

DRAFTING DEFINITIONS FOR EMERGING CONCEPTS AND TERMS UNDERGOING SEMANTIC SHIFT WITHIN THE ARTES KNOWLEDGE BASE: A PROTOCOL FOR INTEGRATING LLMs INTO TERMINOLOGICAL ANALYSIS BY EXPERIMENTAL APPROACH

Naujų sąvokų ir semantiniu poslinkiu pasižyminčių terminų apibrėžčių rengimas ARTES žinių bazės kontekste: didžiųjų kalbos modelių integravimo į terminologinę analizę eksperimentinis protokolas

MOJCA PECMAN

Université Paris Cité

E-mail: mojca.pecman@u-paris.fr

ORCID ID: <https://orcid.org/0000-0002-6753-1936>

Fields of research: ARTES knowledge database, corpus-based terminology, semantic neologisms, drafting terminological definitions, generative AI, prompt engineering for terminology.

<https://doi.org/10.35321/term32-02>

ABSTRACT

This paper presents a protocol for evaluating and integrating generative AI (GenAI) tools in the framework of the terminological analysis of emerging, semantically unstable terms, absent from established term bases. Implemented within the ARTES knowledge base, the protocol supports Master's students in translation at Université Paris Cité, in their task consisting of conducting a terminological analysis required for their dissertation. The study focuses particularly on terms displaying semantic instability and variation, thereby giving rise to semantic neologisms, and on evaluating the effectiveness of GenAI in retrieving existing definitions and drafting new terminological definitions for such terms. A survey of students' GenAI use and an experimental study on the concept of *data pollution* illustrate the approach. Findings show corpus-linguistic tools help structure conceptual knowledge and critically assess GenAI outputs, confirming the need for human oversight. The study proposes a model for prompt construction and evaluation, a systematic process for building a collection of effective prompts, and a methodology that combines LLMs and corpus linguistics' techniques for terminology management.

Copyright © 2025 Mojca Pecman. Published by the [Institute of the Lithuanian Language](#). This is an Open Access article distributed under the terms of the [Creative Commons Attribution Licence](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. // Išleido [Lietuvių kalbos institutas](#). Šis straipsnis yra atviros prieigos, platinamas pagal „Creative Commons“ priskyrimo licencijos sąlygas, leidžiančias neribotai naudoti, platinti ir atkurti turinį bet kokioje laikmenoje, nurodant autorių ir šaltinį.

Received: / **Gauta:** 2025-09-10. **Accepted:** / **Priimta:** 2025-09-22.

KEYWORDS: ARTES knowledge database, corpus-based terminology, semantic neologisms, drafting terminological definitions, generative AI, prompt engineering for terminology.

ANOTACIJA

Šiame straipsnyje pristatomas generatyvinio DI (GDI) įrankių vertinimo terminologinės analizės kontekste ir jų integravimo į terminologinės analizės sistemą, taikomą naujiems, semantiškai nestabiliems, ir pripažintas terminų bazes neįtrauktiems terminams, protokolas. Jis, įgyvendinamas ARTES žinių bazėje, palengvina darbą Paryžiaus universiteto vertimo magistrantūros studentams, atliekantiems terminologinę analizę, reikalingą baigiamiesiems darbams. Tyrime daugiausia dėmesio skiriama terminams, kurie pasižymi semantiniu nestabilumu ir variantiškumu, todėl tai lemia semantinius neologizmus, bei GDI veiksmingumo atrenkant esamas šių terminų apibrėžtis ir rengiant naujas šių terminų apibrėžtis įvertinimui. Metodą iliustruoja atlikta studentų apklausa dėl GDI naudojimo ir eksperimentinis sąvokos *data pollution* (klaidingų duomenų įrašymas) tyrimas. Rezultatai rodo, kad tekstynų lingvistikos įrankiai padeda struktūrizuoti konceptualias žinias, kritiškai įvertinti GDI išvestis ir patvirtina žmogaus atliekamos priežiūros poreikį. Tyrime siūlomas užklausų kūrimo ir vertinimo modelis, sistemingas procesas veiksmingų užklausų rinkiniui kurti bei didžiųjų kalbos modelių ir tekstynų lingvistikos technikas suderinanti terminologijos valdymo metodologija.

ESMINIAI ŽODŽIAI: ARTES žinių bazė, tekstynais grindžiama terminologija, semantiniai neologizmai, terminologinių apibrėžčių rengimas, generatyvinis DI, užklausų kūrimas terminologijai.

INTRODUCTION

In this study, we present the protocol for GenAI-assisted terminological analysis, implemented within the ARTES project¹. The protocol is specifically designed to address emerging concepts that are not yet documented in existing terminological databases but are increasingly used in specialised literature. Emerging concepts must first and foremost be defined. The ARTES project focuses on creating the necessary linguistic resources and knowledge to address such concepts, which often give rise to terminological variation – whether formal or semantic – and are generally absent from established term bases (cf. L’Homme 2024b). In this context, terms are analysed using a corpus-based methodology and recorded in the ARTES knowledge base², which functions as both a term base and a pedagogical tool for teaching terminology to future translators at Université Paris Cité. Moreover, ARTES is actively used by Master’s students as part of their Master's dissertation in terminology. Every year students compile comparable corpora to conduct in-depth terminological analysis and draft term records to support specialised translation.

This study aims to evaluate the contribution of large language models (LLMs) to terminological management in this specific educational context and highlight the importance of conducting a corpus-based analysis prior to interacting with generative artificial intelligence (GenAI) tools. The study also illustrates the contribution of terminology to knowledge engineering and its connection, in this context, with artificial intelligence (Condamines 2022). Using a pilot corpus composed of texts representing various contexts of specialised and semi-specialised communication in English, we analyse the concept of *data pollution* by means of Knowledge-Rich Contexts (KRCs) (Meyer 2001).

¹ Available at: <https://altae.u-pariscite.fr/artes-aide-a-la-redaction-de-textes-scientifiques>.

² Available at: <https://artes.app.univ-paris-diderot.fr/artes-symfony/web/app.php>.

The meaning of the term *data pollution* began to shift in 2018, making it a valuable candidate for studying processes of semantic change and the challenges involved in defining semantic neologisms. While the term is increasingly used in English (alongside its more established counterpart, *digital pollution*, which is well-documented in term bases), *it* has yet to be recorded in major terminological databases with its newly emerging meaning.

This study outlines the phases of an experimental protocol designed to explore and assess the contribution of GenAI tools to terminological work, with particular focus on drafting definitions. The goal is to undertake an informed adaptation of the terminology curriculum in the Master's program for the 2025–2026 academic year. The integration of GenAI tools into term management is also motivated by students' evolving needs identified through a survey whose results are presented in this study. It is also driven by the requirements of the European Master's in Translation (EMT), joined by Master ILTS³ in 2009. The latest version of the EMT Competence framework (2022), covering the period 2023–2028, places particular emphasis on the integration of technologies and the need for future translators to continuously update their skills in translation tools and technologies. In light of recent advances in AI and neural machine translation (NMT), this aspect has become central to current training needs.

The first section of the study is devoted to presenting background context with particular emphases on the interconnectedness between terminology, corpus linguistics and AI. We also present the ARTES framework for terminological analysis and the key step in this process: drafting terminological definitions. We also introduce the experimental approach for integrating LLMs to corpus-based term analysis. The second section presents the methodology for devising a protocol for GenAI integration and the survey conducted among Master's students. The last section provides the analysis of the protocol designed for collecting and drafting terminological definitions by integrating GenAI through critical approach. We also illustrate the final stage of the analysis leading to drafting a term record for the term *data pollution* in the ARTES knowledge base, and the model designed within this protocol for collecting the most efficient prompts for terminology.

TERMINOLOGY, ARTIFICIAL INTELLIGENCE AND THE ARTES KNOWLEDGE BASE

In this section we review the latest evolutions in terminology as a science and practice driven by the progress in AI. We then present the ARTES platform for drafting term records in line with ISO standards and FAIR principles, and the possibilities it offers for interacting with AI. We also present the central role of definition in terminological analysis, particularly in the case of emerging concepts and terms.

³ Master Industrie de la langue et traduction spécialisée (ILTS), UFR EILA, UPCité.

Terminology and Artificial Intelligence

According to Leonardi (2025: 148), the landmarking work by Eugen Wüster, the founder of terminology science, is a “part of a long tradition of interventions in natural languages aiming at improving their representative and communicative efficiency.” Moreover, Leonardi (2025) explains that “this tradition continues in contemporary formalised models developed in Natural Language Processing (NLP) that are at the basis of Artificial Intelligence applications”. Terminology and AI are therefore intrinsically linked. Indeed, the convergence between Terminology and AI can be traced back to 1993 when Didier Bourigault and Anne Condamines set a research group “Terminology and Artificial Intelligence (TIA)”. At the time, terminology-related AI was rooted in knowledge engineering and explorations of the interactions between corpora and terminology by implementation of symbolic approach using patterns for searching candidate terms and conceptual relationships (Meyer *et al.* 1992; Bourigault *et al.* 2001; Condamines, Rebeyrolle 2001; Condamines 2005). Consequently, Terminology and AI share the same goal: knowledge modelling.

Furthermore, the intersection between terminology and AI through their shared goal – knowledge modelling – is rooted in the emergence of corpus linguistics in the late 1980s, which provided access to digital corpora and tools for their exploration. Over time, corpus linguistics became a leading approach in terminology studies (Pecman, Kübler 2022). It has also contributed to the development of the textual approach to terminology (Bourigault, Slodzian 1999), which is often seen as a reaction to the traditional prescriptive methodology of the Wüsterian school (Humbley 2022). This connection between Terminology, Corpus Linguistics and AI is also visible in the use of the terms such as *computational terminology* and *corpus-based terminology*. In the view of Condamines (2022), this historical alignment between terminology and AI pursuing common goals, and using common tools such as corpora or large collections of language data, has progressively eroded, reflecting the shifting paradigms and accelerated development within the field of AI which culminated in 2020s with the introduction of Large Language Models (LLMs). Nevertheless, language data, which serves as a reservoir for knowledge modelling, is another shared element between Terminology, Corpus Linguistics and LLMs.

In approximately thirty years, the joint venture between computational linguists, NLP and IT specialists culminated in the development of the research prototype, ChatGPT. Launched in November 2022, ChatGPT quickly evolved into a widely used application, illustrating the rapid public uptake of artificial intelligence innovations. The following years were characterised by the proliferation of different types of LLMs and GenAI tools: ChatGPT, DeepSeek, Perplexity, Gemini, Claude, etc.

The paradigm shift driven by Artificial Intelligence

The paradigm shift brought by AI is ongoing, broadening the scope and possibilities for both research and teaching, henceforth oriented toward the evaluation and integration of AI. In relation to terminology, specialised translation and corpus linguistics, the focus is on critical assessment of the effectiveness and pitfalls of AI-assistance in linguistic analysis, translation and term management. While most of the works underline the need for critical approach, particularly in the context of specialised languages: e.g. Raus, Mattioda 2024; San Martín 2024; Davies 2025; Kübler, Pecman 2025; the articles highly enthusiastic towards the possibilities offered by AI tools can be found too, like Schryver's (2023) study on the effectiveness of ChatGPT in general language lexicography. At the same time, the research aiming to improve the performance of LLMs models is specifically interested in domain-specific terminology, in parallel resources in particular, which serve as a high-quality training data: e.g. Ballier *et al.* 2021; Bénard *et al.* 2023; Zhu *et al.* 2023.

Moreover, in this context of rapid propagation of AI, the increasing number of publications and conferences invite scientists to join the reflection on the role of AI in science and teaching, in professional and social practices, on ethical and GDPR (General Data Protection Regulation) issues, as well as on the issues related to limited training data for languages of lesser diffusion: e.g. Rastier 2021; Casal, Kessler 2023; Finardi 2023; Lommel 2024. The institutions contribute to the debate by organising important events. Interpreting Europe Conference 2025⁴ invited professionals, industry experts, academics and students to the discussion on the future of the interpretation, as a profession, in relation to advances in AI. At the 2025 Artificial Intelligence Action Summit (“Sommet pour l'action sur l'intelligence artificielle”)⁵ held in Paris, over 60 countries signed for the first time a declaration aimed at promoting trustworthy, sustainable, and inclusive AI. At the same time, media and press amplify the debate by bringing the topic into the public sphere, spurring researchers to reach wider audience and considering a broader range of genres and sources: e.g. Falgas, Robert 2023; Finardi 2023; Davies 2025.

It is therefore noteworthy that progress in AI greatly impacts scientific, academic and educational settings. In the context of translation and language studies, the growing influence of AI is significantly reshaping student behaviour, particularly in how they search for and construct definitions. As noted by Ptasznik and Lew (2025), ChatGPT has sparked debate among lexicographers. In their recent study on the impact of AI among Polish students of English, they found that ChatGPT is widely used and regarded as a valuable tool for writing and translation. The authors report that students also turn to it for inspiration, curiosity, and entertainment. They also stress that the future role of AI in the field remains uncertain. As the technological landscape evolves, so do academic practices, creating a need to revise the educational and methodological frameworks to reflect these changes.

⁴ Available at: <https://knowledge-centre-translation-interpretation.ec.europa.eu/en/events/interpreting-europe-conference-2025>.

⁵ Available at: <https://www.elysee.fr/sommet-pour-l-action-sur-l-ia>.

Consequently, much of the current research in terminology, language resource development and dictionary-making is now focused on examining the paradigm shift introduced by AI tools, along with their critical assessment and potential integration into teaching and research. Notable recent initiatives include a special issue of the journal *Terminology*, “Terminology and AI”, scheduled for 2027 and a 2025 workshop in Paris exploring the role of dictionaries in the age of AI.⁶ As Altameemi (2024: 429) stresses, “<...> the integration of CL [Corpus Linguistics] with ChatGPT holds great potential for the understanding of language in the digital age. <...> In other words, instead of being cautious in applying ChatGPT, linguists should examine the importance of merging CL and ChatGPT. Even linguistic academic programmes should consider the importance of applying technology in the study plan of their degrees. Moreover, it is not only linguists who must take this point into account, but scholars in other fields who should consider these aspects and think about the effective utilisation of technology in studying fields of human knowledge.”

