The Role of NLP in the Digital Transformation of Science: An Emerging and Innovative Practice in Research Data Management

Cibele e Silva Souza[a], Monica Dantas[b]

[a]*cbl.ssouza@gmail.com*
[b]*satnadmonica@gmail.com*

Abstract:

The exponential growth of scientific production and the heterogeneity of research data pose complex challenges to their organization, curation, preservation, and reuse. In this context, Natural Language Processing (NLP) emerges as a strategic technology to support Research Data Management (RDM), fostering the alignment of institutional practices with the FAIR principles (Findable, Accessible, Interoperable, Reusable) and the consolidation of open science.

This paper results from a systematic literature review conducted in the Scopus database, from which a corpus of articles articulating NLP, RDM, and text mining techniques was derived. The analysis revealed that most applications are concentrated in the fields of biomedicine and biocuration, while also identifying initiatives in information science, digital curation, and institutional repositories. This mapping not only highlighted the areas of greatest concentration but also uncovered conceptual and methodological gaps, indicating the need to further integrate NLP with RDM practices.

The discussion will address the potential of NLP to automate processes such as metadata annotation, entity extraction, semantic classification, and trend analysis, as well as the technical, ethical, and institutional challenges of its adoption, including algorithmic transparency, system interoperability, and skills development.

By positioning NLP as an emerging and innovative practice in supporting RDM, this contribution seeks to raise awareness among science managers, librarians, data curators, and researchers of its transformative potential. It argues that, while relevant experiences exist, the literature remains incipient at the intersection of NLP and RDM, opening space for new research agendas and for the institutional consolidation of open science.

Keywords: Natural Language Processing (NLP); Research Data Management (RDM); Text Mining / Text Analytics; FAIR Principles

---

Project/Infrastructure/Initiative Designation
Systematic review on the use of Natural Language Processing (NLP) as an emerging and innovative practice for Research Data Management (RDM), based on literature indexed in Scopus and Web of Science.

Target Audience
Repository and data center managers; science managers; librarians; data curators; information technology specialists; data scientists; researchers interested in open science and research data management practices.

Useful Web Links

- FAIR Principles – GO FAIR Initiative
- Scopus – Scientific Database
- Web of Science – Clarivate