Journal of Scientometric Research, 2026; 15(3):780-792.
https://www.jscires.org

Research Article

A Data Integration Pipeline for Research Information Management: Implementation and Performance Assessment in Higher Education Contexts

Luis Ramirez-Polo[1,*], Nestor Caicedo-Solano[2], Fabio Bermejo-Altamar[1]

[1]Productivity and Data, Universidad Autonoma del Caribe, Barranquilla, COLOMBIA.
[2]Industrial Engineering, Universidad del Magdalena, Santa Marta, COLOMBIA.

ABSTRACT

Higher education institutions face persistent challenges when research data remains fragmented across multiple platforms, limiting strategic decision-making and comprehensive research oversight. This issue is particularly acute in emerging economies, where manual collection processes constrain institutional planning and competitiveness in global research evaluation frameworks. This study presents the design, implementation, and validation of an automated data integration pipeline that consolidates heterogeneous research data sources to strengthen institutional research management and analytics. The system integrates Colombia's national research platform (GrupLAC), international bibliometric databases (Scopus and Web of Science), and internal institutional systems. To resolve author disambiguation and affiliation matching challenges, the pipeline employs fuzzy matching algorithms combining FuzzyWuzzy and python-Levenshtein libraries. Implemented at a Colombian university with 536 academic divisions members across six academic divisions**,** the pipeline achieved 99% data coverage while processing 50,000 records in under 5 min, with 94.2% accuracy in author disambiguation and 96.8% in affiliation resolution. Network analysis revealed interdisciplinary collaboration structures across twenty-two active research groups that remained invisible under prior fragmented data management approaches. Results are delivered through a multi-level Power BI dashboard supporting evidence-based institutional decision-making. This work contributes a practical, open-source framework for institutions operating under dual national-international evaluation systems a context underserved by existing commercial CRIS solutions and provides a nationally scalable framework based on standardized ministerial metrics, alongside a methodological blueprint adaptable to similar emerging economy contexts.

Keywords: Data integration, Higher education information systems, Information systems management, Institutional research management, Organizational data analytics, Research information systems.

Correspondence:
Luis Ramirez-Polo
Productivity and Data, Universidad Autonoma del Caribe, Barranquilla, COLOMBIA.
Email: luis.ramirez@uac.edu.co

Received: 05-06-2026;
Revised: 21-08-2026;
Accepted: 18-09-2026.

INTRODUCTION

Higher education institutions generate vast research data that could significantly enhance educational quality and program development, yet fragmented information systems prevent universities from leveraging this knowledge to improve learning experiences and curriculum design. This technological gap particularly affects universities in emerging economies, where integrated research-educational insights could optimize limited resources and strengthen program delivery. Research outputs, academic divisions productivity metrics, collaboration networks, and resource utilization data remain scattered across national research databases, international bibliometric platforms, institutional repositories, and internal management systems. This fragmentation creates significant barriers to comprehensive research oversight and evidence-based institutional planning.

The technological infrastructure gap between data availability and institutional analytical capability has become particularly pronounced in the digital transformation of higher education. While universities generate unprecedented amounts of research-related data, the absence of integrated systems to consolidate, validate, and analyze this information limits institutional capacity for strategic research management. Academic leaders require comprehensive visibility into research productivity patterns, collaboration networks, and resource allocation effectiveness, yet existing systems typically provide only partial, siloed perspectives on institutional research activity. This challenge intensifies in emerging economic contexts, where

OPEN ACCESS

Manuscript TECHNOMEDIA

DOI: 10.5530/jscires.20260148

Copyright Information :
Copyright Author (s) 2026 Distributed under Creative Commons CC-BY 4.0

Publishing Partner : Manuscript Technomedia. [www.mstechnomedia.com]

universities must simultaneously demonstrate research impact to national funding agencies, compete for international recognition, and support educational innovation and program quality improvement. Institutions in countries like Colombia operate within complex evaluation frameworks that combine national research classification systems with international bibliometric standards, creating additional integration requirements that generic research information systems cannot adequately address. Most universities implement fragmented information systems approaches that address individual operational requirements without comprehensive integration capabilities. Student Information Systems (SIS), Learning Management Systems (LMS), Human Resources Information Systems (HRIS), and Research Information Management Systems (RIMS) operate independently, creating data silos that limit institutional analytical capabilities.

Current Research Information Systems (CRIS) provide institutional research data management capabilities for individual institutions, collecting and storing structured data about academic divisions research and scholarly activities. However, these systems typically focus on internal data management without addressing external integration challenges or multi-source validation requirements essential for comprehensive institutional assessment. Learning analytics research has established sophisticated frameworks for educational data integration, particularly in Latin American contexts where (Guerra *et al*., 2020), demonstrated successful dashboard implementation at three universities, though primarily focused on student performance rather than institutional research management. Research by (Hilliger *et al*., 2020), identified specific needs for learning analytics adoption in Latin American universities, revealing gaps in comprehensive institutional data integration approaches. Bibliometric literature addresses author disambiguation and data quality challenges in individual databases, with recent advances by (Krüger and Petersohn, 2023), highlighting persistent difficulties in translating digital infrastructure capabilities into effective evaluation practices but lacks comprehensive solutions for institutional-scale integration across heterogeneous sources.

However, systematic reviews by (Banihashem *et al*., 2022), reveal persistent challenges in implementing comprehensive institutional analytics, while recent frameworks by (Kaliisa *et al*., 2023), provide guidelines for learning analytics dashboard implementation in higher education contexts. The intersection of educational technology and institutional research management remains underexplored, particularly regarding practical implementation frameworks that universities can adapt to local contexts and requirements.

The methodology builds on established principles from current research information systems while incorporating recent advances in educational technology implementation frameworks developed for higher education contexts. The system utilizes fuzzy matching algorithms and automated data validation procedures to achieve high accuracy rates while processing large volumes of institutional research data. The case study reveals collaboration patterns, productivity distributions, and resource allocation dynamics that remained invisible under previous manual data management approaches, demonstrating the practical value of comprehensive research data integration for institutional planning and decision-making.

The study contributes to the growing body of literature on educational technology applications in higher education administration while addressing specific gaps in research information management. The implementation demonstrates how contemporary data integration techniques can be applied to solve practical challenges in university research oversight, providing a foundation for future developments in this rapidly evolving field. This paper focuses specifically on the technological and methodological aspects of research data integration in higher education contexts, with practical validation through implementation at a Colombian university serving as a case study for similar institutions in emerging economies. The scope encompasses technical architecture design, data processing methodology, validation procedures, and assessment of practical utility for institutional research management.

LITERATURE REVIEW

The systematic integration of research data in higher education institutions represents a convergence of three distinct but interconnected research domains: educational technology applications in institutional management, research information systems development, and bibliometric data processing methodologies. This section reviews the current state of knowledge in each area and identifies the gaps that motivate the present study. Educational Technology in Higher Education Research Management. Educational technology research has increasingly focused on data-driven approaches to institutional decision-making, yet most applications concentrate on student learning analytics rather than comprehensive institutional research management (Guerra *et al*., 2020). Demonstrated successful implementation of learning analytics dashboards at three Latin American universities, revealing both the potential and limitations of current approaches. Their study found that while dashboards effectively improved academic support services, significant gaps remained in integrating research productivity data with institutional planning processes.