In this context of pressing need for critical reflection on the role of AI and the development of appropriate practices, this study aims to explore the possibilities for integration of GenAI tools into the process of terminological analysis in the context of the ARTES knowledge base framework. The necessity for such an evaluation is all the more crucial as “with the inevitable integration of AI into terminology work, the distinction between human-created and AI-created content will become increasingly blurred”, as pointed by San Martín (2024).

The ARTES knowledge base and AI

The ARTES knowledge base has been used since 2010 for teaching terminology to Master’s students in Translation at the EILA⁷ department (cf. Pecman, Kübler 2011; Kübler, Pecman 2012; Gledhill, Kübler 2015; Pecman 2021). It provides valuable resources for supporting the teaching and research in terminology and specialised translation. Developed by ALTAE⁸ research team, ARTES consists of two platforms, one for collecting terminological and phraseological resources by translation students and the other for querying the database freely online. ARTES is a term base; however, its enhanced, knowledge-oriented approach to terminology and specialised discourses also qualifies it as a knowledge base (Pecman 2018).

The basic tenet of the ARTES framework is that efficient term management relies on the acquisition of specialist knowledge. This can be achieved by conducting onomasiological and semasiological corpus-driven and -based term analysis, to which, in the ARTES framework, is added knowledge or conceptual networks analysis. It should be emphasised

⁶ *Le dictionnaire face à l’essor de la traduction automatique et des intelligences artificielles génératives*, ISIT, Paris, 12th June 2025. Available at: <https://www.isit-paris.fr/le-dictionnaire-face-a-lessor-de-la-traduction-automatique>.

⁷ Available at: <https://u-paris.fr/eila>.

⁸ Available at: <https://altae.u-pariscite.fr>.

that both, corpus-based terminology (Pecman, Kübler 2022) and corpus-based translation studies (CBTS) (Kübler *et al.* 2024), are characterised by a strong grounding in authentic data drawn from corpora, used to address terminological and translation challenges. Furthermore, in terminology, corpora are particularly useful for identifying and analysis the newly coined terms and semantic neologisms.

Furthermore, the ARTES DB is devised in accordance with ISO standards (i.e. ISO 1087:2019, 704:2022, 12620-1:2022, 5078:2025) and the FAIR data principles (findability, accessibility, interoperability and reusability) initiated by Wilkinson *et al.* (2016) and adapted for terminology by Vezzani *et al.* (2023). ARTES provides valuable resources for NMT training (cf. SPECTRANS (Ballier *et al.* 2021; Zhu *et al.* 2023) and MaTOS (Bénard *et al.* 2023) projects). The aligned or parallel data is namely available for five types of items: terms, collocations, definitions (terminological and encyclopaedic), translation notes and subject or domain.

In the current 2025-2026 academic year, the ARTES framework is being extended to exploring LLMs integration into terminological analysis. The protocol is intended to take into account the various phases of term analysis, that is, for identifying terms and terminological variants, finding collocations, equivalent terms, definitional contexts, KRCs and for drafting definitions and notes. To ensure a critical and experimental approach to evaluating GenAI output, emphasis will remain on acquiring knowledge about terms through a corpus-based approach and developing the necessary skills for producing high-quality definitions. In terminological analysis, definitions play central role in the process of knowledge acquisition. Consequently, it is important to explore the effectiveness of GenAI tools for identifying relevant existing definitions and drafting terminological definitions.

Drafting terminological definitions with LLMs and corpus-based approach

With the emergence of terminology as an independent discipline following Wüster's (1968) proposal for drafting specialised dictionaries, the definition assumed a central role as a key element in term analysis. In the introduction to his *English-French dictionary of Machine Tools*, Wüster (1968: 2.15) identifies the definition as the first and foremost element essential for achieving terminological precision. The definition is also the backbone of the onomasiological approach (from concept to linguistic unit) in terminological analysis. Defining specialised concepts constitutes a first step in acquiring both specialised knowledge and terminological expertise. In the Master's programs offered at the EILA department, future specialised translators, interpreters and researchers in Languages for Specific Purposes (LSP) develop this skill through courses on terminology.

Collecting and drafting terminological definitions for the ARTES database is one of the requirements for record design. It consists in drafting an original definition with the purpose to harmonise the data stored in the term base and to provide definitions for previously undefined concepts or the concepts for which only definitional contexts are available. The model for drafting definitions follows ISO standards 1087:2019 and 704:2022,

widely used by terminologists. Students draft definitions by selecting the relevant information provided in definitional contexts and KRC as well as by analysis of semantically related terms and semantic networks. They also collect existing definitions when they are available in reputable term bases (UNTERM, IATE, Termium, etc., cf. Figure 7).

Drafting definitions for newly emerging concepts or neologisms, specifically in relation to the terms exhibiting semantic instability, is highly challenging. In the ARTES project, particular attention is given to situations where the use of a term starts changing and showing a shift in meaning, thereby leading to the formation of a new concept. Evaluating the contribution of LLMs to their analysis is expected to be a complex task, as discussed by San Martín (2024) who observes: “If crafting traditional definitions is labor-intensive, the consideration of contextual and functional constraints makes the task even more time-consuming. This is an important barrier to the creation of flexible terminological definitions. Generative Artificial Intelligence (GenAI) tools, especially those powered by Large Language Models (LLMs) such as ChatGPT, can remove these barriers by reducing the time and effort required to create them. However, the impact of GenAI can extend well beyond this, as it can profoundly transform the methods and purposes underlying the creation and consumption of terminological definitions.”

For our study, we selected the term *data pollution* which we analysed through a corpus-based approach prior to tests presented in this paper on AI-assisted terminological analysis. The term *data pollution* is a valuable candidate term for the purposes of the present study because it exhibits semantic shift and is absent from currently available terminological resources. Moreover, the case study of this term enables a comprehensive assessment of the ARTES framework for term management, the advantages of corpus-based terminological analysis and the efficacy of LLMs in supporting this process.

As is well-known, terms undergoing semantic shift subtly alter the underlying knowledge paradigm, making it particularly difficult to distinguish the newly emerging meaning from the original one. Nevertheless, such terms reflect broader conceptual trends and are of considerable terminological significance (cf. Pecman 2012; 2014). Typically referred to as *semantic* neologisms (as opposed to *formal* neologisms), and sometimes as neosemes (Renouf 2020), this type of term results from a range of linguistic phenomena, such as polysemy and microsenses, which make them particularly challenging to study (L’Homme 2024a; 2024b). As Lombard *et al.* (2023) point out: “novel word senses are called ‘semantic neologisms’ and “they result from semantic extension by means of polysemy”. Moreover, L’Homme (2024a: 216) explains that “polysemy is a prevalent phenomenon with which terminologists are often confronted”, making the management of polysemous terms in terminological *resources* essential “in order to avoid ambiguity in communication”.

The following section presents the general methodology employed for devising the experimental protocol on the integration of GenAI tools to term record drafting, with emphases on terminological definitions as core elements provided in term records.

METHODOLOGY FOR INTEGRATING GenAI TOOLS INTO THE ARTES CORPUS-BASED FRAMEWORK

Integrating LLMs into the process of finding and drafting term records and definitions in particular is expected to save time and enhance the quality of definitions (cf. San Martín 2024). Nevertheless, in our approach, corpus-based analysis is conducted before introducing LLMs to be able to effectively integrate and assess them.

Davies (2025) explains, in his critical evaluation of “how well the predictions of Large Language Models (or LLMs, like ChatGPT and Gemini) matched up with actual corpus data”, that “a corpus is much more immersive and it’s a much more connected experience than what are often just the barebones displays in LLMs.” <...> “Full-featured corpora provide an immersive learning environment with extensive links between words.” Moreover, Davies suggest: “both LLMs and corpora have their advantages and the best is probably to use LLMs in conjunction with corpus data, since in many cases these two ‘sources of data’ complement each other quite well”. Thus, to enable effective interaction with LLMs and support a critical evaluation of their output, a thorough understanding of the linguistic data under analysis is essential.

Consequently, in our experimental approach, the interaction with LLMs relies on corpus-based terminology, combined with prompt engineering for terminology. According to Boonstra (2025: 7), “prompt engineering is the process of designing high-quality prompts that guide LLMs to produce accurate outputs. This process involves tinkering to find the best prompt, optimizing prompt length and evaluating a prompt’s writing style and structure in relation to the task. In the context of natural language processing and LLMs, a prompt is an input provided to the model to generate a response or prediction.”

For prompt engineering, we used the templates proposed by Boonstra (2025) originally developed for Gemini-pro, and by Schulhoff *et al.* (2025), which we adapted to cater for a critical approach to LLMs and to serve our objective of compiling a collection of effective prompts for term management. We thus excluded the parameters that are designed to influence model behaviour, such as “temperature” which controls the degree of randomness in token selection, because they are not supported by GenAI tools tested here.⁹ They can however be adjusted with prompt phrasing: e.g. “Be creative and unexpected.” encourages high-temperature behaviour while “Be concise and accurate.” encourages low-temperature behaviour. In term management, for finding relevant authentic data, low-temperature behaviour may be expected to yield better results. For our experimental purposes, we added a number of fields to Boonstra’s template, namely: prompting technique¹⁰, date, comment and relevance. We also adopted the possibility to define the limit diversely, by a number of characters, tokens or items.

⁹ They are supported by Application Programming Interface (API), which allows for customising the applications using LLMs.

¹⁰ The different types of prompts according to Boonstra (2025) include: zero-shot (a prompt with no examples provided), one-shot (with one example provided), few-shot (with several examples provided), system

In the last Section, we present the exploration of GenAI integration for finding definitional contexts and drafting definitions. The protocol is designed to allow experimentations in relation to specific steps of term management of the global ARTES framework:

1. Verifying if the term is recorded in existing term bases and collecting the relevant information
2. Interrogating the existing corpora for the term and retrieving the relevant information
3. Collecting relevant sources for prospecting the candidate term and specialised corpus design
4. Retrieving definitional contexts and KRC in designed specialised corpus
5. Contacting experts to resolve potential difficulties
6. Analysing, structuring and classifying the information gathered
7. Drafting an original definition
8. Creating an entry for a term in the ARTES database

We first carried out a thorough analysis of the term *data pollution* in textual sources before testing GenAI tools. To create the corpus, we conducted an extensive search for texts across several platforms¹¹, including the European Commission website, by using *data pollution* as the main key term, along with several additional terms (*digital pollution*, *digital transition* and *digital technology*) that helped us delimit the concept of *data pollution* and consider its circulation within the fields of environment and technology.

As our focus in this study is on exploring the possibilities for the integration of GenAI tools into terminological analysis, we have decided to construct a small specialised corpus of manageable size. The designed corpus comprises 22 texts (amounting to 276,415 tokens), which we sorted by year, discipline and genre before uploading them to Sketch Engine for exploration (see Figure 1 below). The size of the corpus is also limited because we targeted the term *data pollution* as used in its newly emerged meaning, which is less frequently found, as is often the case with neologisms.

prompting (consisting in providing an additional task to the system), role prompting (by stating the role and task), contextual prompting (by providing the general context), step-back prompting (consists of three steps: first a prompt serving to mobilise general knowledge, then a step-back prompt looking to find the most relevant examples, the last one consists in returning to the original prompt by including the output of the step back as context). We added to this list follow-up question, which is different from step-back prompting, as it consists in refining the output by subsequent question(s).

¹¹ The searched platforms include: Semantic Scholar (<https://www.semanticscholar.org>), Google Scholar (<https://scholar.google.com>), PLOS One (<https://journals.plos.org/plosone>), ScienceDirect (<https://www.sciencedirect.com>), European Commission (<https://commission.europa.eu> et <https://data.europa.eu/en>), Europresse (<https://www.europresse.com>), and Factiva (<https://www.dowjones.com/professional/factiva>).

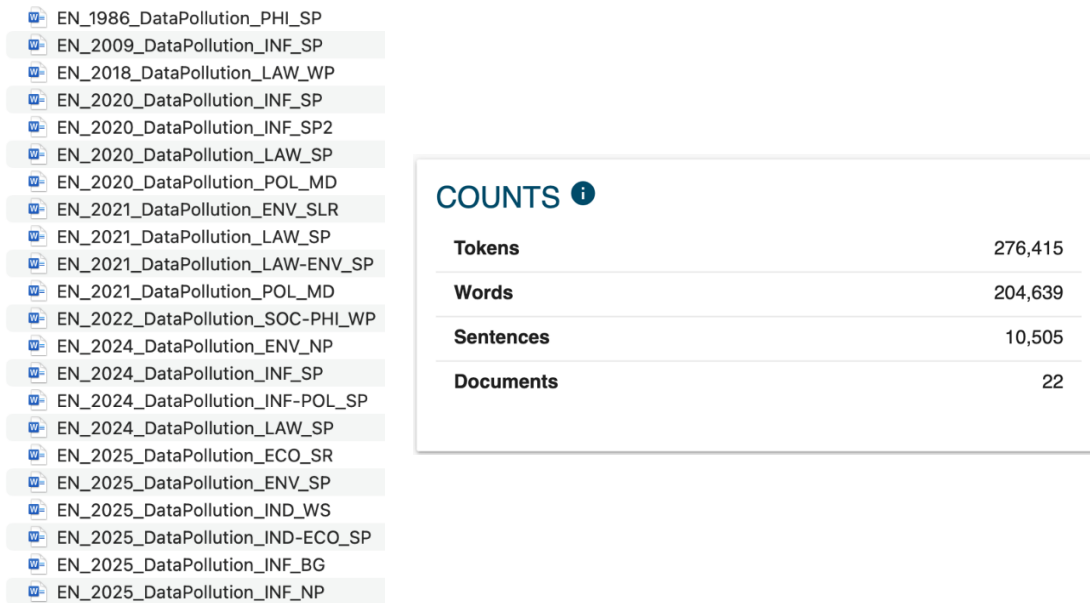


Figure 1. Collected texts sorted by year, discipline and genre (on the left) and corpus size (on the right)

The chronological span covered by the corpus is 1986 to 2025, as shown in Figure 1 and Table 1. Nevertheless, only one text dates from 1986 and another from 2009, while the majority of texts are published between 2018 and 2025, suggesting that *data pollution* is a recently emerging concept.

