scale
ML for search ranking and recommendations	Customer-facing relevance	Not applicable	Yes (ranking only)	Partial	Production-scale
Title classification research (contemporaneous)	Product title categorization	No	Yes (offline/experimental)	No	Research-scale
Attribute extraction pipelines	Structured attribute identification	No	Yes (batch processing)	No	Experimental
Auto ASIN Creation Platform (this work)	End-to-end catalog onboarding	Yes (confidence-driven)	Yes (operational gating)	Yes (continuous audit)	Full production-scale

Table 4: Summary of Related Work and Positioning of This Contribution [1, 3, 15, 16]

8. Competitive and Industry Implications

8.1 Anticipating Industry-Wide Adoption

After this work, public documents and engineering blogs from other big marketplaces show that they are increasingly using similar methods, like automated checks for listing quality, human reviews for exceptions, and ML-assisted validation. The approach described here anticipated these trends by applying ML not to customer-facing ranking but to operational quality gating—a less explored application at the time. The overall move in the industry towards using automated systems for managing catalogs supports the design decisions made for this platform and highlights the ongoing importance of confidence-driven design, as seen in marketplaces around the world.

8.2 Generalizability of the Approach

The principles embedded in the Auto ASIN Creation platform generalize beyond the specific context of Amazon's catalog. Any large-scale system that requires quality decisions to be made consistently at high volume—whether in catalog management, document processing, or content moderation—can benefit from the same confidence-driven, audit-backed automation framework. The basic idea that human opinions can be turned into tasks that machines can learn to do, and that machines can take over human reviews when they are sure enough, is important and affects how systems are built in many different fields.

10.48047/jocaaa.2026.35.03.38

9. Discussion

9.1 Lessons from Production Deployment

This work demonstrates that large-scale operational workflows can benefit from the same ML rigor traditionally reserved for user-facing features. Key lessons from the deployment include the need to make decisions based on confidence to safely increase automation, the use of continuous auditing to maintain quality over time, and the benefit of viewing human judgments as learnable signals instead of just manual tasks. These lessons are directly applicable to other catalog engineering challenges and to quality-enforcement problems across adjacent domains [11][12].

9.2 Limitations and Future Directions

While the platform achieved substantial automation coverage, confidence-based routing inherently preserves a residual population of submissions that require human review. Future work could explore increasingly fine-grained confidence models that reduce this residual further, as well as multimodal approaches that incorporate product imagery alongside text and metadata. Recent improvements in large language models since the first release have made it possible to improve title validation by using better representation learning. This could lead to more accurate results in different categories and languages without needing special features for each one.

Conclusion

The Auto ASIN Creation platform and ML-based Title Validation system represent an original, production-scale contribution to large-scale catalog engineering. By transforming a human-first workflow into a confidence-driven automated system, the work enabled Amazon to scale its marketplace efficiently while preserving quality and trust. The ideas presented anticipated broader industry adoption and continue to inform best practices in catalog onboarding and data quality automation, particularly in how companies can leverage automated systems to enhance their operational efficiency and maintain high standards of data integrity. This article documents original work conducted within Amazon Retail Business Services between 2012 and 2014 and is intended to present the technical and architectural contributions in a scholarly format suitable for expert evaluation.

References

[1] ELIAS DRITSAS AND MARIA TRIGKA, "Machine Learning in E-Commerce: Trends, Applications, and Future," *IEEE Access*, June 12, 2025. Available: https://ieeexplore.ieee.org/iel8/6287639/10820123/11009009.pdf

[2] Yaswanth Jeganathan, "Enhancing Product Descriptions Using Large Language Models in E-Commerce," *International Journal for Research Trends and Innovation (IJRTI)*, June 06, 2025. Available: https://www.ijrti.org/papers/IJRTI2506181.pdf

[3] Elias Dritsas and Maria Trigka, "Exploring the Intersection of Machine Learning and Big Data: A Survey," *MDPI (Data)*, February 07, 2025. Available: https://www.mdpi.com/2504-4990/7/1/13

