Contribution ID: **55** Type: **Presentation**

Challenges and Strategies for Ensuring the Quality and Reliability of Scientific Data at BrCris

Tuesday 14 October 2025 16:00 (11 minutes)

Introduction

The understanding of the development of scientific disciplines, knowledge dissemination, and technological evolution is predominantly informed by analyzing scientific publications, collaborative networks, and patent records. Brazil holds a prominent position in Latin America's scientific production, becoming a key regional player and talent attractor. The growing digitization of academic output has resulted in the availability of extensive datasets for scientific analysis, leading to the establishment of systems known as Current Research Information Systems (CRIS). CRIS aggregates scientific information, facilitating comprehensive data analysis on research ecosystems at various institutional and national levels.

BrCris, modeled after CRIS standards, consolidates information related to Brazilian scientific output—researchers, institutions, publications, and projects—and enables comprehensive bibliometric and scientometric analyses. However, integrating data from diverse sources generates considerable challenges related to harmonization and consistency, compromising analysis accuracy.

Persistent identifiers (e.g., DOI, ORCID) are vital to data accuracy, yet records often lack these identifiers or contain errors, complicating accurate data linkage and author attribution. Therefore, strategies for data disambiguation, deduplication, and metadata validation are paramount for improving BrCris reliability.

Common Data Quality Issues in BrCris

Integrating diverse data sources into BrCris presents several recurrent issues:

1. Record Duplication:

Duplicate records occur when identical authors or publications appear repeatedly due to absent or inconsistent identifiers and metadata discrepancies. This redundancy distorts co-authorship networks and artificially inflates productivity statistics, impairing scientific analyses.

2. Inconsistent Persistent Identifiers:

Incorrect or missing DOI, ORCID, and ISSN identifiers complicate normalization processes, author recognition, and citation attribution. Approximately 7% of Brazilian researchers on Plataforma Lattes have registered ORCID IDs, limiting the interoperability and accuracy of author-publication associations.

3. Entry and Standardization Errors:

Manually entered textual fields (authors'names and institutions) contain variations, inconsistent abbreviations, and typographical errors. For example, Universidade Federal de Minas Gerais can appear in multiple formats, complicating aggregation and consistency. These variations significantly impair accurate network analyses and institutional evaluations.

4. Outdated and Incomplete Records:

Many records become outdated due to researchers' negligence in updating profiles or linking publications to identifiers. Incompleteness in repository metadata indexing further reduces visibility, impacts accurate performance indicators, and weakens data-driven policy formulation.

These challenges profoundly influence the reliability and accuracy of scientific performance assessments, negatively affecting policymaking and resource allocation processes based on BrCris data.

Impact on Bibliometric and Scientometric Analysis

Data quality directly impacts bibliometric and scientometric accuracy, crucial for scientific evaluation, institutional assessments, and policy formulation. Duplicate records inflate productivity indicators, while inconsistent standardization underestimates scientific output, distorting accurate assessments of research impact.

Mismanaged co-authorship networks, caused by inadequate disambiguation, lead to fragmentation or artificial collaborations, misrepresenting institutional performance. Inconsistent article indexing impacts accurate

citation attribution, compromising bibliometric analyses, such as citation counts and impact factor measurements.

The integration of multiple sources, each with varying metadata standards, exacerbates interoperability issues, complicating accurate identification of international collaborations. Such inaccuracies can lead to misguided resource allocation, skewing policy decisions and affecting institutional funding equitably.

Strategies for Improving Data Quality

To mitigate these challenges, several technical solutions have been implemented in BrCris:

• Record Deduplication:

Applying machine learning algorithms, heuristic rules, and clustering techniques efficiently identifies and merges duplicate records. OpenRefine and similar tools help standardize metadata, improving database accuracy.

• Researcher and Institutional Disambiguation:

Utilizing persistent identifiers (ORCID for researchers and ROR for institutions) substantially reduces ambiguity and enhances tracking capabilities. Integrations between Lattes and OpenAlex demonstrate effective correlation between author identities and institutional affiliations.

• Data Validation and Certification:

Implementing international metadata standards (CERIF) and automated API-based validation procedures contributes significantly to enhancing metadata consistency. Cross-referencing data from trusted platforms (Oasis.Br, OpenAIRE Research Graph) supports automatic correction, ensuring reliability.

• Integration with Bibliometric Tools:

Strengthening integration with analytical tools (VOSviewer, Gephi, Visão) enables systematic anomaly detection and data quality monitoring through interactive dashboards, identifying inconsistencies for proactive resolution within BrCris.

The continuous adoption and refinement of these strategies significantly bolster BrCris's credibility and enhance the precision and robustness of its bibliometric and scientometric analyses.

Concluding Remarks

Data quality is fundamental for BrCris effectiveness, reliability, and usefulness in scientific evaluation. Implementing robust strategies for deduplication, disambiguation, and metadata validation is essential to ensuring data integrity.

BrCris has significant potential as a reliable scientific information ecosystem if systemic challenges are adequately addressed. Adopting international metadata standards and persistent identifiers enhances global interoperability, data synchronization, and credibility.

Recent studies demonstrate data quality directly affects metrics such as productivity, scientific impact, and research collaboration. Innovative technological solutions such as machine learning and semantic analysis provide viable pathways for further enhancing BrCris capabilities.

Beyond data management, BrCris represents an initiative to strengthen Brazil's scientific information infrastructure, promoting greater transparency and data accessibility. Improvements in data quality not only enhance domestic scientific credibility but also facilitate international collaboration and recognition.

Despite significant advancements, persistent challenges remain, especially concerning intentional or unintentional errors and misuse of identifiers. Issues like deliberate duplicate entries and incorrect authorship attribution underscore the importance of complementing automated solutions with active community involvement and institutional policies promoting transparency and responsible data management.

Ultimately, maintaining data quality within BrCris demands continuous technological and methodological improvements combined with community participation, ensuring the accuracy, reliability, and global competitiveness of Brazilian scientific information.

Primary authors: SOUZA, MARCEL (IBICT); Mr RODRIGUES, Thiago (Ibict); Mr SEGUNDO, Washington (Ibict)

Presenter: SOUZA, MARCEL (IBICT)

Session Classification: Presentations Session 5: Rigorous, responsible and reproducible science in the era of FAIR data and AI / Infrastructures to Support Data-Intensive Research

Track Classification: SciDataCon2025 Specific Themes: Open research through Interconnected, Interoperable, and Interdisciplinary Data