Table 1. Publication years of the texts selected for the corpus

Year	Number of texts according to publication year
1986	1
2009	1
2018	1
2020	4
2021	4
2022	1
2024	4
2025	6
Total no. of texts	22

The concept of *data pollution* is addressed from various perspectives and appears to be of interest across multiple domains or disciplines, as shown in Figure 1 and Table 2. Four texts are associated with two disciplines. The range of disciplines and the distribution of the concept across them indicate that *data pollution* predominantly falls within the fields of Computer Science, Cybersecurity and AI. They also suggest that it is a complex concept and a significant issue that needs to be addressed from multiple angles.

Table 2. Disciplines to which the texts selected for the corpus tend to belong

Code	Disciplines	Number of texts belonging to a discipline	Number of texts belonging to another discipline
INF	Computer Science, Cybersecurity, AI	7	
POL	Political Science	2	
SOC	Sociology	1	
PHI	Philosophy	1	1 (with SOC)
LAW	Law	5	1 (with INF)
ENV	Environment	3	1 (with LAW)
IND	Industry, Business, Services	2	
ECO	Economy, finance	1	1 (with IND)
Total no. of texts		22	

In parallel, Figure 1 and Table 3 illustrate the wide variety of document types in which the term *data pollution* appears. Half of the identified genres (SP, SLR, WP, SR) fall within the scope of the scientific register, indicating a tendency for this concept to be employed in scientific register. The remaining genres (NP, MD, WS, BG) demonstrate its circulation from the scientific to the public sphere, further emphasising the importance of this concept from scientific, economic and socio-political standpoints.

Table 3. Text genres to which the texts selected for the corpus tend to belong

Code	Text genres	Number of texts belonging to a genre
SP	scientific paper	12
SLR	systematic literature review	1
WP	white paper	2
SR	scientific report	1
NP	news paper	2
MD	ministerial declaration	2
WS	web site	1
BG	blog	1
Total no. of texts		22

In this study, we thus focus on the steps 4 and 7 of the general ARTES framework, by testing the LLMs and GenAI tools the most frequently used by our students revealed by the survey presented in the next section.

SETTING UP THE EXPERIMENTAL PROTOCOL FOR AI-ASSISTED TERMINOLOGICAL DEFINITIONS DRAFTING

Survey of Master's students' current use of GenAI tools

At the end of 2024–2025 academic year, we conducted a survey among 50 Master's students: 25 Master 1 and 25 Master 2 students who attended classes on terminology in both semesters. The students are introduced into corpus-based terminology and term management within the first year of Master. In the second year, they undertake a project, as part of their Master's dissertation, involving extensive terminological analysis of a specific specialised discourse.

Up to 2024–2025 academic year, the terminology courses, on both levels, did not systematically include the LLMs for the purposes of term management. However, the same students attended and completed a series of courses related to terminology, including information retrieval, corpus design and specialised translation, in which NMT systems are introduced, and progressively LLMs too. It was thus interesting to conduct a survey to evaluate whether the students used LLMs for terminology, and which ones, in order to better prepare the protocol for the current 2025–2026 academic year.

The survey reveals an emerging tendency among Master's students to use GenAI tools when drafting term records: 26 students out of 50 use them; however, it also indicates that Master 2 students use these tools more frequently (Figure 2).

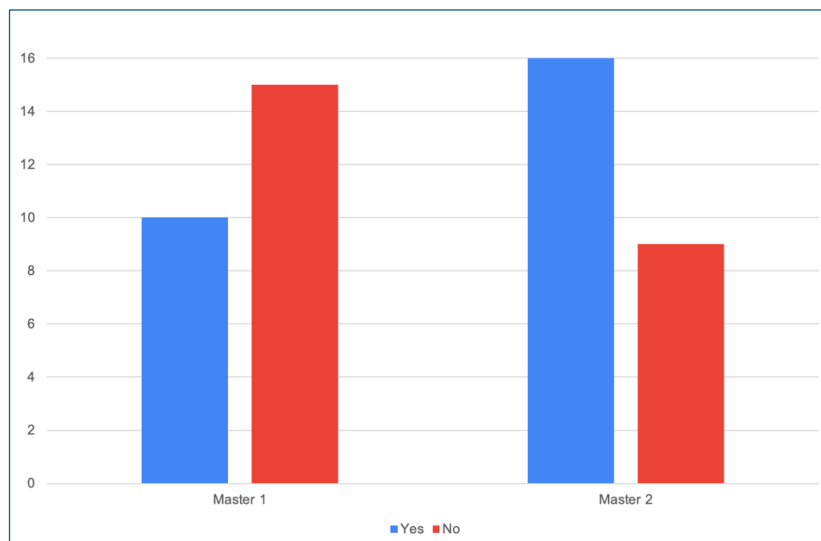


Figure 2. The tendency of Master 1 and Master 2 students to use GenAI tools in drafting term records

Although the total number of respondents (50) is too low to draw firm conclusions, we can make some useful observations to inform protocol design. For instance, among the 25 Master 1 students, only 10 reported using GenAI tools, compared to 16 out of 25 Master 2 students. This corresponds to 40% and 64% usage respectively, suggesting a possible trend of increased adoption of GenAI tools at more advanced stages of study, though further research with a larger sample would be needed to confirm this tendency.

The difference is not unexpected, as Master 2 students are engaged in a terminological project as part of their Master's theses. It is also unsurprising that students explored the use of these tools independently even though they were not introduced within the ARTES framework during the 2024–2025 academic year.

Moreover, the majority of students use these tools, although not all have been offered courses on their integration into linguistic and language analysis (Figure 3).

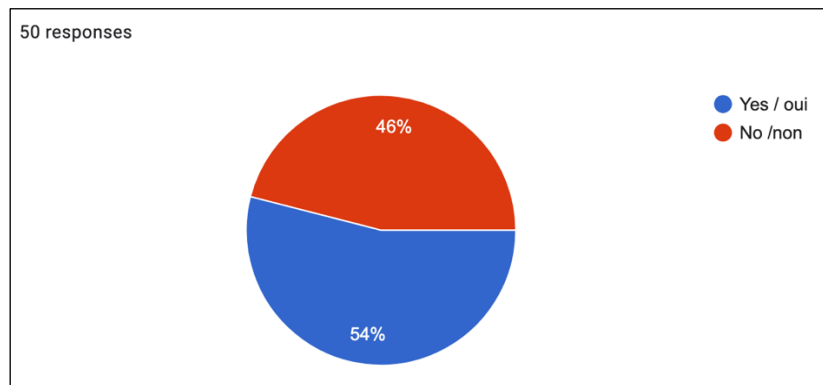


Figure 3. The place of GenAI tools in Master's courses showing how many students attended classes in which GenAI tools were introduced

Similarly, the majority of students searched on their own for relevant information to learn how to use GenAI tools efficiently (Figure 4).

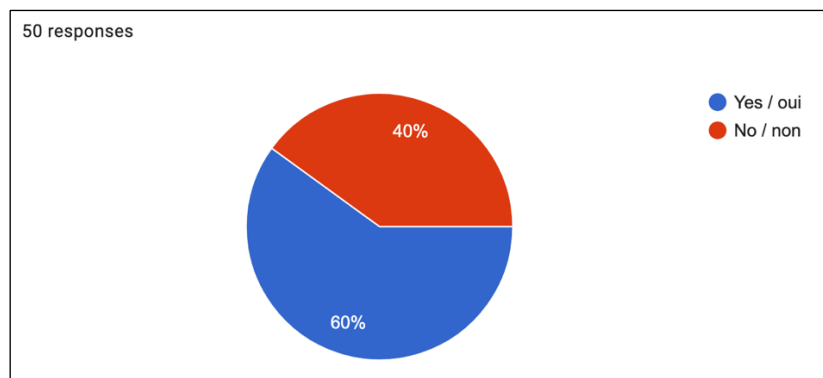


Figure 4. The tendency of Master's students to search on their own for relevant information to effectively use GenAI tools

Consequently, this survey highlights the pressing need to provide Master's students with a protocol for the efficient integration of GenAI tools into the process of drafting term records.

Furthermore, the survey indicates that students' primary need is associated with drafting definitions, with 63% reporting the use of GenAI tools for this purpose. Additionally, 55.6% of students use GenAI tools to help them identify the sources for corpus design and semantically related terms (Figure 5).

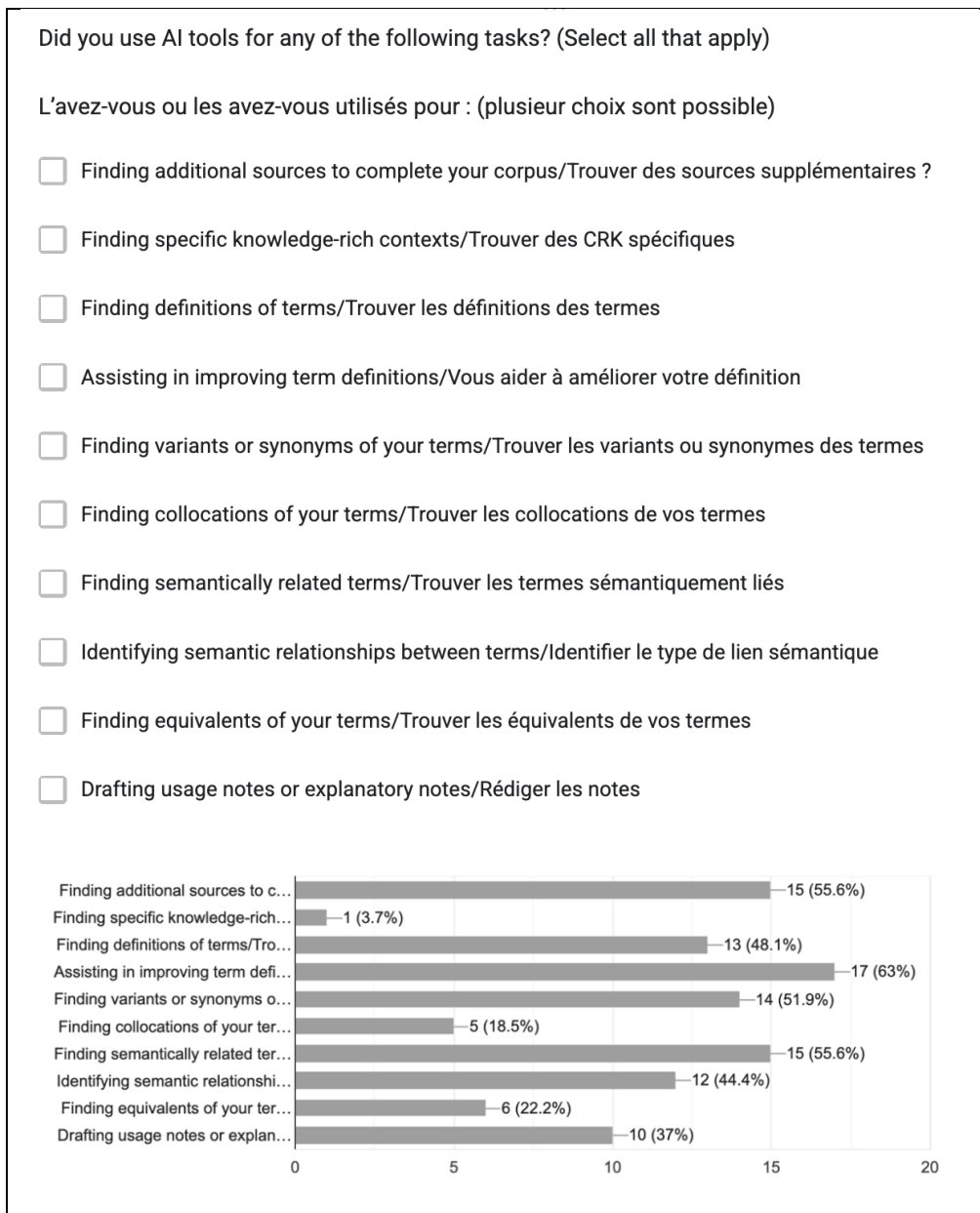


Figure 5. The tendency of Master's students to use GenAI tools for drafting definitions, as well as for identifying sources for corpus and semantically related terms

To further account for the tools students tend to use, we also inquired about their preferred GenAI tools. The survey reveals an overwhelming preference for ChatGPT (85,2%), followed by Perplexity (11,1%) and DeepSeek to a lesser extent (3,7%) (Figure 6).

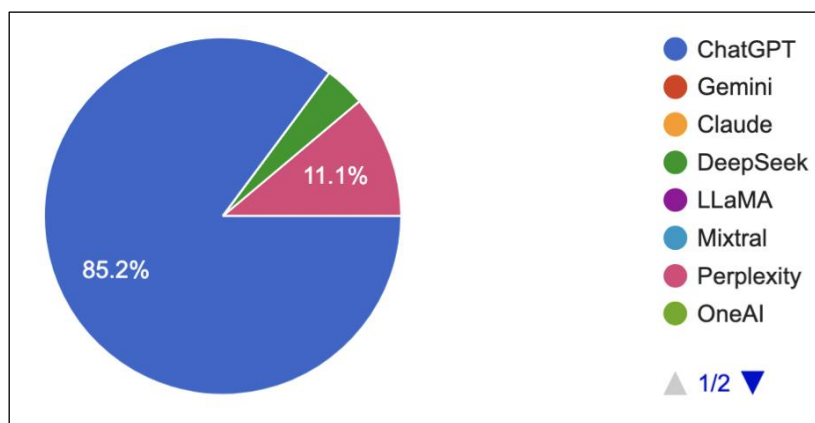


Figure 6. The most frequent GenAI tools used by Master 1 and Master 2 students for drafting term records

In this study, we thus focus on testing ChatGPT by OpenAI, as well as DeepSeek developed by the Chinese company of the same name, and Perplexity developed by Perplexity AI which uses different LLMs, namely GPT developed by OpenAI and Claude by Anthropic. In the next sections, we present a systematic sequential protocol for their integration into the ARTES term management workflow.

