Journal of Digital Terminology and Lexicography
Vol. 1, No. 1 (2025) DOI: 10.25430/pupj.jdtl.1752566034

Terminology-Augmented Generation (TAG): Foundations, Use Cases, and Evaluation Paths

Giorgio Maria Di Nunzio

Department of Information Engineering, University of Padua, Via Gradenigo 6a, Padova, 35131, Italy.

Contributing authors: giorgiomaria.dinunzio@unipd.it;

Abstract

This position paper introduces the concept of *Terminology-Augmented Generation* (TAG) as a new paradigm for integrating curated terminology resources into generative AI workflows. Inspired by but distinct from Retrieval-Augmented Generation (RAG), TAG emphasizes structured knowledge, multilingual precision, and expert-defined term usage as key drivers for high-quality, domain-sensitive language generation. We examine the architectural motivations for TAG, contrast it with RAG in terms of control, explainability, and accuracy, and outline use cases relevant to terminology work—such as term extraction, multilingual alignment, and automatic definition generation. By aligning TAG with ongoing evaluation initiatives, including CLEF's SimpleText and BioASQ GutBrain tasks, as well as earlier efforts like TermEval 2020, we argue that TAG is not only theoretically grounded but practically measurable. We further discuss speculative extensions such as Lexical-Augmented Generation (LAG) and the importance of interoperability for implementing both the TAG approach and its evaluation.

Keywords: Terminology-Augmented Generation, TAG, Generative AI, Evaluation

1 Introduction

The rapid development of generative artificial intelligence (GenAI) presents a great opportunity for science and, more specifically, for the field of terminology (Stokel-Walker & Van Noorden, 2023). Large language models (LLMs) have demonstrated an impressive capacity to generate syntactically coherent, fluent, and contextually appropriate text across a variety of domains. However, their outputs often suffer from terminological inconsistencies, factual inaccuracies, and a lack of transparency with respect to linguistic or conceptual sources.[1] A possible approach to mitigate this problem is Retrieval-Augmented Generation (RAG), one of the most influential paradigms in generative AI, which combines neural generation with real-time access to external document repositories (Arslan, Ghanem, Munawar, & Cruz, 2024). Originally introduced as a way to overcome the limitations of static model knowledge, RAG allows systems to ground their outputs in contextually retrieved text passages. While powerful in general-purpose applications such as question answering and open-domain dialogue, RAG also has notable limitations: the retrieved content is often noisy, or simply not relevant, and not optimized for terminological precision or multilingual consistency. These limitations are especially problematic in high-stakes domains such as law, medicine, and science, where accurate and traceable terminology is critical.

[1] See, for example, the evaluation of text generated for machine translation in the Workshop on Statistical Machine Translation (WMT) forum, https://aclanthology.org/venues/wmt/

Table 1 Kaleidoscope proposal for key infrastructural capabilities of termbanks for supporting generative systems

Structured Data Model	Termbanks have a precise data model that stores relevant information in a clean, structured, and accessible way.
Precise Access	Classical terminology methods enable deterministic access through exact search, filtering options, and controlled retrieval.
Flexible Formats	LLMs require lightly structured data. Termbanks can generate such formats, including Markdown, JSON, and prose.
Real-Time API Access	Terminology searches can be performed rapidly and directly via APIs, avoiding slower access to previously embedded content.

The problems that both GenAI and RAG show have actually sparked growing interest among terminologists, lexicographers, and domain experts[2] in developing structured approaches to guide and evaluate generative systems using curated terminology resources. In fact, despite its shortcomings, RAG offers architectural insights that could be fruitfully adapted to the terminological context. Its modular design, separating retrieval from generation, suggests a way to insert controlled terminological access into LLM pipelines by means of elaborated prompt engineering (Chaubey, Tripathi, Ranjan, & Gopalaiyengar, 2024). Moreover, its ability to dynamically adapt to external knowledge sources points to the possibility of leveraging termbanks and lexical databases via APIs in real time. These connections have inspired a growing community of practitioners to explore the idea of Terminology-Augmented Generation (TAG).

