

DE ARTIGOS A TAGS

Tagbots Semânticos Potencializados por IA Generativa

FROM ARTICLES TO TAGS

Semantic Tagbots Powered by GenAI

Wesley da Silva Santos

Afya

E-mail: wesley.ssantos@afya.com.br

Mariana Cardoso Moura

Afya

E-mail: mariana.cardoso@afya.com.br

Vanessa Maria Barbosa

Afya

E-mail: vanessa.maria@afya.com.br

Patricia Nogueira Morgado

Afya

E-mail: patricia.morgado@afya.com.br

RESUMO

A gestão da informação e a organização do conhecimento enfrentam o desafio de transformar conteúdo textual não estruturado em metadados consistentes e recuperáveis, semanticamente vinculados a vocabulários específicos de domínio. Em contextos de saúde digital, esse desafio é intensificado pela variação terminológica, por menções secundárias a assuntos e por diferentes níveis de relevância informacional. Avanços recentes em inteligência artificial generativa (IA generativa) oferecem alternativas promissoras para a extração, classificação e padronização semântica de entidades textuais. Este artigo apresenta uma abordagem metodológica para desenvolver tagbots semânticos apoiados por IA generativa, avaliada por meio de um estudo de caso no Portal Afya. O problema aplicado consiste em estabelecer vínculos semânticos entre os artigos do portal e uma taxonomia interna de doenças baseada na Classificação Internacional de Doenças, 11ª Revisão (CID-11). A metodologia adota a Design Science Research, compreendendo ciclos iterativos de construção, refinamento e avaliação do artefato, com melhorias sucessivas nos critérios de extração, pontuação, seleção e padronização de tags. Os resultados indicam que a combinação de extração automatizada, critérios explícitos

XII Congresso Internacional em Tecnologia e Organização da Informação

Guimarães, 24 de junho a 10 de julho de 2026

ISSN: 2965-4130

DOI: 10.5281/zenodo.23100200

de relevância e aproximação a uma taxonomia de domínio pode gerar metadados mais consistentes para a recuperação da informação. Como contribuição, o estudo propõe um processo metodológico adaptável a outros contextos informacionais que exigem a organização semântica de conteúdo textual.

Palavras-chave: gestão da informação; organização do conhecimento; inteligência artificial generativa; estudo de caso; design science research.

ABSTRACT

Information management and knowledge organization face the challenge of transforming unstructured textual content into consistent, retrievable metadata semantically linked to domain-specific vocabularies. In digital health contexts, this challenge is intensified by terminological variation, secondary subject mentions, and varying levels of informational relevance. Recent advances in generative artificial intelligence (GenAI) offer promising alternatives for the extraction, classification, and semantic standardization of textual entities. This paper presents a methodological approach for developing semantic tagbots supported by GenAI, evaluated through a case study on the Afya Portal. The applied problem consists of establishing semantic links between portal articles and an internal disease taxonomy based on the International Classification of Diseases, 11th Revision (ICD-11). The methodology adopts Design Science Research, comprising iterative cycles of artifact construction, refinement, and evaluation, with successive improvements in the criteria for extraction, scoring, selection, and standardization of tags. Results indicate that combining automated extraction, explicit relevance criteria, and approximation to a domain taxonomy can generate more consistent metadata for information retrieval. As a contribution, the study proposes a methodological process adaptable to other informational contexts requiring the semantic organization of textual content.

Keywords: information management; knowledge organization; generative artificial intelligence; case study; design science research.

REFERÊNCIAS

- ARONSON, A. R. Effective mapping of biomedical text to the UMLS metathesaurus: The MetaMap program. *Proceedings of the AMIA Symposium* p. 17–21, 2001.
- BROWN, T. B.; MANN, B.; RYDER, N. *et al.* Language models are few-shot learners. *In: Advances in Neural Information Processing Systems*. vol. 33, p. 1877–1901, 2020.

HODGE, G. Systems of knowledge organization for digital libraries: Beyond traditional authority files. Tech. rep., Council on Library and Information Resources, Washington, DC, 2000.

LEAMAN, R.; ISLAMAJ DOĞAN, R.; LU, Z. DNorm: Disease name normalization with pairwise learning to rank. **Bioinformatics**, v. 29, n. 22, p. 2909–2917, 2013. DOI: <https://doi.org/10.1093/bioinformatics/btt474>.

LEAMAN, R.; LU, Z. TaggerOne: Joint named entity recognition and normalization with semi-markov models. **Bioinformatics**, v. 32, n. 18, p. 2839–2846, 2016. DOI: <https://doi.org/10.1093/bioinformatics/btw343>.

LEWIS, P.; PEREZ, E.; PIKTUS, A. *et al.* Retrieval-augmented generation for knowledge-intensive NLP tasks. *In: Advances in Neural Information Processing Systems*. vol. 33, p. 9459–9474, 2020.

MARCHIORI, P. Z. A ciência e a gestão da informação: compatibilidades no espaço profissional. **Ciência da Informação**, v. 31, n. 2, p. 72–79, 2002.

NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY. **Artificial intelligence risk management framework**: Generative artificial intelligence profile. National Institute of Standards and Technology: Tech. Rep. NIST AI 600-1, 2024. DOI: <https://doi.org/10.6028/NIST.AI.600-1>.

PEFFERS, K.; TUUNANEN, T.; ROTHENBERGER, M. A. *et al.* A design science research methodology for information systems research. **Journal of Management Information Systems**, v. 24, n. 3, p. 45–77, 2007.

SCHNEIDER, E. T. R. de S. J. V. A.; KNAFOU, J. *et al.* BioBERTpt - a Portuguese neural language model for clinical named entity recognition. *In: Proceedings of the 3rd Clinical Natural Language Processing Workshop*. p. 65–72. Association for Computational Linguistics, 2020.

WORLD HEALTH ORGANIZATION. ICD-11: International classification of diseases 11th revision, 2026. Disponível em: <https://icd.who.int/>.