Recent developments in learning analytics dashboard design have emphasized the importance of providing actionable insights rather than merely descriptive statistics. A study by (Susnjak *et al*., 2022) investigated current approaches to LA (Learning Analytics) dashboarding and found that most LA dashboards employ only surface-level descriptive analytics, while few utilize predictive analytics to provide actionable guidance. This limitation reflects

broader challenges in translating data availability into meaningful institutional intelligence.

The adoption of learning analytics in higher education faces substantial challenges that mirror those encountered in research data integration (Hilliger *et al.*, 2020). Conducted a comprehensive mixed-methods analysis of learning analytics needs across Latin American universities, identifying critical barriers including fragmented data sources, limited technical expertise, and lack of systematic implementation frameworks. Their findings revealed that 73% of surveyed institutions collected research data but only 31% had implemented integrated analysis systems, highlighting the persistent gap between data availability and analytical capability. Contemporary research reveals increasing sophistication in understanding student engagement through learning analytics (Bergdahl *et al.*, 2024). Conducted a systematic review of student engagement in higher education learning analytics, revealing the complex and multifaceted nature of engagement measurement and the need for comprehensive analytical frameworks that can capture diverse dimensions of institutional activity.

Systematic reviews of learning analytics implementation reveal recurring themes relevant to institutional research management (Banihashem *et al.*, 2022). Analyzed the role of learning analytics in enhancing feedback practices, finding that successful implementations required careful attention to data quality, user experience design, and institutional change management processes.

Privacy and ethical considerations in learning analytics implementation present challenges directly applicable to research data integration systems (Jones, 2019). Proposed models for establishing informed consent mechanisms to promote student privacy and autonomy in learning analytics contexts. Their framework addresses critical issues of data privacy, user autonomy, and institutional accountability that are equally relevant to research data integration systems, particularly when managing sensitive academic divisions research and productivity data.

Recent frameworks for learning analytics dashboard design provide methodological guidance applicable to research management contexts (Kaliisa *et al.*, 2023). Comprehensive guidelines for planning, designing, implementing, and evaluating learning analytics dashboards in higher education, emphasizing the importance of stakeholder engagement, ethical considerations, and systematic evaluation procedures. This work builds on Teaching Analytics research by (Ndukwe and Daniel, 2020), established theoretical approaches combining teaching expertise, visual analytics, and design-based research to support diagnostic pedagogical abilities using data and evidence.

However, critical analysis of learning analytics implementation reveals significant limitations in current approaches (Guzmán-Valenzuela *et al.*, 2021), argue that learning analytics tends to underplay the complexity of teaching-learning processes and frequently focuses on analytics without sufficient attention to actual learning outcomes. Research Information Management Systems (RIMS) have emerged as specialized technological solutions for the systematic collection, organization, and utilization of institutional research data in higher education contexts (Givens, 2016). RIMS collect and store structured data about academic divisions research and scholarly activities for one institution, with the intention of repurposing the information in a variety of ways, fundamentally transforming how universities manage and leverage their research intelligence for strategic decision-making.

The theoretical foundation of research information management builds on principles of organizational knowledge management and information systems theory. Research Information Management (RIM) is defined by the independent bodies OCLC Research and euroCRIS as "the aggregation, curation, and utilization of information and research" (Givens, 2016) establishing a comprehensive framework that extends beyond simple data storage to encompass systematic approaches for transforming research data into institutional intelligence.

Current Research Information Systems (CRIS) represent the most mature implementation of research information management principles in higher education. A Current Research Information System (CRIS) is a database or other information system to store, manage and exchange contextual metadata for the research activity funded by a research funder or conducted at a research-performing organization (Schöpfel *et al.*, 2017). These systems address the fundamental challenge of research data fragmentation by providing integrated platforms that consolidate diverse research information sources into coherent analytical capabilities.

The implementation of CRIS systems reveals significant variability across institutional contexts and regulatory environments. Globally, the RIM ecosystem is quite mature in locales with national requirements to collect and report the outputs of institutional research activities (Dvořák *et al.*, 2019), while institutions in other contexts face greater implementation challenges due to varying reporting requirements and technological infrastructure capabilities. Contemporary research on CRIS implementation emphasizes the importance of integration capabilities with external data sources. Current research information systems are increasingly being used to standardize and ease documentation, communication, and administration of research (Dvořák *et al.*, 2019). With broad coverage and sufficient completeness, data quality, and standardization, CRIS systems can also be used as data sources for studies of institutional research patterns and strategic planning processes.

The Colombian context presents unique opportunities for research information management development, particularly given the recent evolution of sustainable development initiatives in higher education institutions. Recent analysis of sustainable development evolution in Colombia (Altamar *et al.*, 2025) demonstrates the increasing importance of systematic approaches to research management in emerging economy contexts, where institutions must simultaneously address local development priorities while maintaining international research competitiveness. However, significant challenges persist in CRIS implementation across diverse institutional contexts. Current Research Information Systems (CRIS) and Institutional Repositories (IR) were developed as clearly distinguished systems, with different objectives and functionalities, standards and data models, and for different needs and user groups (Schöpfel *et al.*, 2017), creating integration challenges that require sophisticated technical approaches to resolve data inconsistencies and achieve comprehensive coverage.

The integration of research data within CRIS frameworks presents technical challenges related to data quality control and multi-source validation. Recent research and publications on the integration of research data in Current Research Information Systems (CRIS) highlight persistent difficulties in achieving high accuracy rates while maintaining operational efficiency suitable for institutional requirements (Schöpfel *et al.*, 2017). These challenges are particularly acute when integrating national evaluation systems with international bibliometric databases and internal institutional sources, creating the need for specialized solutions like the pipeline presented in this study.

Contemporary research on bibliometric data processing emphasizes the importance of robust data cleaning and validation procedures. Author disambiguation represents one of the most challenging aspects of bibliometric analysis, particularly when integrating data across multiple sources with varying naming conventions, affiliation formats, and temporal coverage. Standard exact-matching approaches typically achieve only 60-70% coverage rates in real-world institutional contexts, necessitating more sophisticated fuzzy matching techniques. Recent advances in digital infrastructure for bibliometric analysis reveal both opportunities and limitations in current approaches. (Krüger and Petersohn, 2023) analyzed how evaluative bibliometrics practitioners navigate digital tools, identifying persistent difficulties in translating technological capabilities into effective evaluation practices. Their findings demonstrate that while digital infrastructure offers unprecedented opportunities for comprehensive assessment, substantial barriers remain in institutional contexts where fragmented manual processes continue to dominate. The integration of national and international bibliometric sources presents additional complexity for institutions operating within multiple evaluation frameworks. Colombian institutions must reconcile data from national databases (GrupLAC) with international sources (Scopus, Web of Science) while maintaining accuracy and avoiding double-counting of research outputs. This multi-source integration requirement exceeds the capabilities of standard bibliometric analysis tools and necessitates custom solutions.

This study addresses these gaps by developing an integrated approach that combines advances from all three research domains to create a comprehensive solution for institutional research data management. The theoretical framework builds on established principles from educational technology for user experience design, research for data architecture, and bibliometric analysis for data quality control and validation procedures. The practical contribution lies in demonstrating how these theoretical frameworks can be combined into a working system that addresses real-world institutional requirements while maintaining technical efficiency and user accessibility. The implementation provides empirical evidence of the institutional insights that emerge from comprehensive data integration, contributing to both theoretical understanding and practical guidance for similar institutional contexts.