Using corpus-based terminology to test the efficiency of GenAI tools

As emphasised in the previous section, in the experimental protocol we propose, the terminological analysis by corpus-based approach is conducted prior to interacting with LLMs to ensure the ability to critically assess GenAI output. We thus conducted the analysis through all the stages of the ARTES framework, that is the stages 1 to 7, listed in the previous section, namely using SketchEngine¹² (Kilgarriff *et al.* 2014) for specialised corpus design and exploration, alongside corpora integrated in the SketchEngine and two more tools with integrated corpora, Google Books Ngram Viewer¹³ (Michel *et al.* 2011) and Netspeak¹⁴ (Riehmman *et al.* 2012).

Figure 7 illustrates the results of the phase 1, consisting in verifying if the term is recorded in existing term bases. The term *data pollution* is rarely recorded in well-known term bases, as only one entry was found in IATE, created in 2021, for the original meaning of the term.

SEARCHED TERM: <i>data pollution</i>			
SEARCHED DATE: 25/05/2025			
TERM BASE	RECORD	COMMENT	RELEVANCE
Termium Plus	none	NA	NA
Vitrine linguistique	none	NA	NA

¹² Available at: <https://www.sketchengine.eu>.

¹³ Available at: <https://books.google.com/ngrams>.

¹⁴ Available at: <https://netspeak.org>.

IATE	1 record	Record creation date: 2.3.2021, Meaning: the 1 st meaning of the term. Provided fields and information: definition, contexts, sources (from 2018 and 2019), equivalents in 20 languages (amongst which French: <i>pollution des données</i>)	medium
WIPO	none	NA	NA
UNTERM	none	NA	NA
IGI Global Dictionary	none	NA	NA

Figure 7. Results of the search for the term *data pollution* in existing term bases

As the content of the record is devoted to the original meaning of the term, this finding is considered of medium relevance for the targeted record design. Figure 8 shows the record content in IATE.

The screenshot shows the IATE record for the term "data pollution". At the top, it indicates the entry ID "3593027" and the source "Consilium". A warning note states "Not to be confused with: 3592145 data poisoning". The term "data pollution" is highlighted in blue. The record includes several fields:

- Term reference:** European Commission, Independent High-Level Expert Group on Artificial Intelligence, *Ethics guidelines for trustworthy AI*, April 2019
- Definition:** injection of maliciously crafted training data samples into a training set, causing the AI system to learn an incorrect model and subsequently misclassify testing samples
- Definition reference:** Council-EN, based on: Yinzhi Cao et al. *Efficient Repair of Polluted Machine Learning Systems via Causal Unlearning*. ACM Asia Conference on Computer and Communications Security, Incheon, Republic of Korea, 4–8 June 2018 (2.3.2021)
- Language level note:** In the case of **data pollution** the data is tampered with (most likely AFTER the development of the system) for either commercial or malicious purposes, whereas **data poisoning** refers to a deliberate attack carried out DURING the development of an AI system (making the system inherently corrupt. This strict distinction between the two concepts is not always made, however, and the terms are sometimes used interchangeably.
- Language level note reference:** Council-EN based on information from Artificial Intelligence FCT
- Term in context:** A major attack, called **data pollution**, injects maliciously crafted training data samples into the training set, causing the system to learn an incorrect model and subsequently misclassify testing samples. The most recent real-world example is Microsoft's AI powered chatbot Tay. Tay learned racism because some Twitter users interacted with Tay using offensive, racist words, and these words were included in Tay's training set [...]. A natural solution to a **data pollution** attack is to remove the polluted data from the training set and relearn a clean model. Unfortunately, the training set of a real-world machine learning system can contain millions of samples; it is thus hopeless for an administrator to manually inspect all of them to weed out the polluted ones. This overwhelming amount of manual cleaning required is perhaps why Microsoft brought Tay offline for repair but has yet to bring it back online.'
- Context reference:** Yinzhi Cao et al. *Efficient Repair of Polluted Machine Learning Systems via Causal Unlearning*. 2018 ACM Asia Conference on Computer and Communications Security, Incheon, Republic of Korea, June 4–8, 2018 <http://yinzhicao.org/unlearning/casual-unlearn-asiaccs18.pdf> (2.3.2021)

The record also shows the creation date (2.3.2021 10:37) and modification date (2.3.2021 10:44).

Figure 8. Record of the term *data pollution* in IATE displaying information on the original meaning of the term

Nevertheless, the data provided can be used to create a term record for *data pollution* with its original meaning (which could be glossed as ‘the act of polluting the data’) in order to contrast it with the record dedicated to *data pollution* used with the new meaning (which could be glossed as ‘the pollution caused by the data’). The complexity of defining *data pollution* with its newly emerged meaning (sometimes reformulated by *pollution by data*), is particularly complex because not only it overlaps with the meaning of the term *digital pollution* but also because of the simultaneous existence of its original meaning (sometimes reformulated by *pollution of data*, *polluted data*). Gathering information from existing resources and replicating them in term records drafted in ARTES lends coherence to analysis, by taking into account the wide scope of information on a term (Figure 9).

DEFINITION	injection of maliciously crafted training data samples into a training set, causing the AI system to learn an incorrect model and subsequently misclassify testing samples
SOURCE	IATE, Council-EN, based on Yinzhi Cao <i>et al.</i> 2018: 1
ISONYME	data poisoning
COMMENT	the first meaning of the term
RELEVANCE	medium

Figure 9. Information selected from IATE for the ARTES DB

The absence of the term in existing term bases illustrates the usefulness of corpus-based approach. Figure 10 shows the definitions retrieved from corpora by using discourse markers for definitional contexts (such as “is a”, “is the”, “refer to”, etc.):

1	DEF. KRC	Data pollution is the set of harms generated by economic activities related to data collection, storage and use, which transversally impacts individuals and their living environments.
	SOURCE	Mori et al. 2024: 155
	COMMENT	the new meaning of the term, found in more recent literature
	RELEVANCE	high
2	DEF. KRC	Data pollution is the interrelated adverse impact that the generation, storing, handling, and processing of digital data has on our natural environment, social environment, and personal environment. It is the unsustainable handling, distribution, and generation of data resources.
	SOURCE	Hasselbalch et al. 2022: 9–10
	COMMENT	the new meaning of the term, found in recent literature
	RELEVANCE	high
3	DEF. KRC	Internet contents (documents, emails, chats, images, videos, etc.) that are posted on the Internet are often disseminated and replicated on different peers or servers, generating what we refer to as Data Pollution.
	SOURCE	Castelluccia & Kaafar 2009: 1
	COMMENT	the first meaning of the term, found in less recent documents
	RELEVANCE	medium
4	DEF. KRC	In what follows the term “data pollution” is taken to refer to the accumulation of all “contaminations” or “distortions” which can result from working with data in the information technology field.

	SOURCE	Zimmerli 1986: 291
	COMMENT	the first meaning of the term, found in in less recent documents
	RELEVANCE	medium

Figure 10. Definitional contexts retrieved from corpus

Although several definitional contexts were found, only the first two refer to the recent meaning occurred by semantic shift. These definitional contexts provided in expert-to-expert communication are of high relevance for targeted term record design. The first one is particularly of relevance as it comes from a paper published in proceedings of an international conference.

The second definition relevant for drafting the targeted term record comes from a white paper (Hasselbalch 2022), published subsequently to a monograph (Hasselbalch 2021), where the issue of *data pollution* is addressed within the framework of an independent initiative¹⁵.

The definitional context 3 and 4 are considered of medium relevance as they point to the original meaning of the term.

In the next step, the acquisition of knowledge on the concept of data pollution through corpus consisted in analysing concordances of the term, and more specifically in identifying KRCs. Corpus analysis allowed us to retrieve multiple contexts from which we selected the most relevant ones for understanding the term, and its semantic relation to other terms characteristic of this discourse. These contexts are also useful for providing the examples on use of the term. Figure 11 shows a sample of collected KRCs (with the selected parts in bold for interacting with LLMs, presented in the last Section):

1	KRC	Recognizing that data pollution is also a public problem degrading an entire ecosystem and not merely the individual spheres of the data givers , offers a new and rich perspective on the existing solutions—and introduces new ones.
	SOURCE	Ben-Shahar 2018: 133
	COMMENT	the new meaning of the term, found in recent literature
	RELEVANCE	high
2	KRC	Digital information is the fuel of the new economy. But like the old economy’s carbon fuel, it also pollutes. Harmful “data emissions” are leaked into the digital ecosystem, disrupting social institutions and public interests. This article develops a novel framework—data pollution—to rethink the harms the data economy creates and the way they have to be regulated.
	SOURCE	Ben-Shahar 2018: 104

¹⁵ The Data Pollution & Power – Initiative. Available at: <https://www.datapollution.eu>.

	COMMENT	the new meaning of the term, found in recent literature
	RELEVANCE	high
3	KRC	Two traditional usages of the term data pollution can thus be combined. Firstly, data pollution can be understood as the adverse impact on personal and social environments, for instance on individual rights, such as data protection or the right to private life, and on democratic institutions and balances of power. Secondly, data pollution can be understood as the material adverse effects on our natural environment, e.g., the carbon footprint of big data.
	SOURCE	Hasselbalch 2022: 22
	COMMENT	the new meaning and the original meaning of the term, put in contrast
	RELEVANCE	high
4	KRC	In the white paper, data pollution is addressed similarly as not only one type of environmental impact, but rather as the interrelated adverse effects on delicate balances in our natural, social and personal ecosystems and environments. As described, the term data pollution is currently used to emphasise the very real and material adverse environmental impact of big data on these environments. As follows, the goal of a new ‘green movement’ for big data is ‘data sustainability’, which cuts across the SDGs with sustainability considerations connected to the various environmental changes caused by the volume and diversity of big data, ranging from its effects on the natural landscape to our decisions and democracy.
	SOURCE	Hasselbalch 2022: 22–25
	COMMENT	the example shows the entanglement of the new meaning and the original meaning of the term
	RELEVANCE	high

Figure 11. Retrieval of Knowledge Rich Contexts (KRCs) in a specialised corpus

The KRCs 1 to 4 illustrate the term used with a new meaning.¹⁶ In the following step, on the bases of the collected definitional contexts and KRCs (Figures 10–11), we can make a proposal for a terminological definition, presented in Figure 12:

DRAFTED DEFINITION	specific type of pollution related to information technologies and digital era characterised by the increasing generation of data which is transforming professional, social and everyday life of individuals and their environment in the way that appears not to be in accordance with their capabilities or needs for retrieving relevant information (295 characters)
SOURCE	term record author name, based on Mori et al. 2024, Hasselbalch et al. 2022 and Ben-

¹⁶ We also collected the relevant KRCs illustrating the first meaning of the term for drafting the record of the concurrent term in order to link the two records and provide a note on the semantic link between the two terms.

	Shahar 2018
RELEVANCE	high

Figure 12. Drafted terminological definition based on corpus retrieval and KRCs

At this stage of our experimental analysis, we observed a cognitive bias: the restriction to interact with LLMs before finalizing corpus-based analysis became an obstacle as it turned out to require substantial self-control. In the next step, we thus introduced LLMs for finding KRCs and relevant sources, KRCs-informed definition drafting, revising of (human) term definitions, and for drafting definitions through a “loop” interaction with LLMs.

Introducing GenAI tools to the process of term analysis

Finding Knowledge-Rich Contexts (KRC) and relevant sources

We first tested LLMs capacity to find definitional contexts. The target was to write a prompt which would return the most relevant definitional contexts in quotes followed by a reference. After several attempts, the prompting technique illustrated in Figure 13 yielded useful results. However, it did not return specifically definitional contexts, which in the case under scope is not surprising. Corpus-based analysis showed that the definitional contexts for *data pollution* are rare. The prompt returned useful KRCs and relevant references for corpus design.

