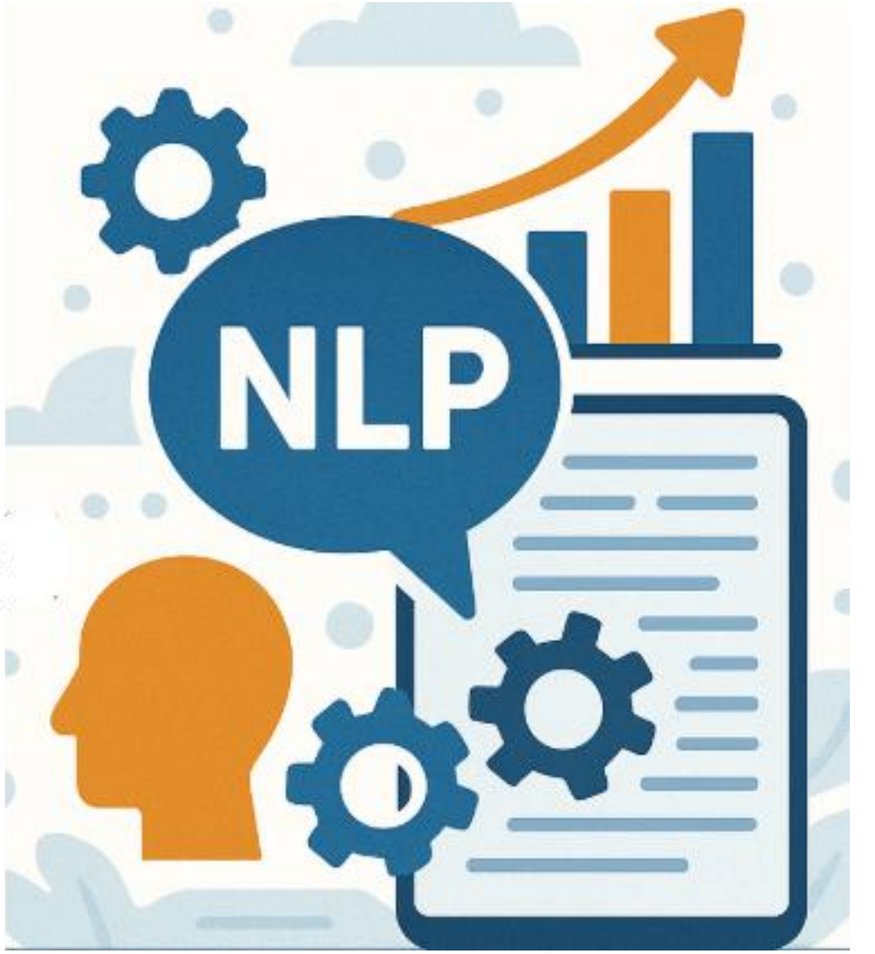
FÓRUM
GESTÃO DE DADOS
DE INVESTIGAÇÃO

The role of NLP in the digital transformation of science:

An emerging and innovative practice in research data management

Cibele e Silva Souza



O papel do PNL na transformação digital da ciência

- O crescimento exponencial da produção científica e a **heterogeneidade** dos ecossistemas de dados de pesquisa criaram **desafios complexos** para a curadoria, preservação e reutilização de dados.

- Nesse contexto, o Processamento de Linguagem Natural (PLN) surge como **estratégica para apoiar** a Gestão de Dados de Pesquisa (GDP).

- O PNL contribui para **alinhar as práticas institucionais** com os Princípios FAIR (Localizável, Acessível, Interoperável, Reutilizável) e para promover a Open Science.

Lacunas

- Apesar dos avanços dos princípios FAIR e da Ciência Aberta, ainda **existem lacunas** na **interoperabilidade semântica e na automação** da curadoria de dados.
- O **gerenciamento manual de metadados não é mais sustentável** em ecossistemas digitais de grande escala.

Objetivos

- **Mapear** e analisar o papel do PNL (Natural Language Processing) na Gestão de Dados de Investigação.
- **Identificar tendências** temáticas, áreas disciplinares e redes de colaboração.
- Discutir os **desafios conceptuais** e técnicos para a integração do PNL e dos princípios FAIR (Interactive Information and Communication Access) nas infraestruturas científicas.

Metodologia

Revisão sistemática da literatura e análise bibliométrica

Database

Scopus
Web of Science
Periodo: 2014–2025

Search descriptors

"Natural Language Processing", *"Research Data Management"*, *"FAIR Principles"*, and related terms (*text mining*,

Critério

Open access, English language, peer-reviewed.