METHODOLOGY

Overall System Architecture

The data integration pipeline addresses the multi-source integration challenges identified in the literature review by implementing a comprehensive four-stage process: data extraction, integration and validation, database modeling, and interactive visualization (Figure 1). The architecture builds on established frameworks from educational technology research while addressing specific requirements for research data management in emerging economy higher education contexts.

The system integrates three primary data streams to create a comprehensive research evaluation database: Colombia's national research platform (GrupLAC) for institutional validation, international bibliometric databases (Scopus, Web of Science) for citation metrics, and institutional databases for academic divisions and programs information. This multi-source approach addresses the integration challenges identified by Hilliger *et al.* (2020) in their analysis of Latin American university data management needs.

The technical architecture responds to persistent difficulties in translating digital infrastructure capabilities into effective evaluation practices, as documented by (Krüger and Petersohn, 2023) Rather than implementing separate systems for each data source, the integrated approach enables comprehensive analysis across institutional, national, and international research metrics.

Data Source Integration Strategy

Colombian National Research Framework Integration

The pipeline's foundation integrates with Colombia's comprehensive national research evaluation system administered by Minciencias (Science and Technology Ministry). This system operates through formally recognized research groups that serve as fundamental organizational units for research evaluation within Colombian universities. Each research group maintains an official profile on the GrupLAC platform containing detailed research information, research lines, and comprehensive research product listings according to standardized national typologies.

The standardized classification framework categorizes research outputs into specific typologies including scientific articles, books, book chapters, technological products, patents, and knowledge transfer activities. Each category receives relative weights and points values reflecting their contribution to national research objectives and institutional evaluation criteria. This systematic approach enables comparative assessment across institutions while maintaining consistency with national science and technology policy priorities.

The public GrupLAC platform presents a critical limitation for institutional evaluation: while displaying comprehensive research product inventories, it does not indicate which products have achieved formal validation by Minciencias reviewers or their assigned point values. This restriction significantly limits utility for accurate institutional performance assessment, as validation status and scoring represent crucial dimensions for research evaluation compliance and strategic planning.

International Bibliometric Database Integration

The system integrates data from Scopus and Web of Science to provide comprehensive coverage of internationally indexed research outputs. This integration addresses the challenge of combining national evaluation standards with international bibliometric measures while managing complex institutional relationships where researchers participate across multiple groups, programs, and faculties.

Integration with international databases enables validation of research outputs claimed in national systems while providing additional metrics including citation counts, journal impact factors, and collaboration network analysis. The system maintains

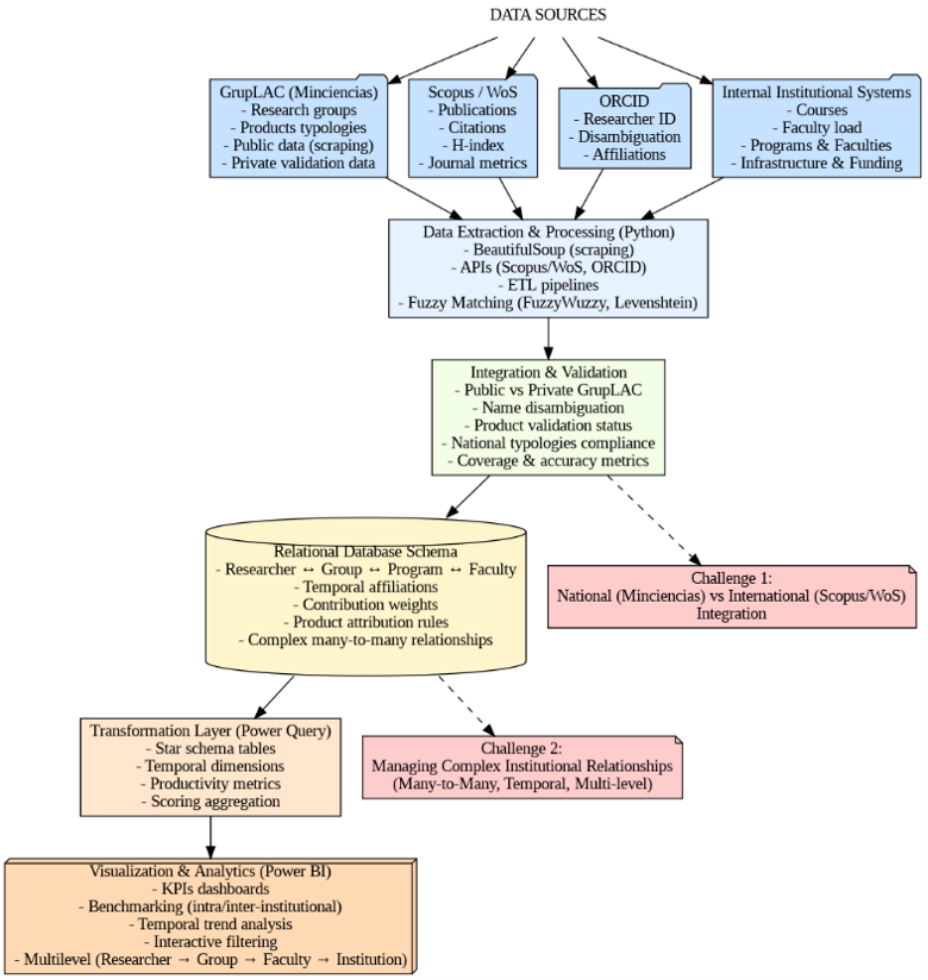


Figure 1: System Architecture Overview.

detailed records of validation status for each research product across all integrated sources.

Institutional Data Integration

Comprehensive internal institutional data integration creates complete researcher profiles enabling sophisticated multi-dimensional analysis across organizational structures. Internal academic information systems provide essential data including course assignments by semester, program affiliations, academic divisions appointments, and credit hour distributions across teaching, research, and service activities.

The system captures information about research infrastructure availability, laboratory affiliations, equipment access, and funding support that contextualizes individual and group research performance within available resources. This comprehensive approach addresses the multi-faceted nature of academic productivity identified in contemporary research on institutional factors affecting research performance.

Technical Implementation

Data Extraction and Processing

The system employs Python for all data extraction operations, utilizing BeautifulSoup for GrupLAC web scraping and API connections for Scopus and Web of Science integration. The critical technical challenge involves developing robust cross-referencing algorithms that accurately align public GrupLAC data with private validation files and international bibliometric records (Figure 2).

Fuzzy Matching Implementation

The achievement of high data coverage requires sophisticated fuzzy logic algorithms to address common data quality issues in academic databases. These challenges include orthographic variations in author names, incomplete name entries, and inconsistent punctuation that traditionally cause significant information loss in exact matching approaches.

The solution employs an integrated system combining FuzzyWuzzy, BeautifulSoup, and python-Levenshtein libraries. The python-Levenshtein library provides algorithmic foundation through efficient edit distance calculations, quantifying minimum operations required to transform one string into another. FuzzyWuzzy leverages this fundamental metric to implement advanced comparison strategies including simple ratio, partial ratio, token sort, and token set matching (Figure 3).