By using a combination of system prompting and step-back prompting, we obtain not only the results relevant for the new meaning of the term, but also in the desired format. By restricting the search to specific type of literature (scientific) and limiting number of definitions to look for (three), we constrained LLM to produce the most relevant output (see Prompt 2 in Figure 13). Moreover, we first asked one question to allow the LLM to activate the general knowledge (Prompt 1 in Figure 13), before asking it to find the most relevant examples. Figure 13 presents the test conducted on ChatGPT:

GOAL	Finding KRCs and relevant references
MODEL	ChatGPT-4-turbo
DATE	13/05/2025
TECHNIQUE	Step-back prompting, system prompting
LIMIT	3 definitions
PROMPT 1	What is the difference between “data pollution” defined in these terms “specific type of pollution related to information technologies and digital era characterised by the increasing generation of data which is transforming professional, social and everyday life of individuals

	and their environment in the way that appears not to be in accordance with their capabilities or needs for retrieving relevant information” and “digital pollution”?
PROMPT 2	Find three most relevant definitions of the term “data pollution” in scientific literature and provide the sources for each definition found. Return the definitions in the list with the exact wording found in the texts followed by the bibliographical reference of the source and its URL.
OUTPUT 1	<p>The terms “data pollution” and “digital pollution” are related but distinct in focus and scope:</p> <p style="padding-left: 40px;">Data Pollution</p> <p>Focus: The harmful effects of excessive or mismanaged data—its collection, storage, processing, and use.</p> <p>Key Concerns:</p> <ul style="list-style-type: none"> Overload of irrelevant, biased, or harmful data Environmental and social costs of big data infrastructures Impacts on human attention, privacy, and decision-making Disruption of social and informational ecosystems <p>Analogy: Comparable to chemical pollution—contamination of the information environment with harmful byproducts.</p> <p style="padding-left: 40px;">Digital Pollution</p> <p>Focus: The broader environmental and resource impact of digital technologies.</p> <p>Key Concerns:</p> <ul style="list-style-type: none"> Energy consumption and carbon footprint of devices, servers, and networks E-waste from hardware and devices Lifecycle impact of producing and disposing of digital tools <p>Analogy: Comparable to industrial pollution—physical and ecological damage caused by digital infrastructure.</p> <p style="padding-left: 40px;">Summary of the Difference</p> <p>Data pollution is about the <i>content and consequences of data itself</i> and its socio-environmental effects.</p> <p>Digital pollution is about the <i>environmental cost of the digital systems</i> that store, transmit, and process data.</p> <p>Would you like a side-by-side table comparison for clarity?</p>
OUTPUT 2	<p>“Digital information is the fuel of the new economy. But like the old economy’s carbon fuel, it also pollutes. Harmful ‘data emissions’ are leaked into the digital ecosystem, disrupting social institutions and public interests.”</p> <p>Ben-Shahar O. 2018: <i>Data Pollution</i> 679, University of Chicago Public Law & Legal Theory Paper Series. Available at: https://chicagounbound.uchicago.edu/public_law_and_legal_theory/699/</p> <p>“Data pollution is the interrelated adverse impact that the generation, storing, handling and processing of digital data has on our natural environment, social environment and personal environment. It is the unsustainable handling, distribution and generation of data resources.”</p> <p>Hasselbalch G. <i>et al.</i> 2022: <i>Data Pollution & Power – White Paper for a Global Sustainable Development Agenda on AI</i>. Available at: https://www.researchgate.net/publication/362174889_Data_Pollution_Power_-_White_Paper_for_a_Global_Sustainable_Development_Agenda_on_AI_by_Gry_Hasselbalch_with_contributions_from_the_Data_Pollution_Power_DPP_Group_at_Bonn_Sustainable_AI_Lab</p>

	<p>“The concept of data pollution includes information pollution or ‘infollution’ (Özdemir 2016), as coined by Oram 1984. However, information pollution is different to some extent; ‘the contamination of information sources with irrelevant, redundant, unsolicited and low-value information’ or as Cai and Zhang (1996) quote: ‘... is a pile of widespread yet unwanted messages and that one day, these messages could deeply influence the social life bearing negative results.”</p> <p><i>Data Pollution and Taxation</i> 2022. Available at: https://www.academia.edu/76896486/Data_Pollution_and_Taxation</p>
COM- MENT	<p>1st not a definitional context, a KRC, a quote, source provided and well-constructed</p> <p>2nd a definitional context, a quote, source provided and well-constructed</p> <p>3rd not a definitional context, a KRC, a quote, misleading (consists of two definitions of another term), source provided but poorly constructed, not a scientific register (academic register)</p>
RELE- VANCE	high

Figure 13. Prompt engineering for term management: finding KRC and relevant references with ChatGPT

Although only one of four definitional contexts found by corpus-based approach was returned, alongside two KRCs, the output is relevant for the analysis of the concept of *data pollution*. First, the output shows that corpus-based approach cannot be bypassed for selecting the best definitional contexts and KRCs, as well as for effectively evaluating the LLM’s output. Second, ChatGPT returned one source that was not identified during the phase of corpus design. Although it belongs to academic rather than scientific register¹⁷, it can prove useful in the subsequent analysis on semantically related terms, and be used for augmenting the corpus: the discursive marker “includes” infers that *information pollution* is a meronym of *data pollution*, and the markers “or” and simple quotes that *infollution* is a synonym of *information pollution*. Nevertheless, this third proposal made by ChatGPT as a definitional context is misleading because it contains two definitions of another concept (*information pollution*) and no definition of the concept searched for.

Moreover, in a previous prompt, we used the same instruction without specifying that the expected output are the quotes, which resulted in reformulated quotes and misleading data (see Figure 24 in Appendix).

KRC-informed definition drafting

The next test consisted in asking LLMs to draft a definition supported by a selection of KRCs. We selected 3 KRCs: the definitional context 1 (in Figure 10) and the extracts from KRC 1 and 4 (appearing in bold in Figure 11). Figure 14 shows this test with ChatGPT, DeepSeek and Perplexity.

¹⁷ In the ARTES framework, Master’s dissertations are consulted when specialist sources are lacking. In general, they are not used for constructing specialised corpora. They are considered as belonging to academic rather than scientific register.

115b bG115a	KRC-informed definition drafting		
DATE	13/05/2025		
TECH- NIQUE	Zero shot		
LIMIT	300 characters		
PROMPT	<p>Draft the terminological definition of the term “data pollution” in no more than 300 characters on the bases of these knowledge rich contexts:</p> <p>Data pollution is the set of harms generated by economic activities related to data collection, storage and use, which transversally impacts individuals and their living environments.</p> <p>Data pollution is also a public problem degrading an entire ecosystem and not merely the individual spheres of the data givers.</p> <p>Data pollution is addressed as not only one type of environmental impact, but rather the interrelated adverse effects on delicate balances in our natural, social and personal ecosystems and environments. The term data pollution is currently used to emphasise the very real and material adverse environmental impact of big data.</p>		
MODEL	ChatGPT-4-turbo	DeepSeek-V3	Perplexity
OUTPUT	<p>Data pollution is the harmful byproduct of data-related economic activities—such as collection, storage, and use—that disrupts natural, social, and personal ecosystems, affecting both individuals and the broader environment in ways comparable to traditional environmental pollution.</p>	<p>“Data pollution refers to the systemic harms caused by data-related economic activities, degrading natural, social, and personal ecosystems through adverse environmental, collective, and individual impacts beyond mere privacy concerns.”</p> <p>(298 characters)</p> <p>Key features captured:</p> <ol style="list-style-type: none"> 1. Systemic harm – Emphasizes large-scale degradation, not just individual effects. 2. Economic root – Links to data collection, storage, and use. 3. Transversal impact – Affects natural, social, and personal environments. 4. Beyond privacy – Posi- 	<p>Data pollution refers to the harms caused by economic activities involving data collection, storage, and use, which degrade natural, social, and personal ecosystems. It is a public problem with interconnected adverse effects extending beyond individuals to entire environments.</p>

		tions it as an ecological/public issue. This aligns with the framing of data pollution as an environmental and societal challenge. Let me know if you'd like any adjustments!	
COMMENT	adequately formulated, relatively clear, fairly concise	adequately formulated, relatively clear, fairly concise, with additional information	inadequately formulated (two utterances), complex (by the choice of hypernym which is less clear), with high alignment to KRC examples, fairly concise
RELEVANCE	high	high	low

Figure 14. Prompt engineering for term management: KRC-informed definition drafting with ChatGPT, DeepSeek and Perplexity

The three outputs appear relevant for term management, and this method can be regarded as suitable for drafting terminological definitions with LLMs' assistance. Indeed, providing KRCs guides the LLMs to produce a definition of the newly emerging meaning for the term. The outputs by ChatGPT and DeepSeek appear as more appropriate, in comparison to Perplexity. In appendix, in Figure 25, we provide an example of the risks of using insufficiently constrained prompts when managing terms undergoing semantic shift with LLMs.

Revising and improving bio-definition by LLMs

In the next test, we asked GenAI tools to improve the definition we drafted without the help of AI-tools, on the bases of collected information, namely various KRC, and which we propose to call “bio-definition” (Figure 15):

GOAL	Improving bio-definition
DATE	13/05/2025
TECHNIQUE	Zero shot
LIMIT	300 characters
PROMPT	Improve this definition of the term “data pollution”: specific type of pollution related to information technologies and digital era characterised by the increasing generation of data which is transforming professional, social and everyday life of individuals and their environment in the way that appears not to be in accordance with their capabilities or needs for retrieving relevant information

MODEL	ChatGPT-4-turbo	DeepSeek-V3	Perplexity
OUTPUT	<p>Data pollution is a form of digital-era pollution arising from the excessive and often uncontrolled generation of data, which overwhelms individuals' and systems' ability to retrieve, process, or use relevant information effectively, disrupting professional, social, and daily life.</p> <p>Would you like this tailored further for a specific audience (e.g., academic, policy, general public)?</p>	<p>“Data pollution refers to the excessive, irrelevant, or low-quality data generated in the digital age, overwhelming individuals and systems beyond their capacity to process or extract meaningful information, leading to inefficiencies and misinformation.”</p> <p><i>(298 characters)</i></p> <p>Key improvements:</p> <ol style="list-style-type: none"> 1. More focused – Clearly defines the problem (excessive, irrelevant, or low-quality data). 2. Stronger cause-effect link – Explains how it overwhelms systems and individuals. 3. Practical impact – Highlights consequences (inefficiencies, misinformation). 4. More concise – Fits within the 300-character limit while maintaining clarity. <p>Would you like any further refinements?</p>	<p>Data pollution is a specific form of pollution in the digital era, characterized by the excessive and uncontrolled generation of data that overwhelms individuals and systems, hindering their ability to access relevant, accurate, and meaningful information in professional, social, and everyday contexts.</p> <p>Related questions:</p> <p>How does data pollution affect decision-making processes</p> <p>What are the long-term consequences of data pollution on society</p> <p>How can individuals protect themselves from data pollution</p> <p>What industries are most affected by data pollution</p> <p>How does data pollution influence data privacy laws</p>
COMMENT	adequately formulated, clear, fairly concise	adequately formulated, complex, fairly concise, with additional information	adequately formulated, complex, fairly concise, with additional information
RELEVANCE	high	medium	medium

Figure 15. Prompt engineering for term management: improving bio-definition with ChatGPT, DeepSeek and Perplexity

As in the previous test, all three outputs appear relevant and the method suitable for generating terminological definitions with LLMs assistance. The output by ChatGPT appears again more appropriate, less complex, in comparison to DeepSeek and Perplexity.

Drafting definitions through LLM-bio loop interaction

This final test consisted in drafting a bio-definition inspired by the interaction with Gen-AI tools, by selecting lexical choices carrying the key semantic components of the concept. The resulting definition was submitted to an LLM for final refinement. This circular or loop interaction with LLM (which we propose to name LLM-bio loop) yields an interesting result but of limited applicability for term record drafting if we take into account the complexity of a produced definition (Figure 16). “LLM-bio loop” thus designates iterative back-and-forth interactions between AI output and human revision until reaching a desired result.

GOAL	Revising a definition
MODEL	Perplexity
DATE	13/05/2025
TECHNIQUE	Zero shot, follow-up question
LIMIT	NA
PROMPT 1	Revise: a form of pollution arising from excessive, irrelevant, or low-quality data generated in the digital age, impeding individuals to process and extract meaningful information, hence degrading natural, social, and personal ecosystems and causing a public problem with interconnected adverse effects extending beyond individuals to entire environments
OUTPUT	a form of pollution caused by excessive, irrelevant, or low-quality data generated in the digital age, which hinders individuals from processing and extracting meaningful information, thereby degrading natural, social, and personal ecosystems, and creating a public problem with interconnected effects extending beyond individuals to entire environments
COMMENT	good quality comprehensive definition, clear, relatively concise (308 characters)
RELEVANCE	high
PROMPT 2	What have you changed in a revised version of my sentence for clarity and flow?
	<p>Here’s what I changed in your sentence for clarity and flow:</p> <ol style="list-style-type: none"> Subject clarity: Changed “arising from” to “caused by” for a more direct and active construction. Verb usage: Replaced “impeding individuals to process” with “hinders individuals from processing” for grammatical correctness and smoother readability. Parallel structure: Broke up the sentence to clarify the sequence—first stating the cause (excessive, irrelevant, or low-quality data), then the effect on individuals, and finally the broader impact on ecosystems and society. Conciseness: Streamlined phrases such as “hence degrading” to “thereby degrading” to improve logical flow. Consistency: Ensured that the sentence maintains a consistent subject and verb

	agreement throughout.
6.	Readability: Reordered some elements to make the sentence easier to follow and less cumbersome.

Figure 16. Revision of a definition drafted by LLM-bio loop interaction, by Perplexity with a follow-up question on the revision process

This last test also demonstrates LLMs’ possible pedagogical value for crafting precise and well-formulated definitions, with an iterative follow-up question which provides detailed justifications for each edit on the process of improving clarity.

Evaluation of the quality of definitions obtained by various approaches

We finally submitted the selection of definitions to appraisal by Master’s students. The different modes used for obtaining them were not informed in the query constructed to collect their feedback and appraisal of the quality of definitions. We asked them to rate each definition on a scale from 1 to 5 (with 5 being the highest), but without requiring them to rank the definitions from lowest to highest.

We selected the ‘expert definition’¹⁸ retrieved in specialised corpus (1), an LLM’s drafted definition informed by KRC (2), revised bio-definition by an LLM (3), the definition produced by LLM-bio loop mode consisting in selecting the best material (4), and bio-definition drafted after completing corpus-based term analysis (5), as illustrated in Figure 17.

1	EXPERT DEFINITION	the set of harms generated by economic activities related to data collection, storage and use, which transversally impacts individuals and their living environments
2	KRC-INFORMED LLM’s DEFINITION	the harmful byproduct of data-related economic activities—such as collection, storage, and use—that disrupts natural, social, and personal ecosystems, affecting both individuals and the broader environment in ways comparable to traditional environmental pollution
3	LLM-IMPROVED BIO-	a form of digital-era pollution arising from the excessive and

¹⁸ In the ARTES framework, the ‘expert definitions’ are expert definitional contexts retrieved from specialised corpora. They are initially recorded in the database without any modifications, in order to provide evidence for the definitions proposed by experts. The criteria for selecting and categorising them as definitions rely on the structure and content of the context, which must conform to terminological definition standards. Another criterion that can additionally be used is the evaluation of the expert’s intention to provide a definition, as illustrated in our example with the definition by Mori *et al.* (2024). In the next step, these ‘expert definitions’ or expert definitional contexts are used to produce terminological definitions by applying the necessary modifications and improvements to meet the requirements of ISO standards for drafting definitions and to achieve the desired clarity of definition according to the targeted audience. If an expert definitional context does not conform to terminological definition standards, but is nonetheless considered as potentially useful for understanding a concept, it is recorded in the database as a contextual example rather than as a definition.