Ferramenta

Bibliometrix package in **R (version 4.3.2)**.

Final *corpus*

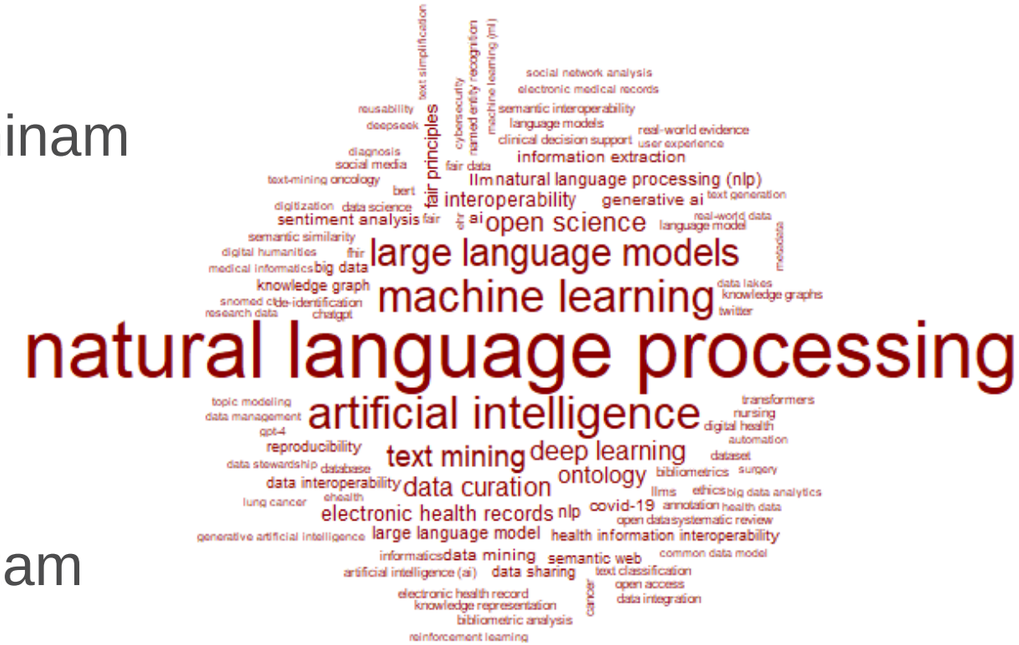
510 publications directly linking NLP, RDM, and FAIR frameworks.

FÓRUM
GESTÃO DE DADOS
DE INVESTIGAÇÃO

Principais tendências identificadas:

PREDOMÍNIO:

Os domínios **biomédicos** e de **biocuradoria** dominam as aplicações de PNL *(extração de entidades, anotação de metadados).*



EMERGEICiA :

- A **ciência da informação** e os *repositórios institucionais* são campos emergentes que utilizam PNL para a automação FAIR.

- **Crescimento notável entre 2021 e 2025**, paralelo à ascensão dos **Modelos de Linguagem** (LLMs).

- A curadoria digital e a integração baseada em ontologias destacam o papel da PNL no enriquecimento semântico.

FÓRUM
GESTÃO DE DADOS
DE INVESTIGAÇÃO

Desafios identificados:

Principal conclusão: A adoção do PNL requer não só ferramentas tecnológicas, mas também **governaça institucional, transparência e capacitação humana.**

Fragmentação semântica

Fragmentação semântica em vocabulários disciplinares.

Interoperabilidade

Falta de ontologias interoperáveis e padrões de metadados.

Lacunas nas políticas

Lacunas nas políticas institucionais sobre IA e PNL em fluxos de trabalho de dados de pesquisa.

Escassez de competências

Escassez de profissionais qualificados que combinem curadoria digital e linguística computacional.

Conclusões

O PNL (Processamento de Linguagem Natural) está se tornando uma **infraestrutura cognitiva da ciência digital**, dando suporte à automação, interoperabilidade semântica e reprodutibilidade.

Mais do que uma ferramenta técnica, o PNL representa **um mediador** entre o conhecimento humano e os sistemas de dados.

A integração de PNL (Processamento de Linguagem Natural), RDM (Medida de Dados de Pesquisa) e FAIR (Acesso Justificado, Interoperabilidade e Redução de Infraestrutura) está criando uma **fronteira interdisciplinar transformadora** que possibilita ecossistemas de pesquisa mais inteligentes, conectados e transparentes.