The term TAG has recently begun circulating in the terminology community, notably through public communications by Kara Warburton[3] and Klaus Fleischman. [4] TAG was very likely introduced for the first time in a blog post around 2024 by Fleischman himself at Kaleidoscope.[5] However, at least initially, TAG was a great intuitive keyword that resonated with RAG but it lacked a consistent definition, theoretical foundation, or implementation framework. More recently, at the Multilingual Digital Terminology Today (MDTT 2025) conference, the Kaleidoscpe team presented the first research paper with the acronym TAG in it and with a clearer specification of what TAG could be (Lackner, Vega-Wilson, & Lang, 2025), also following their previous efforts made to specify the main elements of TAG itself (see Table 1).[6]

While the term Terminology-Augmented Generation appears to echo Retrieval Augmented Generation in form, their operational foundations diverge significantly. RAG is performed on the retrieval of large-scale unstructured text chunks at generation time, relying on contextual similarity in vector space and often yielding opaque or underspecified sources. In contrast, the TAG paradigm, when grounded in termbanks, leverages precise, structured data models with deterministic access mechanisms. Instead of retrieving loosely relevant paragraphs, TAG systems can extract formally defined concepts, multilingual equivalents, and controlled relationships via real-time API queries. These termbank resources offer both machine-readable formats and filtered, curated entries, enabling generation that is transparent, terminologically faithful, and explainable. Rather than mimicking RAG's architecture, a well-defined TAG model should be understood as a complementary paradigm built on the strengths of terminological infrastructures.

In this paper, we aim to clarify the notion of TAG by contrasting it with RAG, and articulating what a terminology-aware generative system should look like. Our contribution is divided in four parts: first, in Section 2, we provide a conceptual foundation for TAG by analyzing its potential architecture, data sources, and integration patterns with LLMs. Then, in Section 3, we survey related evaluation initiatives, such as the CLEF 2024 SimpleText track and the CLEF 2025 BioASQ GutBrain pilot, that offer concrete mechanisms for assessing the quality of generation with respect to term extraction, definition generation, and relation identification. In Section 4, we consider the role of lexical resources and propose a complementary paradigm of Lexical-Augmented Generation (LAG), aimed at controlling lexical variation and stylistic output. Finally, in Section 5, we give some concluding remarks for establishing shared evaluation benchmarks and infrastructure for TAG.

[2] In this paper, by "domain experts", or just "experts", we refer to professionals with deep, field-specific knowledge, such as medical practitioners, legal scholars, or biodiversity researchers, who contribute to specialized terminology and lexicography.

[3] https://www.linkedin.com/posts/karawarburton_genai-terminology-ai-activity-7263315028143939585-XKhq/

[4] https://www.linkedin.com/posts/klauskaleidos_genai-terminology-ai-activity-7263483874716905472-HsNb/?utm_source=share&utm_medium=member_android

[5] https://kaleidoscope.at/en/blog/ai-and-terminology/

[6] There is, however, a previous paper written in German that presents the idea of Terminology Augmented Generation, https://aktuelles.dttev.org/veranstaltungen/dtt-symposion-2025/DTT2025_Sa04_Fleischmann-Lang.pdf

2 From RAG to TAG

RAG is a prominent architecture in the field of LLMs that combines the strengths of neural generation with external knowledge retrieval. RAG systems augment generation by dynamically querying external document collections at inference time. This enables the system to draw on up-to-date or domain-specific information that may not be embedded in the model's training data.

In order to give just a flavor of how LLMs changed radically the world of Natural Language Processing and Information Retrieval, we need to make a step back. Before the emergence of LLMs and RAG, information retrieval systems largely relied on "sparse retrievers" which functioned through keyword matching. These systems index documents based on the frequency of exact word occurrences (and more elaborate statistical functions), favoring literal overlap between the user's query and candidate documents (Bailey, Moffat, Scholer, & Thomas, 2017; Di Nunzio & Vezzani, 2022; Marchesin, Di Nunzio, & Agosti, 2021). While efficient, sparse retrievers struggled with synonymy and semantic variation. The advent of "dense retrievers" marked a significant shift: instead of matching words, they map both queries and documents into high-dimensional vector spaces using neural network encoders, typically transformer-based models (Gillioz, Casas, Mugellini, & Khaled, 2020). Relevance is then computed through vector similarity, allowing for a more flexible and contextual-oriented retrieval. This innovation laid the groundwork for RAG architectures, where dense retrieval is used to dynamically select contextually relevant passages that guide language generation.[7]

The standard RAG workflow consists of two main stages (Fan et al., 2024): retrieval and generation. First, a dense retriever identifies relevant passages or documents from a large corpus based on the semantic similarity to the input query. These documents are then passed, along with the original prompt, to a generative model (for example, GPT) that produces a response grounded in the retrieved content. This architecture is particularly effective for tasks such as open-domain question answering, summarization, and chatbots, where the accuracy and recency of information are crucial.