	DEFINTION	often uncontrolled generation of data, which overwhelms individuals' and systems' ability to retrieve, process, or use relevant information effectively, disrupting professional, social, and daily life
4	LLM-BIO LOOP DEFINTION	a form of pollution caused by excessive, irrelevant, or low-quality data generated in the digital age, which hinders individuals from processing and extracting meaningful information, thereby degrading natural, social, and personal ecosystems, and creating a public problem with interconnected effects extending beyond individuals to entire environments
5	BIO-DEFINTION	specific type of pollution related to information technologies and digital era characterised by the increasing generation of data which is transforming professional, social and everyday life of individuals and their environment in the way that appears not to be in accordance with their capabilities or needs for retrieving relevant information

Figure 17. Selection of definitions submitted to the appraisal by Master's students

Figure 18 shows the results of Master 1 and 2 students' appraisal of the results of different approaches to crafting definitions, during the survey conducted at the end of 2024-2025 academic year.

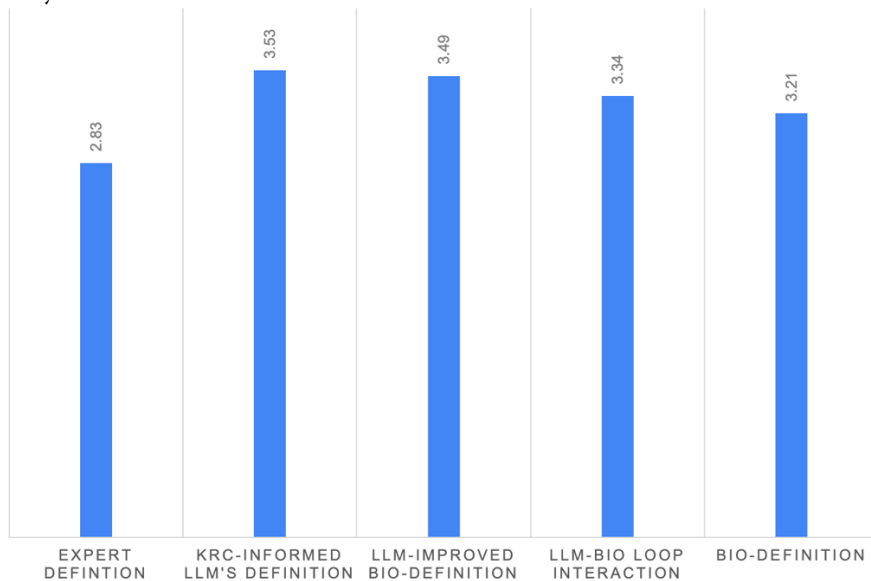


Figure 18. Appraisal of the results of different approaches to crafting definitions by Master's students

The differences are relatively subtle, with however a slight preference for the three definitions in the middle, crafted in the interaction with GenAI tools. The appraisal by students thus highlights the role of human intervention and corpus-based approaches in crafting high-quality definitions, while acknowledging LLMs' potential as supplementary tools for improving the definitions.

However, scores also reveal limitations in relying solely on LLMs for drafting definitions, even with KRC-informed method which scored the best result, as KRC-

informed LLM’s definitions consisted in selecting the KRC from corpora by human-only intervention, before devising a prompt.

Additionally, results highlighted the need to provide explicit terminological definitions alongside expert ones, which may not adhere to standard definition guidelines, as the expert definition scored the lowest result.

However, a comparison of the two groups (Figure 19) reveals a different perspective: Master 1 students tend to prefer definitions in which the role of LLMs is more prominent (specifically, the KRC-informed LLMs’ definition and the LLM–bio loop definition), whereas Master 2 students prefer the ones where human input is stronger (such as the bio-definition and the LLM-improved bio-definition).

Additionally, Master 2 students are more critical of the expert definition than Master 1 students, noting that it does not provide a clear hypernym. The difference in academic level between the two groups appears to be a significant factor. This survey thus provides evidence that Master 2 students acquire the necessary skills for a critical approach to definitions.

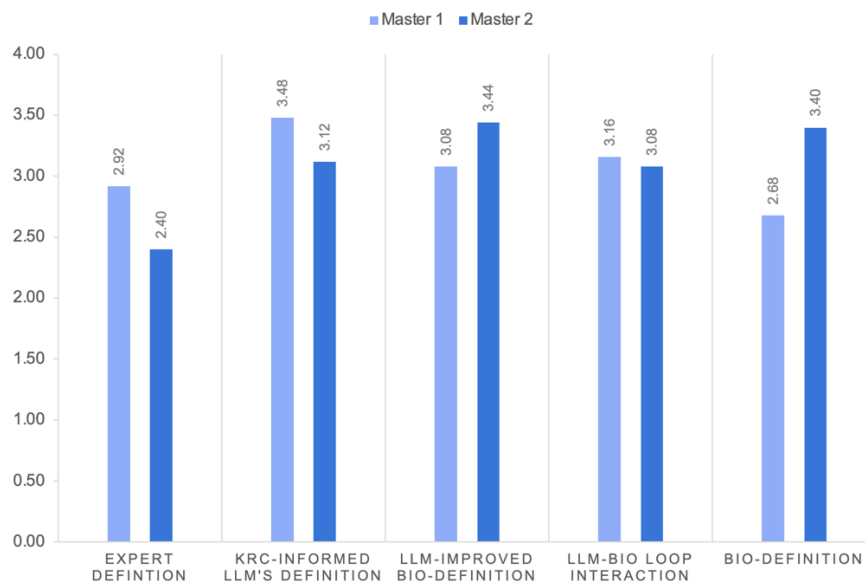


Figure 19. Appraisal of the results of different approaches to crafting definitions by Master 1 and 2 students

Ongoing monitoring of these tendencies in both groups will allow for further investigation on these observations.

Term record drafting and creating a collection of efficient prompts for terminology

The final steps of our study consisted in drafting the term record for *data pollution* in the ARTES DB by taking into account the integration of LLMs into the process, and

devising a scheme for collecting the most efficient prompts for term management and their evaluation.

Figure 20 presents the term record for *data pollution* in the ARTES database, which, alongside IATE, is currently the only online database to offer an entry for this term. Notably, ARTES is the first to include *data pollution* with its current meaning, reflecting the recent semantic shift.

The screenshot displays the ARTES database interface for the term 'data pollution 2'. At the top, there are logos for Université Paris Cité and ALTAE (Langues et Littératures Contemporaines). The main header reads 'Base ARTES AIDE À LA RÉDACTION DE TEXTES SCIENTIFIQUES'. Below this, the term 'data pollution 2' is listed with the note 'nom, non renseigné, terme, fiche complète'. The domain is identified as 'L'environnement et la société. Conservation et protection en général [CDU 2004]'. The title of the master thesis is 'Risks and impacts of the digital era on individuals and the environment', and the record author is Mojca Pecman, Professeure, ALTAE, UFR EILA, UPCité, 2025. A navigation bar includes buttons for MEANING, USE, TRANSLATION, and VIEW. The 'MEANING' section is expanded to show a 'Definition:' with four bullet points: 1) a form of digital-era pollution arising from the excessive and often uncontrolled generation of data, which overwhelms individuals' and systems' ability to retrieve, process, or use relevant information effectively, disrupting professional, social, and daily life [Source: LLM-improved bio-definition]; 2) Data pollution is the interrelated adverse impact that the generation, storing, handling, and processing of digital data has on our natural environment, social environment, and personal environment. It is the unsustainable handling, distribution, and generation of data resources. [Source: Hasselbalch Gry 2022: Data Pollution and Power – White Paper for a Global Sustainable Agenda on AI, The Sustainable AI Lab, Bonn University. (https://uni-bonn.sciebo.de/s/bYOsFyNiZ9sPIY4/download)]; 3) Data pollution is the set of harms generated by economic activities related to data collection, storage and use, which transversally impacts individuals and their living environments. [Source: Mori Leonardo, Alizée Francey, Tobias Mettler 2024: Data Pollution: Definition and Policy Responses. In 16th IFIP WG 8.5 International Conference, Electronic Participation (ePart 2024), Ghent, Belgium, September 3–5, 2024, Proceedings, Pages 147–162.]; 4) specific type of pollution related to information technologies and digital era characterised by the increasing generation of data which is transforming professional, social and everyday life of individuals and their environment in the way that appears not to be in accordance with their capabilities or needs for retrieving relevant information [Source: Mojca Pecman, ALTAE URP 3967, UFR EILA, Université Paris Cité (2025)].

Figure 20. The ARTES DB interface showing the record for *data pollution* with information on its newly emerging meaning

The screenshot in Figure 20 shows the definitions selected for the term record: the expert definition, the bio-definition, and the LLM-improved bio-definition (corresponding to definitions 1, 5, and 3 in Figure 17). Although the expert definition was rated as less successful by Master's students, its inclusion in ARTES is important, as it represents a proposal by subject-matter experts. The bio-drafted definition is mandatory within the ARTES framework, which accounts for the selection of definition 5 as well. We also decided to make a proposal for informing that one of the definitions was constructed with LLM's assistance, in order to avoid blurring between human and LLM-assisted textual data generation.

To highlight the term's dual meanings, ARTES provides two separate records: one for the original meaning and another for the newly emerging one (Figure 21-22), and an extensive note on the concurrent terms (Figure 22).

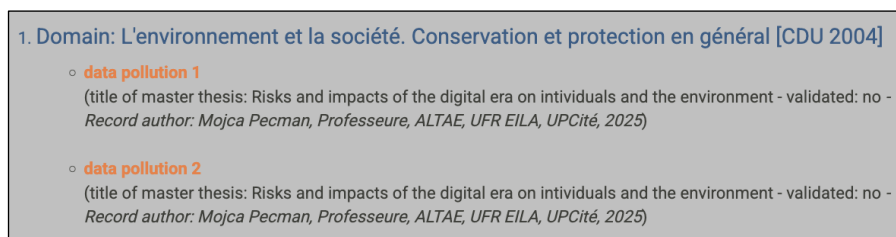


Figure 21. The ARTES DB interface showing the two records for *data pollution* to distinguish the original from the new meaning

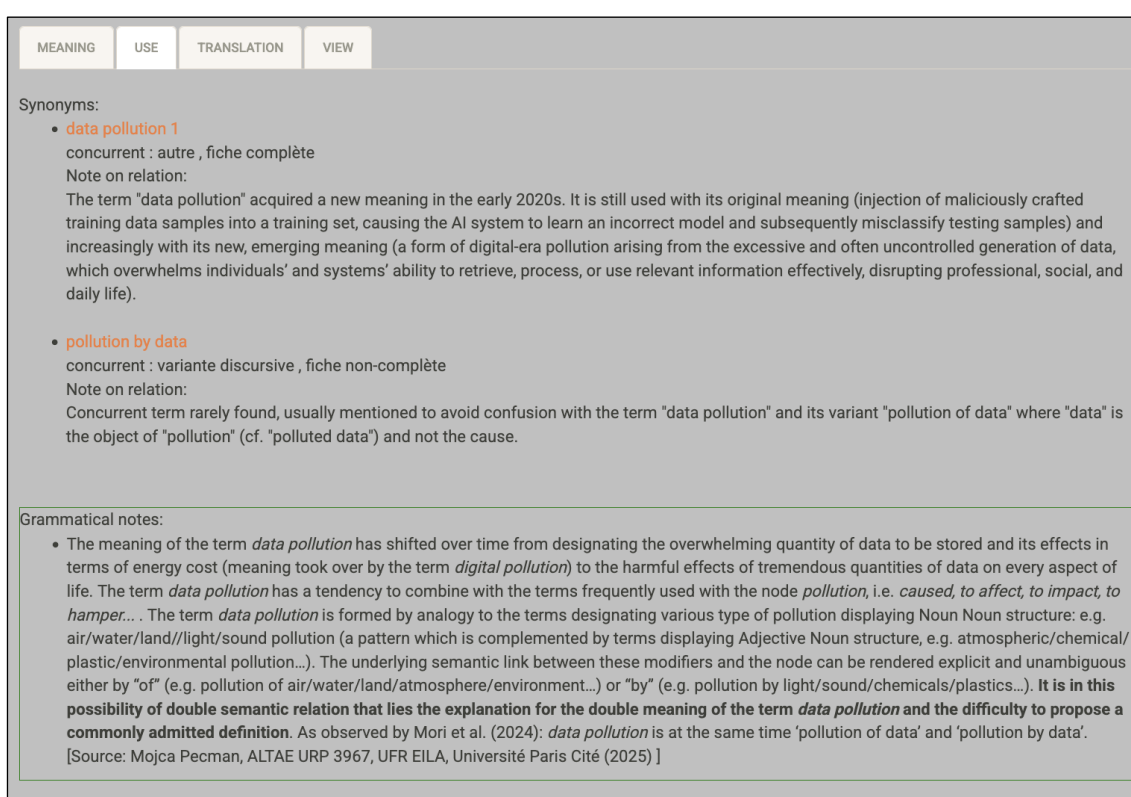


Figure 22. The ARTES DB interface showing the entry for *data pollution 2*, the newly emerged concept, along with the information on *data pollution 1* defined as a concurrent term (and concept) including an explanatory note

Finally, Figure 23 shows the following elements of the scheme devised for creating a collection of efficient prompts for LLMs-assisted term management within ARTES framework by experimental approach: the homepage of the prompt collection interface, a text field for recording the designed prompt, a text field for providing an example of the output produced, the required output size (in characters or words), an assessment of the prompt's quality (based on the obtained output), and the type of prompt technique used.