[4] Boris Otto, "Managing the business benefits of product data management: the case of Festo," *Journal of Enterprise Information Management (Emerald)*, April 13, 2012. Available: https://www.emerald.com/jeim/article/25/3/272/197401/Managing-the-business-benefits-of-product-data

[5] Heng-Chang Lin, et al., "Integrating Product Information Management (PIM) with Internet-Mediated Transactions (IMTs)," *ResearchGate (Preprint)*, October 09, 2014. Available:

10.48047/jocaaa.2026.35.03.38

https://www.researchgate.net/publication/220147123_Integrating_Product_Information_Management_PIM_with_Internet-Mediated_Transactions_IMTs

[6] HAROLD COTACALLAPA, et al., "A Flat-Hierarchical Approach Based on Machine Learning Model for e-Commerce Product Classification," *IEEE Access*, May 30, 2024. Available: https://ieeexplore.ieee.org/iel7/6287639/10380310/10530036.pdf

[7] Joel Bäcker, Victor Winkelmann, "Automated product categorization using transformer models," *Lund University Publications (Student Papers)*, February 02, 2024. Available: https://lup.lub.lu.se/student-papers/record/9147956/file/9147960.pdf

[8] Joel Antony and Amal K Jose, "AI-Powered Product Categorization for E-Commerce," *Zenodo*, 2025. Available: https://zenodo.org/records/15673727/files/AI-Powered%20Product%20Categorization%20for%20E-Commerce.pdf?download=1

[9] Noor Mathivanan, "Performance Analysis of Supervised Learning Models for Product Title Classification," *IAES International Journal of Artificial Intelligence (IJ-AI)*, December 2019. Available: https://ijai.iaescore.com/index.php/IJAI/article/view/20273

[10] Sylvain Goumy and Mohamed-Amine Mejri, "Ecommerce Product Title Classification," *CEUR-WS.org*, July 12, 2018. Available: https://ceur-ws.org/Vol-2319/ecom18DC_paper_5.pdf

[11] Tarun Gupta, Supriya Bansal, "Machine Learning Algorithms for Predictive Analytics in E-Commerce," *International Journal of Science and Research (IJSR)*, August 15, 2020. Available: https://www.ijsr.net/archive/v9i8/SR24302212009.pdf

[12] Chandra Prakash Gupta, "Predictive Analytics Application in E-Commerce: An Artificial Intelligence-Driven Approach to Success," *IEEE Xplore*, June 24, 2025. Available: https://ieeexplore.ieee.org/document/11042327/

[13] Luiz Felipe Sales, et al., "Multimodal deep neural networks for attribute prediction and applications to e-commerce catalogs enhancement," *Multimedia Tools and Applications (Springer)*, April 24, 2021. Available: https://www.researchgate.net/publication/351083612_Multimodal_deep_neural_networks_for_attribute_prediction_and_applications_to_e-commerce_catalogs_enhancement

[14] Norsyela Muhammad Noor Mathivanan, et al., "A Mini Review on Supervised Learning Models for E-Commerce Product Classification," *International Journal of Advanced Trends in Computer Science and Engineering (IJATCSE)*, November 15, 2019. Available: https://www.warse.org/IJATCSE/static/pdf/file/ijatcse99914sl2020.pdf

[15] ADAM WASILEWSKI, et al., "Enhanced E-Commerce Personalization Through AI-Powered Content Generation Tools," *IEEE Access*, March 21, 2025. Available: https://ieeexplore.ieee.org/iel8/6287639/10820123/10925355.pdf

[16] KAPILYA GANGADHARAN, et al., "From Data to Decisions: The Power of Machine Learning in Business Recommendations," *IEEE Access*, January 29, 2025. Available: https://ieeexplore.ieee.org/iel8/6287639/10820123/10849522.pdf