Despite its success, RAG also faces several limitations. The retrieved content is not guaranteed to be relevant for the initial query and this may generate hallucinations or irrelevant outputs. Moreover, RAG systems generally do not support fine-grained control over terminology, definitions, or multilingual variants, factors that are critical in high-stakes applications such as healthcare, law, and translation.

These shortcomings motivate the exploration of alternative or complementary paradigms. In contrast to RAG, which prioritizes scalable retrieval from broad sources, TAG leverages structured, curated knowledge from terminology resources such as termbanks. This shift opens the door to more transparent, domain-anchored, and controllable language generation workflows, which we explore in detail in the following sections. We propose that TAG should be defined as a generative architecture that directly integrates specialized knowledge – according to the dual conceptual and linguistic dimensions of terminology science – into the language generation process.. Unlike RAG, which retrieves unstructured text fragments based on vector similarity, TAG interfaces with resources such as multilingual termbanks, ontologies, glossaries, and domain-specific concept systems. These sources are curated by experts and encode not only terms but also natural language definitions, usage contexts, conceptual hierarchies, and interlingual mappings.

Architecturally, a TAG system may comprise several key components:

- A terminology access layer that supports structured queries to terminology resources;
- A filtering and reasoning module that aligns retrieved terminological data with the input context;
- A generation module that conditions its output on the retrieved terms, definitions, and constraints, either through prompt engineering, fine-tuning, or adapter layers;
- A module to support human-in-the-loop workflows, enabling terminologists to verify, correct, or extend term usage dynamically during content generation.

TAG can support a wide range of tasks central to terminological workflows, particularly in domains where precision, multilingual consistency, and expert validation are essential. Unlike traditional NLP approaches, TAG enables generation that is conditioned on structured terminological data, improving both reliability and traceability. Below, we outline several high-impact use cases:

[7] The terms 'sparse retriever' and 'dense retriever' refer to the mathematical concept of vector of numbers with lots of zero values (sparse) or with very few zero values (dense), respectively.

- Term extraction with disambiguation in multilingual corpora: TAG systems can assist in identifying candidate terms across large corpora while leveraging terminological databases to resolve ambiguities. For example, in the medical domain, distinguishing between "stroke" as a cerebrovascular event versus a physical movement is critical; TAG can anchor interpretations using definitions from medical ontologies (e.g., SNOMED CT[8]).
- Automatic generation of concept definitions: TAG can generate or revise definitions that follow domain-specific templates, taking into account hierarchical position, scope notes, and usage contexts. In legal terminology, for instance, TAG can help draft jurisdiction-specific definitions of terms like "contract" or "liability" that are aligned with authoritative sources.
- Relation extraction at conceptual and lexical levels: TAG systems can support the identification of conceptual relations, such as hierarchical links between broader and narrower concepts as well as lexical relations between terms, including term variants or abbreviations. This dual-level approach enables both taxonomic structuring and the harmonization of terminological variants across languages.
- Multilingual term alignment and translation support: TAG can align terms across languages by grounding them in shared conceptual representations and curated multilingual termbanks. This is particularly valuable for translation workflows in domains such as international law or pharmaceutical regulation, where terms must be equivalent and legally compliant across jurisdictions.

These use cases illustrate TAG's potential not just to automate existing terminological tasks, but to enhance them by offering more contextualized, accurate, and user-controllable outputs. We envision TAG as a tool that complements the expertise of terminologists, accelerating their work while maintaining high standards of quality and traceability.

As the next sections will show, evaluation methodologies inspired by shared tasks such as CLEF SimpleText and BioASQ GutBrain offer a path forward for measuring the effectiveness of TAG systems. These initiatives provide concrete ways to assess not only whether a term is correctly used, relatable, and aligned with expert-curated knowledge. By clarifying the architectural foundations and evaluative strategies of TAG, we aim to establish it as a coherent and actionable paradigm for integrating terminological knowledge into generative AI.

3 Evaluating TAG Systems: Alignment With Evaluation Initiatives

To validate the architectural proposal of TAG, it is essential to anchor its development in robust, task-based evaluation frameworks. Initiatives such as the TermEval 2020 shared task have laid crucial groundwork for systematic evaluation of terminology-related NLP tasks (Rigouts Terryn, Hoste, Drouin, & Lefever, 2020). TermEval 2020 focused on monolingual and multilingual term extraction across English, Dutch, French, and German. It provided manually validated gold standards and addressed domain variation, making it highly relevant for TAG systems that must operate across different languages and subject areas. The evaluation of term candidates based on precision, recall, and F1-score remains directly applicable to the quality control of terminologically grounded generation outputs.