Database Architecture Design

The relational database schema addresses multifaceted relationships inherent in Colombian university research structures. The complexity stems from researchers participating in multiple research groups simultaneously while maintaining teaching responsibilities across various academic programs and service commitments at different institutional levels (Figure 4).

The researcher-group relationship table maintains temporal validity periods, participation levels, research time allocation percentages, and contribution weights enabling accurate performance attribution across different time periods. Research product attribution represents the most complex aspect, as each product must be assigned to a primary research group while maintaining detailed records of all participating researchers, their contribution levels, and institutional resource access.

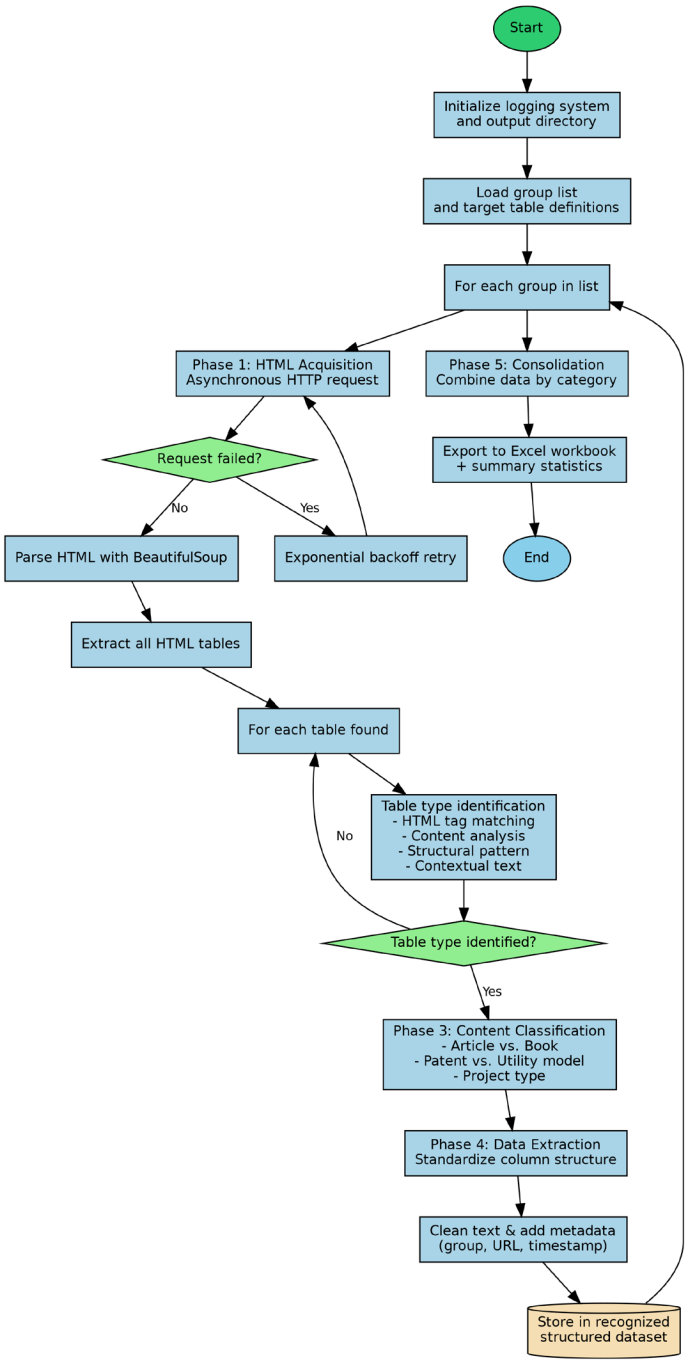


Figure 2: Multi-Source Data Extraction Process.

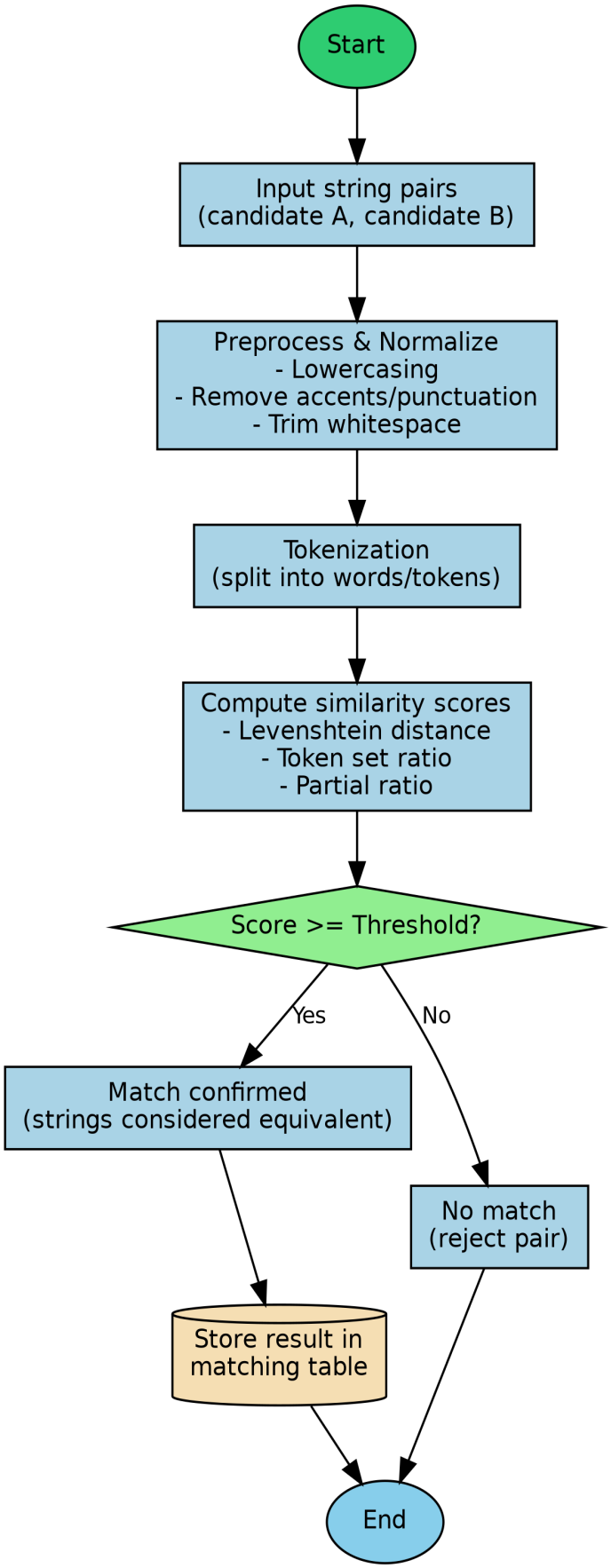


Figure 3: Fuzzy Matching Procedure.

Validation and Quality Control Procedures

Cross-Source Validation Framework

Validation procedures comparing internal database records with external sources implement systematic approaches to ensure data accuracy and completeness. The system employs multiple validation layers including author disambiguation validation, institutional affiliation verification, and publication classification accuracy assessment.

Manual validation sampling provides ground truth data for algorithm performance assessment. Random samples of 1,000 author records undergo manual verification to establish baseline accuracy measures for automated matching procedures. This validation approach follows established practices in bibliometric data quality control while addressing specific challenges of multi-source integration.