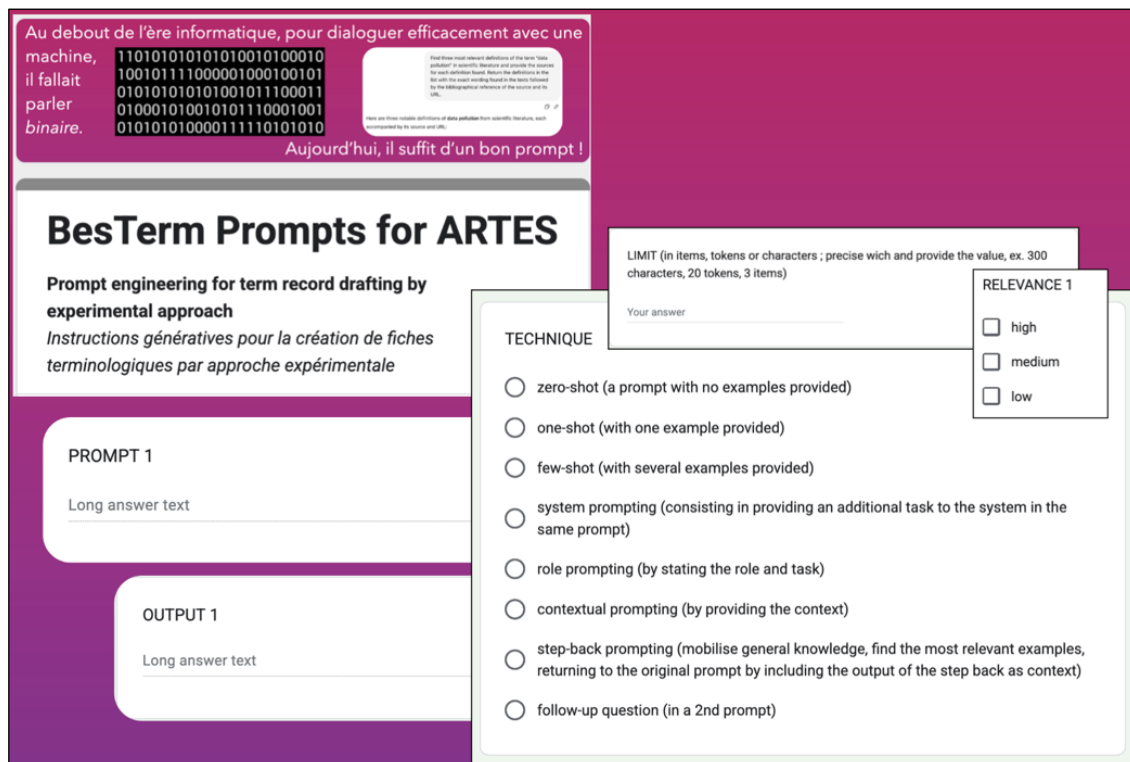


Figure 23. A scheme devised for creating a collection of efficient prompts for LLMs-assisted term management

CONCLUDING REMARKS

This study has explored the potential for integrating GenAI tools into terminological analysis, specifically for drafting term records within the ARTES knowledge base and pedagogical framework used for teaching terminology to Master's students in translation at Université Paris Cité. We selected the term *data pollution* which exhibits semantic shift, as a case study, to conduct an exploratory analysis and construct a protocol for an efficient interaction with LLMs and GenAI tools. This protocol is now being integrated into the curriculum of terminology training within the Master's program. We also conducted an inquiry among 50 Master's students of translation, which showed that half of the students are using LLMs for term record drafting, which further enhances the need to guide them in their efficient use.

The experimental findings of our study highlight the value of corpus-linguistic tools in assisting terminological analysis, enabling the construction and organisation of conceptual knowledge, as well as assessing the quality of the output of GenAI tools. The protocol designed for AI integration to terminological analysis showed that GenAI tools can be useful for revising and improving bio-drafted terminological definitions, but that the human intervention is paramount. These findings align with recent works on integrating AI in teaching and research (cf. Kübler *et al.* 2024; Raus, Mattioda 2024; San Martín 2024; Kübler, Pecman 2025) and reinforce the key recommendations emerging from them,

namely the need to foster a critical approach to AI technologies and to develop efficient methods for evaluating machine-generated output.

The theoretical findings of our study highlight the crucial role of exploratory analyses in the current landscape of rapidly evolving AI tools, particularly in designing effective schemes for integrating GenAI and LLMs. Central to this process is the need for preliminary human analysis. Mastery of domain-specific knowledge and linguistic data is essential for assessing the quality of GenAI output and for crafting effective prompts to address inaccurate responses. In the context of terminology, this requires expertise in specialised concepts and the application of corpus-based approaches to term analysis. Corpus linguistics emerges as a key method for informed interaction with GenAI tools and prompt engineering.

A noteworthy challenge encountered during our study was the cognitive bias introduced by deliberately limiting LLM interactions, which demanded considerable self-discipline. In response, the protocol proposed in this paper advocates a sequential integration of GenAI tools across various stages of information retrieval and terminological analysis.

Future research will focus on testing and refining this protocol with translation students in terminology courses. We aim to expand the methodology to cover all stages of term record drafting, with the objective of enhancing the retrieval and generation of terminologically relevant information in interaction with GenAI tools. One of the protocol's core goals is to develop a curated collection of highly effective prompts for term record creation. Our study has laid the groundwork for this, by offering a model for prompt construction and evaluation, tailored to encompass different language models, task types, and linguistic datasets, and that we will continue to test, adapt and refine in further studies.

REFERENCES

- Altameemi Yaser M. 2024: State-of-the-art Review of the Corpus Linguistics Field from the Beginning Until the Development of ChatGPT. – *Theory and Practice in Language Studies* 14(2), 423–431. Available at: <https://doi.org/10.17507/tppls.1402.13>.
- ARTES. Available at: <https://artes.app.univ-paris-diderot.fr/artes-symfony/web/app.php>.
- Ballier Nicolas, Cho Dahn, Faye Bilal, Ke Zong-You, Martikainen Hanna, Pecman Mojca, Wisniewski Guillaume, Yunès Jean-Baptiste, Zhu Lichao, Zimina-Poirot Maria 2021: The SPECTRANS System Description for the WMT21 Terminology Task. – *ACL Rolling Review, a New Initiative of the Association for Computational Linguistics*, 818–825. Available at: <http://www.statmt.org/wmt21/pdf/2021.wmt-1.80.pdf>.
- Bénard Maud, Mestivier Alexandra, Kubler Natalie, Zhu Lichao, Bawden Rachel, de la Clergerie Eric, Romary Laurent, Huguin Mathilde, Nominé Jean-Francois, Peng Ziqian, Yvon François 2023: MaTOS: Traduction automatique pour la science ouverte [MaTOS: Machine Translation for Open Science]. – *Actes de l'atelier*

- “Analyse et Recherche de Textes Scientifiques” (ARTS)@TALN 2023, 8–15. Available at: <https://aclanthology.org/2023.jeptalnrecital-arts.2.pdf>.
- Boonstra Lee 2025: *Prompt Engineering*. White paper. Available at: <https://www.kaggle.com/whitepaper-prompt-engineering>.
- Bourigault Didier, L’Homme Marie-Claude, Jacquemin Christian 2001: *Recent Advances in Computational Terminology*, Amsterdam/Philadelphia: John Benjamins.
- Bourigault Didier, Slodzian Monique 1999: Pour une terminologie textuelle [Towards a Textual Terminology]. – *Terminologies nouvelles* 19, 29–32.
- Casal J. Elliott, Kessler Matt 2023: Can Linguists Distinguish Between ChatGPT/AI and Human Writing? A Study of Research Ethics and Academic Publishing. – *Research Methods in Applied Linguistics* 2(3), 100068. Available at: <https://doi.org/10.1016/j.rmal.2023.100068>.
- ChatGPT, based on GPT-4-turbo, released on the 6th November 2023, OpenAI. Available at: <https://chatgpt.com>.
- Condamines Anne 2005: Linguistique de corpus et terminologie [Corpus Linguistics and Terminology]. – *Langages* 157. *La terminologie: nature et enjeux*, ed. L. Depecker, 36–47.
- Condamines Anne 2022: Terminologie, Intelligence Artificielle et Psychologie cognitive: réflexions sur les interactions possibles dans l’étude de la variation en langues spécialisées [Terminology, Artificial Intelligence and Cognitive Psychology: Reflections on Possible Interactions in the Study of Variation in Specialised Languages]. – *De Europa. European and Global Studies Journal*, Special Issue on Multilingualism and Language Varieties in Europe in the Age of Artificial Intelligence, Università di Torino, 131–148.
- Condamines Anne, Rebeyrolle Josette 2001: Searching for and Identifying Conceptual Relationships via a Corpus-Based Approach to a Terminological Knowledge Base (CTKB). – *Recent Advances in Computational Terminology*, ed. D. Bourigault, M.-C. L’Homme, C. Jacquemin, Amsterdam/Philadelphia: John Benjamins, 127–148.
- Davies Mark 2025: *Corpora and AI/LLM: Comparison of the Predictions of LLMs (ChatGPT and Gemini) to Actual Corpus Data (mainly from English-Corpora.org)*. White paper. Available at: <https://www.english-corpora.org/ai-llms/english-corpora-with-ai-llms.pdf>, alongside a video posted on the 18th of March 2025. Available at: <https://youtu.be/WAZWoNzhZ9A?si=JT6TAhOmlkXjUoa4>.
- DeepSeek, based on DeepSeek-V3-0324, 深度求索 (DeepSeek). Available at: <https://chat.deepseek.com>.
- EMT Competence Framework 2022. Available at: https://commission.europa.eu/system/files/2022-11/emt_competence_fw_2022_en.pdf.

- Falgas Julien, Robert Pascal 2023: Présenter l'IA comme une évidence, c'est empêcher de réfléchir au numérique (repris sur le titre: Comment le discours médiatique sur l'IA empêche d'envisager d'autres possibles) [Presenting AI as a Fact Prevents Reflection on Digital Technologies (based on the title: How Media Discourse on AI Prevents Considering Other Possibilities)]. – *The conversation*, 19 octobre 2023. Available at: <https://theconversation.com/comment-le-discours-mediatique-sur-lia-empêche-d'envisager-d'autres-possibles-211766>.
- Finardi Kyria 2023: *The Paradox of our Time and Role of Applied Linguists in it*: Conférence donnée à l'occasion du webinaire 2023 de l'AFLA. Available at: <http://www.afla-asso.org/webinaire-2023>.
- Gledhill Christopher, Kübler Natalie 2015: How Trainee Translators Analyse Lexico-Grammatical Patterns. – *Journal of Social Sciences* 11(3): Special issue on Phraseology, Phraseodidactics and Construction Grammar(s), ed. M. I. González-Rey, 162–178.
- Humbley John 2022: The Reception of Wüster's General Theory of Terminology. – *Theoretical Perspectives on Terminology: Explaining Terms, Concepts and Specialized Knowledge*, ed. P. Faber, M.-C. L'Homme, Book coll. "Terminology and Lexicography Research and Practice", Amsterdam/Philadelphia: John Benjamins, 15–36.
- IATE. Available at: <http://iate.europa.eu>.
- IGI Global Dictionary. Available at: <https://www.igi-global.com/dictionary>.
- ISO – TC 37 Language and Terminology – ISO Standard 1087:2019(en) Terminology Work and Terminology Science: Vocabulary. Available at: <https://www.iso.org/obp/ui/#iso:std:iso:1087:ed-2:v1:en>.
- ISO – TC 37 Language and Terminology – ISO Standard 704:2022(en) Terminology Work – Principles and Methods. Available at: <https://www.iso.org/obp/ui/en/#iso:std:iso:704:ed-4:v1:en>.
- ISO/TC 37/SC 3/WG 6 Terminology Management and Artificial Intelligence – ISO Standard 5078:2025(en) Management of Terminology Resources – Terminology Extraction. Available at: <https://www.iso.org/obp/ui/en/#iso:std:iso:5078:ed-1:v1:en>.
- ISO/TC 37/SC 3/WG 6 Terminology Management and Artificial Intelligence – ISO Standard 12620-1:2022 Management of Terminology Resources – Data Categories – Part 1: Specifications. Available at: <https://www.iso.org/obp/ui/en/#iso:std:iso:12620:-1:ed-1:v1:en>.
- Kilgarriff Adam, Baisa Vít, Bušta Jan, Jakubíček Miloš, Kovář Vojtěch, Michelfeit Jan, Rychlý Pavel, Suchomel Vít 2014: The Sketch Engine: Ten Years On. – *Lexicography* 1, 7–36.

- Kübler Natalie, Martikainen Hanna, Mestivier Alexandra, Pecman Mojca 2024: Post-editing Neural Machine Translation in Specialised Languages: the Role of Corpora in the Translation of Phraseological Structures. – *Recent Advances in Multiword Units in Machine Translation and Translation Technology*, ed. I. J. Monti, G. Corpas Pastor, R. Mitkov, C. Manuel Hidalgo-Ternero, *Current Issues in Linguistic Theory* 366, Amsterdam/Philadelphia: John Benjamins, 57–78.
- Kübler Natalie, Pecman Mojca 2012: The ARTES Bilingual LSP Dictionary: from Collocation to Higher Order Phraseology. – *Electronic Lexicography*, ed. S. Granger, M. Paquot, Oxford: Oxford University Press, 187–209.
- Kübler Natalie, Pecman Mojca 2025: *Corpus Linguistics: a Dinosaur or an Elephant in the IA Room? A Word from Terminology and Translation Studies Perspective*: Phd Seminar – Winter school – Foreign Literatures and Languages school, 5–6 February 2025, Università di Verona, Italy.
- Leonardi Natascia 2025: Terminology Science, International Languages, and Knowledge Communication. – *Terminology Throughout History. A discipline in the Making*, ed. K. Warburton, J. Humbley, *Terminology and Lexicography Research and Practice* 24, Amsterdam/Philadelphia: John Benjamins Publishing Company, 148–166.
- L’Homme Marie-Claude 2024a: Managing Polysemy in Terminological Resources. – *Terminology* 30(2), 216–249.
- L’Homme Marie-Claude 2024b: Seven Good Reasons for a Better Account of Fine-grained Polysemy in Terminological Resources. – *Terminologija* 31, 6–23. Available at: <https://journals.lki.lt/terminologija/article/view/2407/2408>.
- Lombard Alizée, Huyghe Richard, Barque Lucie, Gras Doriane 2023: Regular Polysemy and Novel Word-sense Identification. – *The Mental Lexicon* 18(1), 94–119.
- Lommel Arle 2024: The Rise of Large Language Models Informed by not so Large Corpora of Training Data. – *Digital Translation* 11/1, 73–84. Available at: <https://doi.org/10.1075/dt.24006.lom>.
- Meyer Ingrid 2001: Extracting Knowledge-Rich Contexts for Terminography: A Conceptual and Methodological Framework. – *Recent Advances in Computational Terminology*, ed. D. Bourigault, Ch. Jacquemin, M.-C. L’Homme, Amsterdam/Philadelphia: John Benjamins, 279–302.
- Meyer Ingrid, Bowker Lynne, Eck Karen 1992: Cogniterm: An Experiment in Building a Terminological Knowledge Base. – *Proceedings of 5th EURALEX International Congress on Lexicography*, Tampere, Finland. Tampere: Studia Translatologica, 159–172.
- Michel Jean-Baptiste, Shen Yuan Kui, Aiden Aviva Presser, Veres Adrian, Gray Matthew K., Brockman William, The Google Books Team, Pickett Joseph P., Hoiberg Dale,