Recent shared tasks within the Conference and Labs of the Evaluation Forum (CLEF) provide a fertile ground for this even though the specific aim was not the evaluation of TAG. In particular, the CLEF 2024 SimpleText task on Identify and Explain Difficult Concepts (Di Nunzio et al., 2024) and the CLEF 2025 BioASQ GutBrain Information Extraction task (Martinelli et al., 2025) align naturally with the core objectives of TAG: generating terminologically faithful, domain-specific outputs in multilingual settings.

Furthermore, the SemEval series[9] has also hosted tasks related to semantic relations and definition modeling, including work on hypernym discovery and word sense definition generation. These contribute indirectly to TAG by offering structured benchmarks to assess relation extraction and definition generation, two of the central use cases for TAG.

The OntoLex and W3C community-driven initiatives also encourage RDF-based modeling of term relations, which can inform the knowledge graph components of TAG systems. In this context, the Language, Data and Knowledge (LDK)[10] conference series also plays an important role in shaping the standards and evaluation methodologies for lexical and terminological data.

[8] https://www.snomed.org/
[9] https://semeval.github.io/
[10] https://2025.ldk-conf.org/

Table 2 A proposal for TAG architecture components aligned with CLEF evaluation tasks and metrics

TAG Component	CLEF Task Alignment	Evaluation Metrics
Terminology-Driven Prompt Augmentation	SimpleText (CLEF 2024): plain-language definition generation	BLEU, ROUGE, Definition Adequacy, Simplification Fidelity
Terminology-Gated Decoding	SimpleText (CLEF 2024) and GutBrain (CLEF 2025): enforcement of preferred terms and definitions	Term Fidelity Score, Human Acceptability, Use of Preferred Labels
Terminology-Enriched Retrieval and Generation	GutBrain (CLEF 2025): ontology-guided QA and relation extraction	Precision/Recall for Term Matching, Concept Normalization F1, Relation Extraction Accuracy

Together, all these initiatives demonstrate that the evaluation of terminology resources is not only feasible but increasingly standardized. For TAG to mature into a widely adopted methodology, it must leverage such existing infrastructures while supporting for new metrics tailored to terminology-aware generation.

3.1 What TAG Evaluation Can Look Like: Insights from CLEF Shared Tasks

The Conference and Labs of the Evaluation Forum (CLEF)[11] has long served as a hub for shared task evaluation in multilingual and domain-specific information access. As Generative AI methods begin to intersect with terminology-driven workflows, CLEF's structured, community-driven evaluation campaigns offer an ideal testing ground for validating the effectiveness of Terminology-Augmented Generation (TAG). In particular, recent tasks such as SimpleText (CLEF 2024) and BioASQ GutBrain (CLEF 2025) highlight the growing demand for systems capable of producing high-quality, terminologically consistent outputs in specialized domains like healthcare and science. These tasks not only provide realistic test collections but also define concrete success metrics, such as terminological accuracy, multilingual fidelity, and alignment with expert-authored resources that are directly applicable to TAG systems. By aligning TAG development with these initiatives, we can ensure that future systems are not only technically proficient, but also grounded in real-world expectations of terminology use and quality.

The CLEF 2024 SimpleText task focuses on the generation of plain-language definitions for complex biomedical concepts. Here, systems are evaluated on their ability to simplify without distortion, preserve semantic content, and reflect preferred terminological usage. This provides a direct benchmark for assessing TAG systems that incorporate structured prompt augmentation or terminology-aware decoding. By leveraging curated sources such as the Unified Medical Language System (UMLS),[12] or institutional vocabularies, TAG systems can explicitly inject concept definitions, term variants, and disambiguating contexts into the generation pipeline. Evaluation metrics include BLEU, ROUGE, and it would be important to define additional domain-sensitive metrics.

The CLEF 2025 BioASQ GutBrain task further broadens the scope to ontology alignment, concept normalization, and biomedical relation extraction. This directly supports the evaluation of TAG's terminology-enriched retrieval components, where structured ontologies (e.g., Gene Ontology,[13] MeSH,[14] etc.) are indexed and queried to inform generation. Outputs are evaluated not only in terms of their lexical quality but also their structural correctness within known ontological frameworks. Metrics include Precision and Recall for term alignment, Concept Coverage, and Relation Accuracy.