Data Quality Metrics

The system implements comprehensive data quality monitoring including coverage assessment, accuracy measurement, and consistency validation across integrated sources. Quality metrics track processing performance over time and identify potential issues requiring manual intervention or algorithm adjustment.

Temporal validation ensures consistency in researcher-institution relationships across different time periods, addressing usual challenges in academic career tracking where affiliations change over time. The system maintains audit trails for all data modifications enabling transparency in data quality control procedures.

Ethical Considerations

This study was conducted in accordance with the institution's Research Ethics Committee, the body responsible for research oversight at Colombian universities under the national regulatory framework established by Resolution 8430 of 1993 issued by the Ministry of Health. As the study involved institutional administrative data rather than direct human subjects research, it was classified as risk-free research, and a formal waiver of individual informed consent was granted on the basis that the project constituted an internal quality improvement initiative.

All data handling adhered to principles of purpose limitation using data solely for institutional research management data minimization, and transparency, with academic divisions informed of the system's scope and purpose prior to deployment. While individual level data was technically necessary for disambiguation and cross source matching, all outputs and reports presented to administrators displayed exclusively aggregated data, ensuring no individual researcher's productivity or affiliation status was directly exposed through the system's interfaces.

RESULTS

Technical Performance and Data Integration

System Performance Metrics

The implemented pipeline successfully integrated multiple heterogeneous data sources creating a comprehensive research evaluation database. The system processed approximately 50,000 records from all integrated sources with total processing time not exceeding 5 min per cycle, achieving 99% data coverage rate indicating minimal data loss during integration (Table 1).

These performance measures indicate operational efficiency well suited for the present institutional deployment. Processing 50,000 records in 4.7 min with 2.1 GB of memory establishes a solid benchmark for mid-sized universities. While this single benchmark confirms the system manages the institution's current data volume effectively, full system scalability for larger national datasets was not empirically evaluated in this phase. Given that fuzzy-matching procedures become computationally expensive as dataset sizes grow, future deployments with bigger volumes will necessitate multi-scale benchmarking to adequately define algorithmic complexity and memory scaling behavior. Nonetheless, the current design comfortably suits the institution's immediate analytical and reporting requirements. The 99% coverage rate reported with a 95% confidence interval of 98.9%-99.1% (n=50,000), calculated using the normal approximation for proportions; given the large sample size, this estimate carries minimal sampling uncertainty.

To contextualize these metrics, the 4.7 min processing time represents a strategic trade-off between computational speed and integration accuracy. While deterministic (exact) matching algorithms can process this data volume in a fraction of the time, they are unsuitable for this context due to the high variability in Latin American naming conventions (e.g., inconsistent use of double surnames) and unstandardized affiliation strings, which lead to severe data loss. Conversely, relying on manual reconciliation or legacy administrative tools for 50,000 records previously required weeks of human effort. Thus, the fuzzy-matching computational cost is justified, as it completely automates a resource-intensive administrative burden while achieving the 94.2% disambiguation accuracy required for reliable institutional reporting.

Fuzzy Matching Algorithm Performance

Validation procedures achieved high accuracy rates across different matching scenarios. Manual validation of 1,000 randomly selected records provided ground truth data for algorithm performance assessment (Table 2).

These accuracy rates, with 95% confidence intervals calculated using the normal approximation for binomial proportions, represent significant improvement over exact traditional matches approaches, which typically achieve 60 to 70% coverage rates in similar institutional contexts. The narrow confidence intervals (width ≤2.8% points across all three categories) reflect the precision afforded by sample sizes of 500-1,000 manually validated records.

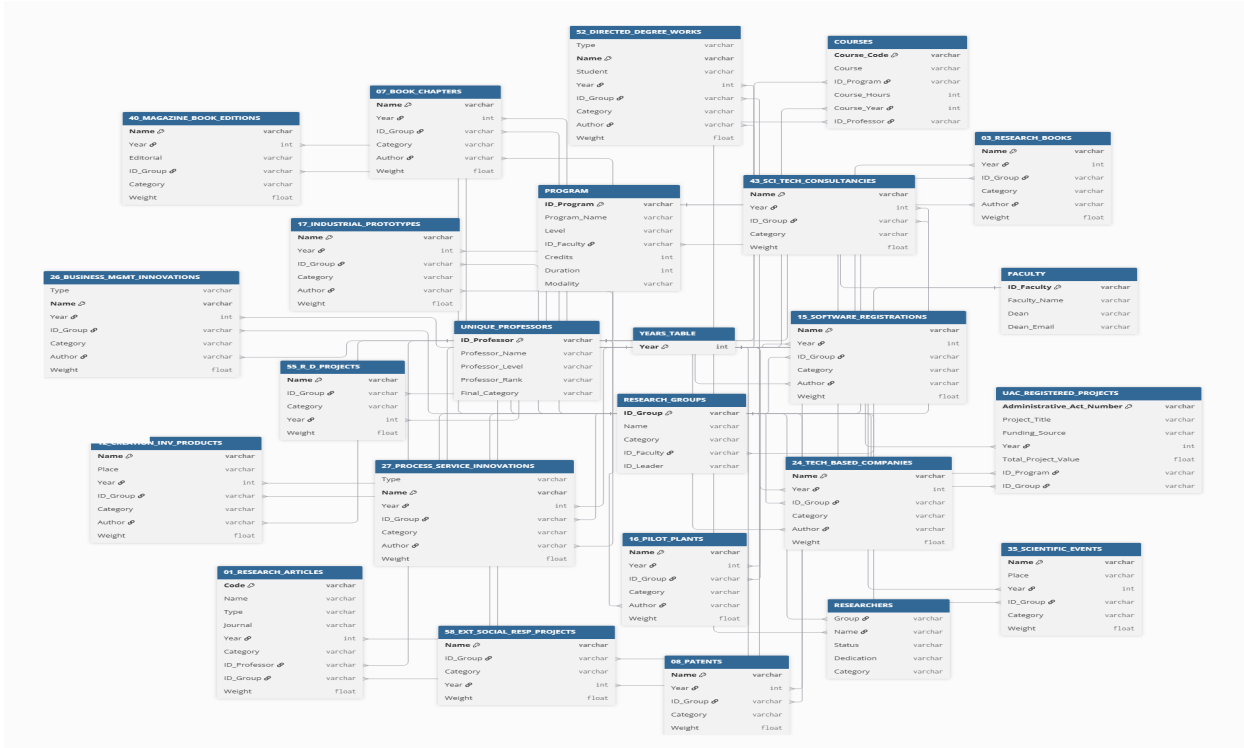


Figure 4: Database Schema Architecture.

Table 1: System Performance Metrics.

Metric	Value	Validation Method
Total records processed	50,000	System logs
Processing time per cycle	4.7±0.3 min	Automated timing
Data coverage rate	99%	Missing data analysis
Memory usage peak	2.1 GB	System monitoring
Data coverage Rate	99%, 95% CI (98.9%-99.1%)	Missing data analysis (n=50,000)

Institutional Data Characterization and Research Landscape Analysis

Academic Structure Analysis

Comprehensive data integration revealed institutional research, landscape complexity, and organizational structure. The analysis encompassed the complete academic ecosystem, providing unprecedented visibility into research capacity and organizational dynamics: six faculties representing primary organizational divisions, forty-eight academic programs distributed across all faculties, 1,360 courses comprising the complete curricular offering, and 536 academic divisions members of whom 197 (36.7%) hold permanent positions.