FÓRUM
GESTÃO DE DADOS
DE INVESTIGAÇÃO

- **Cibele Souza,**
- Phd Corvinus University of Budapest
- Researcher at Cetrad / UTAD

cbl.ssouza@gmail.com

Obrigada!

Referencias

Aria, M., & Cuccurullo, C. (2017). *bibliometrix: An R-tool for comprehensive science mapping analysis*. **Journal of Informetrics, 11**(4), 959–975. https://doi.org/10.1016/j.joi.2017.08.007

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). *On the dangers of stochastic parrots: Can language models be too big?* **Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT)**, 610–623. https://doi.org/10.1145/3442188.3445922

Borgman, C. L. (2015). *Big Data, Little Data, No Data: Scholarship in the Networked World*. MIT Press.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., … Liang, P. (2021). *On the opportunities and risks of foundation models*. **arXiv preprint arXiv:2108.07258**. https://doi.org/10.48550/arXiv.2108.07258

Candela, L., Manghi, P., & Pagano, P. (2023). *FAIR-by-design digital repositories: From open data to interoperable science*. **Data Intelligence, 5**(1), 12–34. https://doi.org/10.1162/dint_a_00153

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of deep bidirectional transformers for language understanding*. **Proceedings of NAACL-HLT 2019**, 4171–4186. https://doi.org/10.48550/arXiv.1810.04805

Dunning, A. C., de Smaele, M. M., & Böhmer, J. K. (2017). *Are the FAIR data principles fair?*. **International Journal of Digital Curation, 12**(2), 177–195. https://doi.org/10.2218/ijdc.v12i2.567

European Commission. (2022). *Turning FAIR into reality: The final report and action plan from the European Commission expert group on FAIR data*. Publications Office of the European Union. https://doi.org/10.2777/54599

Gupta, S., Zaveri, A., & Lehmann, J. (2023). *Semantic enrichment of research data using AI and NLP methods: A systematic review*. **Journal of Web Semantics, 79**, 100774. https://doi.org/10.1016/j.websem.2023.100774

Hey, T., Tansley, S., & Tolle, K. (Eds.). (2009). *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Research.

Lamprecht, A.-L., Garcia, L., Kuzak, M., Martinez, C., Arcila, R., Martin Del Pico, E., … Hooft, R. (2020). *Towards FAIR principles for research software*. **Data Science, 3**(1), 37–59. https://doi.org/10.3233/DS-190026

Leonelli, S. (2018). *Data governance is key to interpretation: Reconceptualizing open data and data reuse*. **History and Philosophy of the Life Sciences, 40**(2), 45. https://doi.org/10.1007/s40656-018-0196-7

Mons, B. (2022). *Data Stewardship for Open Science: Implementing FAIR Principles*. CRC Press.

Mons, B., Neylon, C., Velterop, J., Dumontier, M., da Silva Santos, L. O. B., & Wilkinson, M. D. (2020). *Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud*. **Information Services & Use, 40**(1–2), 13–21. https://doi.org/10.3233/ISU-200084

Neves, M., & Ševa, J. (2022). *Natural language processing in scholarly data curation: From metadata enrichment to FAIR data infrastructures*. **Briefings in Bioinformatics, 23**(2), bbac037. https://doi.org/10.1093/bib/bbac037

Pasquetto, I. V., Randles, B. M., & Borgman, C. L. (2023). *Data as social artifact: Toward a theory of open data practice*. **Information, Communication & Society, 26**(2), 213–233. https://doi.org/10.1080/1369118X.2021.1978917

Peters, I., Konkiel, S., & Haustein, S. (2022). *Open research data and information science: From curation to metrics*. **Journal of Documentation, 78**(6), 1234–1256. https://doi.org/10.1108/JD-11-2021-0245

Tenopir, C., Rice, N. M., Allard, S., Baird, L., & Christian, L. (2020). *Data sharing, management, use, and reuse: Practices and perceptions of scientists worldwide*. **PLOS ONE, 15**(3), e0229003. https://doi.org/10.1371/journal.pone.0229003

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., … Mons, B. (2016). *The FAIR Guiding Principles for scientific data management and stewardship*. **Scientific Data, 3**, 160018. https://doi.org/10.1038/sdata.2016.18

Zaveri, A., Rula, A., Maurino, A., & Hitzler, P. (2021). *FAIR and AI-ready data: The role of metadata and semantics in data reuse*. **Data Intelligence, 3**(3), 343–366. https://doi.org/10.1162/dint_a_00108