In Table 2, we tried to draft a preliminary idea that maps each component of the TAG architecture to the corresponding evaluation opportunities provided by CLEF tasks. By building upon these well-defined evaluation initiatives, TAG can be advanced as more than a conceptual alternative to RAG. It becomes a testable, modular paradigm that supports the terminologist's needs across multiple use cases, grounded in empirical performance against gold-standard terminological data.

[11] https://www.clef-initiative.eu/
[12] https://www.nlm.nih.gov/research/umls/index.html
[13] https://geneontology.org/
[14] https://www.ncbi.nlm.nih.gov/mesh/

4 Can Lexical-Augmented Generation (LAG) Exist as Well?

While Retrieval-Augmented Generation (RAG) emphasizes access to unstructured factual content and Terminology-Augmented Generation (TAG) leverages structured domain-specific resources, a third complementary paradigm can be envisioned: *Lexical-Augmented Generation (LAG)*. This approach would guide generative models through fine-grained lexical knowledge, supporting enhanced control over word choice, style, and linguistic appropriateness.

At this stage, the notion of LAG remains speculative, and we introduce it here primarily as food for thoughts. Unlike TAG, which is beginning to take shape around concrete resources and use cases in terminology, LAG does not yet have a clear architectural definition or community consensus. Nevertheless, it prompts useful questions: could fine-grained lexical resources, such as dictionaries, valency lexicons, or usage patterns, be systematically injected into generation workflows to improve stylistic control, register sensitivity, or fluency? If TAG prioritizes conceptual precision, LAG could, in principle, emphasize surface-level elements.

For example, LAG systems may integrate curated lexical resources such as synonym dictionaries, collocation databases, valency frames, or word sense inventories (e.g., WordNet[15] or BabelNet[16]) into the generation process. This is particularly valuable in tasks that require linguistic variation, paraphrasing, simplification, or stylistic transformation. For example, LAG could be used to adapt output to different reading levels, enforce the use of specific lexical items, or ensure idiomatic usage in translation and cross-cultural communication.

5 Conclusions

In this paper, we introduced the concept of *Terminology-Augmented Generation* (TAG) as a new paradigm for integrating curated terminological knowledge into generative AI workflows. Drawing on the limitations of Retrieval-Augmented Generation (RAG) for high-precision, domain-sensitive applications, and taking advantage of seminal works dedicated to TAG in previous months, we tried to give a better formalization to TAG as a complementary approach that prioritizes accuracy, multilingualism, and conceptual clarity. While TAG is still in its early stages of conceptualization, our discussion has highlighted key design features, plausible use cases, and emerging evaluation pathways.

In particular, we emphasized that robust evaluation is essential for establishing TAG as a meaningful and actionable architecture. Shared international evaluation tasks such as CLEF and SemEval alongside community efforts like TermEval and the LDK/OntoLex ecosystem, provide the ideal ground for developing realistic benchmarks. These initiatives offer not only test collections, but also community-driven metrics that can assess the correctness, relevance, and clarity of terminologically enhanced outputs. We also speculated on the potential for related paradigms such as Lexical-Augmented Generation (LAG), which could emphasize stylistic or lexical appropriateness rather than terminological precision. While still hypothetical, LAG helps to frame a broader conversation about how different layers of linguistic knowledge, from raw documents to lexical and terminology resources, can guide and constrain generative systems. In this context, it is worth mentioning the importance interoperability of lexical and terminological datasets as a critical aspect for the effective implementation and evaluation of TAG systems (Vezzani, Di Nunzio, Salgado, & Costa, 2025). This alignment not only facilitates resource reuse and multilingual consistency but also strengthens the foundations for shared tasks, evaluation campaigns, and generative applications. As TAG matures, such convergence will be instrumental in ensuring that terminology resources are both machine-readable and semantically interoperable across platforms and domains.

Ultimately, our goal is to stimulate debate, experimentation, and community convergence around the idea that terminologists should not merely adapt to generative AI but help shape it. By articulating what TAG should be and how it can be evaluated, we hope to provide a foundation for future research, tooling, and shared tasks at the intersection of terminology, lexicography, and natural language generation.