Distribution analysis revealed significant organizational heterogeneity. The largest academic divisions contain thirty-eight academic programs representing 79% of all institutional programs, indicating concentration of academic offerings. Course distribution demonstrated greater variability, with one academic division managing over five hundred courses, suggesting significant curricular complexity and resource allocation challenges.

Research Personnel and Productivity Analysis

Academic divisions workload analysis revealed 170 academic divisions members (86.3% of permanent academic divisions) actively producing research outputs, indicating high research concentration among academic staff (Table 3). Historical research data integration identified 1,137 researchers contributing to institutional research output over 20-year period.

Cross-academic divisions collaboration analysis revealed 30% of courses delivered across faculty boundaries. In a descriptive comparison across the full faculty database, faculty members teaching across multiple faculties showed a 45% higher rate of inter-group co-authorship than faculty teaching in a single faculty, suggesting that organizational permeability may facilitate research productivity. This estimate should be interpreted as descriptive rather than confirmatory; Section 4.5.4 reports a confidence-interval verification on a validated sub-sample and discusses the conditions under which this association could be more rigorously tested.

Research Output Classification and Multi-Source Integration Results

Multi-Source Data Integration Results

The pipeline processed and classified research outputs across sixty distinct categories providing comprehensive coverage of discipline and methodological diversity. The classification system successfully processed the largest single record containing 70,000 individual data points without system degradation or processing errors.

Cross-source validation comparing internal records with external databases achieved high validation rates across distinct categories of research outputs. Integration with Minciencias validation data provided essential information about product validation status and point values not available through public channels.

The multi-source integration revealed research productivity patterns and validation discrepancies that remained invisible under previous manual data management approaches. This comprehensive view enables more accurate institutional performance assessment and strategic research planning.

Collaboration Network Analysis

Network analysis revealed complex interdisciplinary relationships transcending traditional organizational boundaries (Figure 5). The visualization identified distinct collaboration clusters while highlighting areas for strategic development. The network mapped research groups (gray nodes) and academic programs (yellow nodes), illustrating many-to-many relationships inherent in the institutional research landscape. Strong collaboration clusters emerged between engineering groups and administrative/economic sciences programs, while weaker connections suggest opportunities for targeted relationship development.

Dashboard Implementation and User Interface Performance

Power BI Dashboard Performance

Microsoft Power Query transformation restructured relational data into specialized analytical tables optimized for dashboard consumption. The dashboard architecture balanced executive-level summaries with operational insights, incorporating temporal trend analysis, comparative benchmarking, and interactive filtering across multiple institutional dimensions. Executive interfaces provide high-level institutional performance summaries, while departmental views enable detailed exploration of collaboration patterns and productivity trends.

Table 2: Fuzzy Matching Accuracy Results.

Matching Category	Sample Size	Correct Matches	Accuracy Rate	95% CI
Author disambiguation	1,000	942	94.2%	92.8%-95.6%
Institutional affiliations	500	484	96.8%	95.3%-98.3%
Publication matching	800	785	98.1%	97.2%-99.0%

Table 3: Research Activity Distribution.

Category	Count	Percentage
Total academic divisions members	536	100%
Permanent academic divisions	197	36.7%
Research-active academic divisions	170	86.3% of permanent
Historical researchers	1,137	Over 20-year period
Active research groups	22	Status

System Usability and Institutional Applications

User interface design followed established guidelines from educational technology research, particularly frameworks developed by Kaliisa *et al*., 2023 for learning analytics dashboard implementation. Initial user feedback indicates successful translation of complex research evaluation data into actionable strategic insights. The system demonstrates capacity to reveal analytical insights invisible under traditional evaluation approaches, representing a significant advancement over the manual data management processes previously employed by the institution (Figure 6).

Generalizability, Limitations and Technical Considerations

Case Study Scope and Generalizability

While this case study demonstrates the effectiveness of the proposed pipeline within a specific institutional context, its architectural design adheres strictly to the data standards and evaluation parameters established by the Colombian Ministry of Science, Technology and Innovation (Minciencias) and the Ministry of Education. Because these metrics are mandatory for all higher education institutions in Colombia, the core framework is highly scalable and directly transferable at the national level.

However, its generalizability to international universities requires careful consideration. The methodological framework comprising the four-stage architecture (extraction, integration, modeling, visualization), the use of fuzzy matching for disambiguation, and the principles for dashboard design are highly transferable. Yet, specific implementation details will require adaptation to local data sources, regulatory frameworks, and institutional structures. Future multi-institutional implementations in diverse regulatory environments are necessary to validate and refine this framework across a broader range of contexts.

Data Source Dependencies and Bias

Integration with international bibliometric databases (Scopus and Web of Science) introduces systematic bias toward English-language publications. Analysis revealed that 78% of indexed publications were in English while Spanish-language outputs represented only 12% of total indexed research visibility, potentially undervaluing locally relevant scholarship and regional research contributions.

The system's dependence on privileged access to Minciencias validation data also limits scalability for institutions without similar access arrangements. University leaders must be critically aware that the 'comprehensive' view provided by the pipeline is shaped by the underlying biases of its global data sources, particularly the risk that institutional strategy may inadvertently prioritize internationally indexed research over other valuable forms of scholarly contribution such as locally relevant Spanish-language publications, books, or knowledge transfer activities. Future iterations should incorporate regional databases such as SciELO and Redalyc to present a more balanced view of institutional research impact.

Technical Implementation Requirements

System deployment requires dedicated technical expertise including Python programming capabilities, database management skills, and Power BI proficiency. Ongoing maintenance demands continuous attention to API changes, data source updates, and algorithm refinement, representing substantial institutional resource commitments beyond initial development costs. Institutions with significantly larger academic divisions populations or more extensive historical research records may require system modifications or enhanced computational resources to maintain acceptable performance levels.

Regarding the 45% higher collaboration rate observed among faculty members teaching across multiple academic divisions, it is important to note that this metric represents a descriptive institutional finding derived from the consolidated dataset, rather than a generalized inferential claim. To address potential statistical uncertainty, a significance test was attempted. However, conducting a rigorous inferential test required linking individual authorship records directly to internal academic division assignments using deterministic exact-string matching.

Due to the naming heterogeneities and unstandardized records that originally motivated this pipeline, this exact linkage was only viable for a sub-sample of 106 faculty members. Within

this verified subset, the difference in inter-group co-authorship rates between members operating in multiple divisions (n=97) and those restricted to a single academic division (n=9) did not reach statistical significance (p=0.78, 95% CI: 0.75-1.23). This highlights a critical constraint: the control group (n=9) is statistically underpowered to support reliable population-wide inferences, and the sub-sample suffers from selection bias toward researchers with fewer naming variants. Consequently, while the 45% descriptive metric accurately reflects the current data snapshot, future iterations must extend the fuzzy-matching.

DISCUSSION

The pipeline presented in this study addresses a practical gap that commercial research information management systems have not fully resolved, particularly in emerging economy contexts operating under dual evaluation frameworks. A comparison with established CRIS platforms clarifies the contribution and its boundaries.