[15] https://wordnet.princeton.edu/
[16] https://babelnet.org/

References

Arslan, M., Ghanem, H., Munawar, S., Cruz, C. (2024). A Survey on RAG with LLMs. *Procedia Computer Science*, *246*, 3781–3790, https://doi.org/10.1016/j.procs.2024.09.178

Bailey, P., Moffat, A., Scholer, F., Thomas, P. (2017). Retrieval Consistency in the Presence of Query Variations. *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 395–404). New York, NY, USA: Association for Computing Machinery. (https://doi.org/10.1145/3077136.3080839)

Chaubey, H.K., Tripathi, G., Ranjan, R., Gopalaiyengar, S.k. (2024). Comparative Analysis of RAG, Fine-Tuning, and Prompt Engineering in Chatbot Development. *2024 International Conference on Future Technologies for Smart Society (ICFTSS)* (pp. 169–172). (https://doi.org/10.1109/ICFTSS61109.2024.10691338)

Di Nunzio, G., Vezzani, F., Bonato, V., Azarbonyad, H., Kamps, J., Ermakova, L. (2024). Overview of the CLEF 2024 SimpleText Task 2: Identify and Explain Difficult Concepts. G. Faggioli, N. Ferro, P. Galuščáková, & A.G.S.d. Herrera (Eds.), *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024)* (Vol. 3740, pp. 3129–3146). Grenoble, France: CEUR. (https://ceur-ws.org/Vol-3740/#paper-306)

Di Nunzio, G.M., & Vezzani, F. (2022). Did I Miss Anything? A Study on Ranking Fusion and Manual Query Rewriting in Consumer Health Search. A. Barrón-Cedeño et al. (Eds.), *Experimental IR Meets Multilinguality, Multimodality, and Interaction* (pp. 217–229). Cham: Springer International Publishing. (https://doi.org/10.1007/978-3-031-13643-6_17)

Fan, W., Ding, Y., Ning, L., Wang, S., Li, H., Yin, D., … Li, Q. (2024). A Survey on RAG Meeting LLMs: Towards Retrieval-Augmented Large Language Models. *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (pp. 6491–6501). New York, NY, USA: Association for Computing Machinery. (https://dl.acm.org/doi/10.1145/3637528.3671470)

Gillioz, A., Casas, J., Mugellini, E., Khaled, O.A. (2020). Overview of the Transformer-based Models for NLP Tasks. *2020 15th Conference on Computer Science and Information Systems (FedCSIS)* (pp. 179–183). (https://doi.org/10.15439/2020F20)

Lackner, A., Vega-Wilson, A., Lang, C. (2025). Terminology Augmented Generation: A Systematic Review of Terminology Formats for In-Context Learning in LLMs. F. Vezzani, G. Di Nunzio, E. Loupaki, G. Meditskos, & M. Papoutsoglou (Eds.), *Proceedings of the 4rd International Conference on Multilingual Digital Terminology Today (MDTT 2025)* (Vol. 3990). Thessaloniki, Greece: CEUR. (https://ceur-ws.org/Vol-3990/#short10)

Marchesin, S., Di Nunzio, G.M., Agosti, M. (2021). Simple but Effective Knowledge-Based Query Reformulations for Precision Medicine Retrieval. *Information*, *12*(10), 402, https://doi.org/10.3390/info12100402

Martinelli, M., Silvello, G., Bonato, V., Di Nunzio, G.M., Ferro, N., Irrera, O., … Vezzani, F. (2025). Overview of GutBrainIE@CLEF 2025: Gut-Brain Interplay Information Extraction. G. Faggioli, N. Ferro, P. Rosso, & D. Spina (Eds.), *CLEF 2025 Working Notes. In press.*

Rigouts Terryn, A., Hoste, V., Drouin, P., Lefever, E. (2020). TermEval 2020: Shared Task on Automatic Term Extraction Using the Annotated Corpora for Term Extraction Research (ACTER) Dataset. B. Daille, K. Kageura, & A.R. Terryn (Eds.), *Proceedings of the 6th International Workshop on Computational Terminology* (pp. 85–94). Marseille, France: European Language Resources Association. (https://aclanthology.org/2020.computerm-1.12/)

Stokel-Walker, C., & Van Noorden, R. (2023). What ChatGPT and generative AI mean for science. *Nature*,

614(7947), 214–216, https://doi.org/10.1038/d41586-023-00340-6

Vezzani, F., Di Nunzio, G., Salgado, A., Costa, R. (2025). When LMF and TMF meet: Towards a Unified Markup Framework (UMF). *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, *31*(1), 72–109, https://doi.org/10.1075/term.00084.vez