Commercial systems such as Elsevier's Pure, Clarivate's Converis, and Symplectic Elements provide mature, institutionally validated environments for research data management. Pure and Converis offer robust integration with Scopus and Web of Science and support complex organizational hierarchies, while Symplectic Elements excels at harvesting researcher profiles across multiple external sources. However, these platforms share a critical limitation for Colombian and similar Latin American institutions: none offers native integration with national evaluation frameworks such as Minciencias GrupLAC, nor do they accommodate the specific output typologies and point-weighting systems that national funding compliance requires. Institutions adopting these platforms must either maintain parallel manual processes for national reporting or invest heavily in custom connectors costs that are prohibitive for mid-sized universities in emerging economies.

The pipeline developed here takes a complementary rather than competing position. It is not a replacement for enterprise CRIS infrastructure where that infrastructure exists; rather, it demonstrates that targeted, open-source integration can deliver comparable disambiguation and validation performance 94.2% author disambiguation accuracy versus the 60-70% baseline of exact-match approaches at a fraction of the implementation cost, and with direct alignment to national reporting requirements. This positions the methodology as particularly relevant for institutions in contexts where commercial CRIS adoption is constrained by licensing costs, local regulatory specificity, or limited technical infrastructure.

That said, the comparison also reveals genuine limitations of the present approach. Commercial platforms provide longitudinal vendor support, interoperability standards (notably CERIF), and governance frameworks that a single-institution open-source

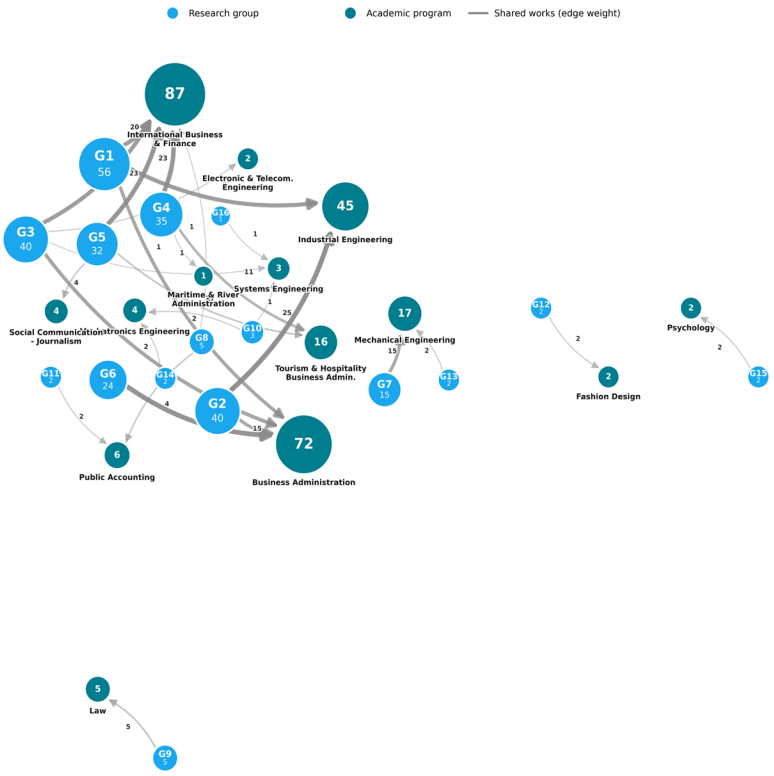


Figure 5: Interdisciplinary Collaboration Network between Research Groups and Academic Programs.

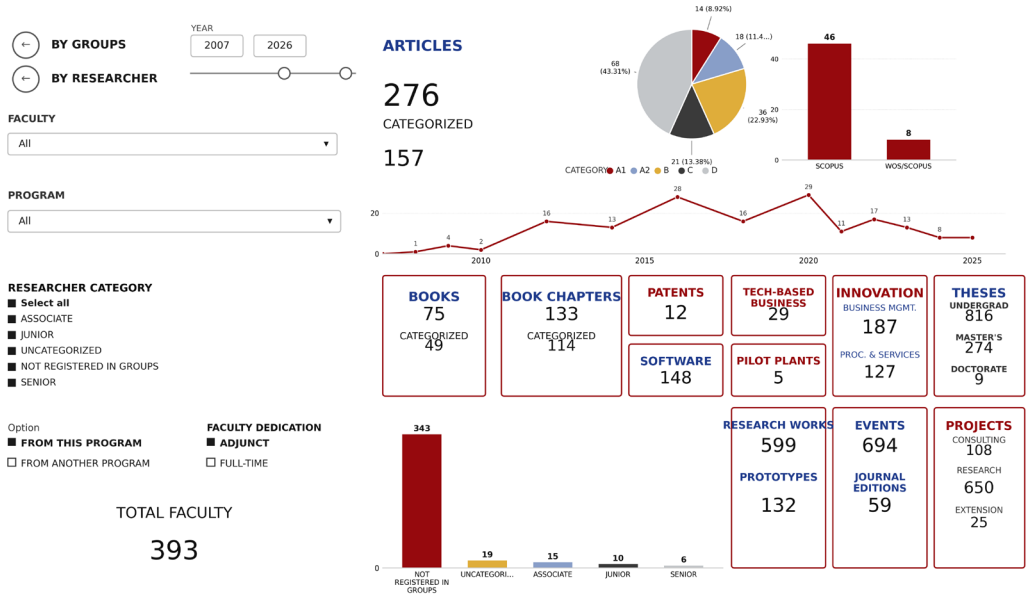


Figure 6: Multilevel Dashboard.

pipeline cannot replicate. The system's dependence on web scraping for GrupLAC data introduces fragility to platform changes, and the fuzzy matching thresholds calibrated for this institution's naming conventions may require significant returning in other contexts. These are not trivial constraints: the 5.8% author disambiguation error rate, while acceptable for strategic planning purposes, could produce meaningful misattributions in high-stakes individual performance evaluations, a use case for which this system was not designed and should not be uncritically applied.

The finding that 78% of indexed publications appear in English, while Spanish-language outputs represent only 12% of visibility, raises a more substantive concern. Systems like Pure and Convers reproduce this same bias by anchoring their metadata quality on Scopus and Web of Science coverage. The difference is that those platforms operate at sufficient institutional scale that the bias is often invisible. A purpose-built pipeline that makes this disparity explicit as this system does creates both a transparency advantage and a responsibility: the dashboard's outputs should be interpreted alongside, instead of, qualitative assessment of locally relevant scholarly contributions. Integration of SciELO and Redalyc in future iterations is not merely a feature enhancement; it is a methodological corrective necessary for equitable institutional evaluation in Latin American contexts.

The combination of automated processing with manual validation provides a practical quality assurance model, though future work should formalize the threshold conditions under which manual review is triggered. The open-source technology approach ensures accessibility for resource-constrained institutions, but sustainability beyond the initial implementation team requires documentation standards and community adoption that remain to be developed. Multi-institutional validation ideally across institutions with different national evaluation frameworks is the most pressing research need before broader transferability claims can be substantiated.

CONCLUSION

This study presented the design, implementation, and validation of a comprehensive data integration pipeline specifically developed for higher education research management. The system successfully addressed critical technical challenges in academic data processing, achieving 99% data coverage while processing 50,000 records within operational time constraints suitable for institutional requirements.

The implementation demonstrated practical solutions for common barriers to institutional research data integration, including author disambiguation (94.2% accuracy), institutional affiliation resolution (96.8% accuracy), and multi-source data validation. These technical achievements provide empirical evidence that sophisticated data integration systems can be successfully implemented in institutional contexts with appropriate methodological approaches and technical resources. Beyond technical performance, the system revealed institutional insights that remained invisible under previous fragmented data management approaches. The identification of collaboration patterns, productivity correlations, and resource allocation dynamics provides empirical foundation for evidence-based institutional decision-making and strategic planning processes.

Because the pipeline was built upon the standardized metrics of the national ministries (Minciencias and MinEducacion), the

system is structurally scalable to other Colombian universities. Furthermore, for similar institutions in other emerging economies facing comparable challenges in research evaluation, the methodology's emphasis on open-source technologies and standard data integration approaches enhances transferability while accommodating local variations in data sources, evaluation frameworks, and institutional structures.

However, significant limitations constrain the broader international applicability of these findings. While nationally scalable, the single institution implementation limits generalizability claims beyond systems modeled similarly to the Colombian framework. Furthermore, dependence on privileged data access and specific technical expertise requirements may restrict adoption by institutions with different resource profiles or regulatory contexts. Future research should focus on multi-institutional validation studies, development of strategies to address bibliometric bias in international databases, and investigation of predictive analytics capabilities for strategic planning support. The integration of natural language processing and machine learning technologies offers promising directions for enhancing system capabilities while maintaining focus on practical utility for institutional research management.

This study serves as a proof-of-concept and a methodological blueprint for other institutions. To advance this field, we issue a twofold call to action:

For Researchers and Developers: We encourage testing and refining this open-source framework, developing standardized connectors, and exploring NLP for automated research classification.

For University Leaders and Administrators: Strategic data integration boosts institutional intelligence. We advocate funding both technical infrastructure and data literacy to ethically enrich the research ecosystem.

This work demonstrates how integrating fragmented data creates strategic analytical capabilities, providing institutions with practical guidance to enhance their research data management.

ABBREVIATIONS

SIS: Student Information Systems; **LMS:** Learning Management Systems; **HRIS:** Human Resources Information Systems; **RIMS:** Research Information Management Systems; **CRIS:** Current Research Information Systems; **LA:** Learning Analytics; **RIM:** Research Information Management; **IR:** Institutional Repositories; **CI:** Confidence Interval; **NLP:** natural language processing; **CERIF:** Common European Research Information Format; **GrupLAC:** Colombia's national research platform for managing researchers' curricula and research information.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

REFERENCES

Altamar, F. B., Polo, L. E. R., & Molina, N. A. (2025). Statistics of sustainable development evolution in Colombia. *World Review of Entrepreneurship, Management and Sustainable Development*, *21*(1), 20-36. https://doi.org/10.1504/WREMSD.2025.143637

Banihashem, S. K., Noroozi, O., van Ginkel, S., Macfadyen, L. P., & Biemans, H. J. A. (2022). A systematic review of the role of learning analytics in enhancing feedback practices in higher education. *Educational Research Review*, *37*, 100489. https://doi.org/10.1016/J.EDUREV.2022.100489

Bergdahl, N., Bond, M., Sjöberg, J., Dougherty, M., & Oxley, E. (2024). Unpacking student engagement in higher education learning analytics: a systematic review. *International Journal of Educational Technology in Higher Education*, *21*(1), 1-33. https://doi.org/10.1186/S41239-024-00493-Y/TABLES/6

Givens, M. (2016). *Keeping up with Research Information Management Systems*. American Library Association. https://www.ala.org/acrl/publications/keeping_up_with/rims

Guerra, J., Ortiz-Rojas, M., Zuñiga Prieto, M. Á., Scheiling García, E., Jiménez Macias, A., Broos, T., Laet, T. De, & Verbert, K. (2020). Adaptation and evaluation of a learning analytics dashboard to improve academic support at three Latin American universities. *British Journal of Educational Technology,* ISSN 0007-1013, 51(4), 973-1001. https://dialnet.unirioja.es/servlet/articulo?codigo=7473677&info=resumen&idioma=ENG

Guzmán-Valenzuela, C., Gómez-González, C., Rojas-Murphy Tagle, A., & Lorca-Vyhmeister, A. (2021). Learning analytics in higher education: a preponderance of analytics but very little learning? *International Journal of Educational Technology in Higher Education*, 18(1), 1-19. https://doi.org/10.1186/S41239-021-00258-X/TABLES/4

Hilliger, I., Ortiz-Rojas, M., Pesántez-Cabrera, P., Scheihing, E., Tsai, Y. S., Muñoz-Merino, P. J., *et al.* (2020). Identifying needs for learning analytics adoption in Latin American universities: A mixed-methods approach. The Internet and Higher Education, 45, 100726. https://doi.org/10.1016/J.IHEDUC.2020.100726

Jones, K. M. L. (2019). Learning analytics and higher education: a proposed model for establishing informed consent mechanisms to promote student privacy and autonomy. International Journal of Educational Technology in Higher Education, 16(1), 1-22. https://doi.org/10.1186/S41239-019-0155-0/METRICS

Kaliisa, R., Jivet, I., & Prinsloo, P. (2023). A checklist to guide the planning, designing, implementation, and evaluation of learning analytics dashboards. International Journal of Educational Technology in Higher Education, *20*(1), 1-22. https://doi.org/10.1186/S41239-023-00394-6/TABLES/2

Krüger, A. K., & Petersohn, S. (2023a). 'I want to be able to do what I know the tools will allow us to do': Practicing evaluative bibliometrics through digital infrastructure. *Research Evaluation*, *31*(4), 475-485. https://doi.org/10.1093/RESEVAL/RVAC009

Krüger, A. K., & Petersohn, S. (2023b). 'I want to be able to do what I know the tools will allow us to do': Practicing evaluative bibliometrics through digital infrastructure. *Research Evaluation*, *31*(4), 475-485. https://doi.org/10.1093/RESEVAL/RVAC009

Ndukwe, I. G., & Daniel, B. K. (2020). Teaching analytics, value and tools for teacher data literacy: a systematic and tripartite approach. International Journal of Educational Technology in Higher Education, 17(1), 1-31. https://doi.org/10.1186/S41239-020-00201-6/FIGURES/6

Schöpfel, J., Prost, H., & Rebouillat, V. (2017). Research Data in Current Research Information Systems. Procedia Computer Science, 106, 305-320. https://doi.org/10.1016/J.PROCS.2017.03.030

Susnjak, T., Ramaswami, G. S., & Mathrani, A. (2022). Learning analytics dashboard: a tool for providing actionable insights to learners. International Journal of Educational Technology in Higher Education, *19*(1), 1-23. https://doi.org/10.1186/S41239-021-00313-7/FIGURES/3

Cite this article: Ramirez-Polo L, Caicedo-Solano N, Bermejo-Altamar F. A Data Integration Pipeline for Research Information Management: Implementation and Performance Assessment in Higher Education Contexts. J Scientometric Res. 2026;15(3):780-92.