- Clancy Dan, Norvig Peter, Orwant Jon, Pinker Steven, Nowak Martin A., Aiden Erez Lieberman 2011: Quantitative Analysis of Culture Using Millions of Digitized Books. – *Science* 331(6014), 176–82. Available at: <https://pubmed.ncbi.nlm.nih.gov/21163965>.
- Pecman Mojca 2012: Tentativeness in Term Formation: a Study of Neology as a Rhetorical Device in Scientific Papers. – *Neology in Specialized Communication*, ed. M. Teresa Cabré Castellví, Rosa Estopá Bagot, Maria C. Vargas Sierra. – Special issue of *Terminology* 18(1), 27–58.
- Pecman Mojca 2014: Variation as a Cognitive Device: How Scientists Construct Knowledge Through Term Formation. – *Terminology* 20(1), 1–24.
- Pecman Mojca 2018: *Langue et construction de connaisSENSes. Energie lexico-discursive et potentiel sémiotique des sciences* [Language and the Construction of Meaning and Knowledge: Lexico-Discursive Energy and the Semiotic Potential of Science]: monographie, préface de M.-C. L’Homme. Paris: Editions L’Harmattan.
- Pecman Mojca 2021: 10th Anniversary of the ARTES Terminological and Phraseological Database. – *Svijet od riječi: terminološki i terminografski ogledi* (“*The world of words: Terminological and terminographic discussions*”), ed. I. Brač, A. Ostroški Anić, Izdavalacko izdanje Instituta za hrvatski jezik i jezikoslovlje (Publication of the Institut of Croatian language and linguistics, Zagreb), 301–324.
- Pecman Mojca, Kübler Natalie 2011: ARTES: An Online Lexical Database for Research and Teaching in Specialized Translation and Communication. – *Proceedings from International Workshop on Lexical Resources (WoLeR) 2011 at ESSLLI*, 1–5 August 2011, Ljubljana, Slovenia, 87–93. Available at: http://alpage.inria.fr/~sagot/pub/WoLeR_2011_proceedings.pdf.
- Pecman Mojca, Kübler Natalie 2022: Text Genres and Terminology. – *Theoretical Perspectives on Terminology: Explaining Terms, Concepts and Specialized Knowledge*, ed. P. Faber, M.-C. L’Homme, Book coll. Terminology and Lexicography Research and Practice, Amsterdam/Philadelphia: John Benjamins, 263–289.
- Perplexity, based on GPT-4o/Claude 4.0 Sonnet, released on the 6th of May 2025, Perplexity AI. Available at: <https://www.perplexity.ai>.
- Ptasznik Bartosz, Robert Lew 2025: *Dictionaries Versus AI Tools Through the Eyes of English Majors: International Journal of Lexicography* 38(2), 140–158. Available at: <https://doi.org/10.1093/ijl/ecaf005>.
- Rastier François 2021: Data vs corpora. – *L’intelligence artificielle des textes. Des algorithmes à l’interprétation* [The Artificial Intelligence of Texts: From Algorithms to Interpretation] 15, ed. D. Mayaffre, L. Vanni, coll. Lettres numériques, Paris: Éditions Honoré Champion, 203–246.

- Raus Rachele, Mattioda Maria Margherita 2024: L'intelligence artificielle en salle de classe: la perception des étudiantes et des étudiants [Artificial Intelligence in the Classroom: Students' Perceptions]. – *Multilinguisme européen et IA: entre droit, traduction et didactique des langues/Multilinguismo europeo e IA tra diritto, traduzione e didattica delle lingue/European Multilingualism and Artificial Intelligence: The Impacts on Law, Translation and Language Teaching*, ed. R. Raus, F. Bisiani, M. M. Mattioda, M. Tonti, De Europa, Special Issue, 221–231.
- Renouf Antoinette 2020: Semantic Neology. Challenges in Matching Corpus-based Semantic Change to Real-world Change. – *Corpora and the Changing Society. Studies in the Evolution of English*, ed. P. Rautioaho, A. Nurmi, J. Klemola, Amsterdam/Philadelphia: John Benjamins, 79–112.
- Riehmann Patrick, Gruendl Henning, Potthast Martin, Trenkmann Martin, Stein Benno, Fröhlich Bernd 2012: WORDGRAPH: Keyword-in-Context Visualization for NETSPEAK's Wildcard Search. – *IEEE Transactions on Visualization and Computer Graphics* 18(9), 1411–1423. Available at: https://downloads.webis.de/publications/papers/riehmann_2012.pdf.
- San Martín Antonio P. 2024: What Generative Artificial Intelligence Means for Terminological Definitions. – *Proceedings of the 3rd International Conference on Multilingual Digital Terminology Today (MDTT 2024)*, Granada: CEUR-WS. Available at: <https://ceur-ws.org/Vol-3703/paper1.pdf>.
- de Schryver Gilles-Maurice 2023: Generative AI and Lexicography: The Current State of the Art Using ChatGPT. – *International Journal of Lexicography* 36(4), 355–387.
- Schulhoff Sander, Michael Ilie, Nishant Balepur, Konstantine Kahadze, Amanda Liu, Chenglei Si, Yinheng Li, Aayush Gupta, HyoJung Han, Sevien Schulhoff, Pranav Sandeep Dulepet, Saurav Vidyadhara, Dayeon Ki, Sweta Agrawal, Chau Pham, Gerson Kroiz, Feileen Li, Hudson Tao, Ashay Srivastava, Hevander Da Costa, Saloni Gupta, Megan L. Rogers, Inna Goncarencu, Giuseppe Sarli, Igor Galyunker, Denis Peskoff, Marine Carpuat, Jules White, Shyamal Anadkat, Alexander Hoyle, Philip Resnik 2025: *The Prompt Report: A Systematic Survey of Prompt Engineering Techniques*. Available at: <https://arxiv.org/pdf/2406.06608>.
- Termium Plus. Available at: <https://www.btb.termiumplus.gc.ca>.
- UNTERM. Available at: <https://unterm.un.org/unterm2/en>.
- Vezzani Federica, Di Nunzio Giorgio, Maria Costa Rute 2023: ISO Standards for Terminology Resources Management. Are They FAIR Enough? – *Digital Translation* 10/2, Amsterdam/Philadelphia: John Benjamins, 233–252. Available at: <https://doi.org/10.1075/dt.00009.vez>.
- Vitrine Linguistique. Available at: <https://vitrinelinguistique.oqlf.gouv.qc.ca>.

- Wilkinson Mark D., Dumontier Michel, Aalbersberg IJsbrand Jan, Appleton Gabrielle 2016: The FAIR Guiding Principles for Scientific Data Management and Stewardship. – *Scientific Data* 3, 160018. Available at: <https://doi.org/10.1038/sdata.2016.18>.
- WIPO. Available at: <https://wipopearl.wipo.int/fr/linguistic>.
- Wüster Eugen 1968: *The Machine Tool./La Machine-Outil. An Interlingual Dictionary of Basic Concepts Comprising an Alphabetical Dictionary and a Classified Vocabulary With Definitions and Illustrations: English-French Master Volume*, Londres, Technical Press. Aslib, 173 p.
- Zhu Lichao, Zimina Maria, Bénard Maud, Namdar Behnoosh, Ballier Nicolas, Wisniewski Guillaume, Yunès Jean-Baptiste 2023: Investigating Techniques for a Deeper Understanding of Neural Machine Translation (NMT) Systems Through Data Filtering and Fine-tuning Strategies. – *Proceedings of the Eighth Conference on Machine Translation*, Singapore, Association for Computational Linguistics, 275–281.

References of the texts on data pollution (used for the corpus and/or appearing in the outputs of GenAI tools)

- Ben-Shahar Omri 2018: Data Pollution. – *University of Chicago Public Law & Legal Theory Paper Series* 679, 104–159.
- Cao Yinzhi, Fangxiao Yu Alexander, Aday Andrew, Stahl Eric, Merwine Jon, Yang Junfeng 2018: Efficient Repair of Polluted Machine Learning Systems via Causal Unlearning. – *Proceedings of 2018 ACM Asia Conference on Computer and Communications Security*, Incheon, Republic of Korea, 4–8 June 2018 (ASIA CCS '18).
- Castelluccia Claude, Kaafar Mohamed Ali 2009: Owner-Centric Networking (OCN): Toward a Data Pollution-Free Internet. – *Proceedings of the 2009 Ninth Annual International Symposium on Applications and the Internet*, 169–172.
- Gertsen Fred 2022: *Data Pollution and Taxation*: Master's Thesis, Executive Master Cyber Security, Defended on the 31st of March 2022, Universiteit Leiden, Netherlands.
- Hasselbalch Gry 2021: *Data Ethics of Power: A Human Approach in the Big Data and AI Era*, Glos (UK): Edward Elgar.
- Hasselbalch Gry 2022: *Data Pollution & Power*, White Paper for a Global Sustainable Agenda on AI, The Sustainable AI Lab, Bonn University.
- Mori Leonardo, Francey Alizée, Mettler Tobias 2024: Data Pollution: Definition and Policy Responses. – *Proceedings of the 16th IFIP WG 8.5 International Conference, Electronic Participation (ePart 2024)*, Ghent, Belgium, September 3–5 2024, 147–162.
- Zimmerli Walther Ch. 1986: Who is to Blame for Data Pollution? On Individual Moral Responsibility With Information Technology. – *Philosophy and Technology II*:

Information Technology and Computers in Theory and Practice, ed. C. Mitcham, A. Huning, Dordrecht/Boston: D. Reidel, 291–305.

APPENDIX

Figure 24 shows a prompt without the specification “with the exact wording found in the texts”, and the output consisting of two erroneous, made-up contexts, where the definition of *information pollution* is credited to *data pollution*. Nevertheless, the proposal by ChatGPT that “data pollution encompasses the contamination of information sources...”, can be regarded as true, if we take into account the semantic relation between *data pollution* and *information pollution* revealed by the discursive marker of meronymy “encompasses”, and according to which *information pollution* is a part of *data pollution*. The reformulation by LLMs “The concept of data pollution includes information pollution” into “data pollution encompasses [definition of *information pollution*] ...” is respectful of knowledge representation. However, in terminological analysis, it is paramount to distinguish semantically related terms, all the more so when they are near-like synonyms.

GOAL	Finding definitional contexts and their references
MODEL	ChatGPT-4-turbo
DATE	13/05/2025
TECHNIQUE	System prompting
LIMIT	3 definitions
PROMPT	Find three most relevant definitions of the term “data pollution” in scientific literature and provide the sources for each definition found. Return the definitions in the list followed by the bibliographical reference of the source and its URL.
OUTPUT	<p>Data pollution refers to the harmful “data emissions” leaked into the digital ecosystem, disrupting social institutions and public interests. This concept rethinks the harms the data economy creates and the way they have to be regulated.</p> <p>Ben-Shahar O. 2018: <i>Data Pollution</i> 679, University of Chicago Public Law & Legal Theory Paper Series. Available at: https://chicagounbound.uchicago.edu/public_law_and_legal_theory/699</p> <p>Data pollution is the interrelated adverse impact that the generation, storing, handling, and processing of digital data has on our natural environment, social environment, and personal environment. It is the unsustainable handling, distribution, and generation of data resources.</p> <p>Hasselbalch G. <i>et al.</i> 2022: <i>Data Pollution & Power - White Paper for a Global Sustainable Development Agenda on AI</i>. Available at: https://www.researchgate.net/publication/362174889_Data_Pollution_Power_-_White_Paper_for_a_Global_Sustainable_Development_Agenda_on_AI_by_Gry_Hasselbalch_with_contributions_from_the_Data_Pollution_Power_DPP_Group_at_Bonn_Sustainable_AI_Lab</p>

	Data pollution encompasses the contamination of information sources with irrelevant, redundant, unsolicited, and low-value information. This includes the impact of data sharing on others, where one's data sharing (or that of corporations) affects others, leading to broader societal harms. <i>Data Pollution and Taxation 2022</i> . Available at: https://www.academia.edu/76896486/Data_Pollution_and_Taxation
COM-MENT	1 st reformulated context, not a quote; source provided and well-constructed 2 nd a definitional context, a quote; source provided and well-constructed 3 rd reformulated context, not a quote, misleading (it is a metonymic definition by silent reference to <i>information pollution</i>); source provided but poorly constructed, not a scientific register (academic register)
RELE-VANCE	low

Figure 24. Unsuccessful prompt for finding definitional contexts and their references with ChatGPT

Figure 25 illustrates the risks associated with using an insufficiently constrained prompt when managing terms undergoing semantic shift or exhibiting semantic instability, notably resulting in conceptual ambiguity. Without using step-back prompting to mobilise LLM’s knowledge, Chat GPT retrieved only the first meaning of the term *data pollution*. It is however not surprising that ChatGPT, as a predictive tool, selected the original meaning, which is more widely used. Nevertheless, this example also emphasises the role of corpus-based approach, without which, using directly GenAI tools can lead to inaccurate terminological analysis. We can also notice that sources are not provided, despite the request, emphasising the importance to constrain the system appropriately to provide them.

GOAL	Providing a definition with sources used to construct it		
DATE	13/05/2025		
TECHNIQUE	Zero shot		
LIMIT	300 characters		
PROMPT	Define the term “data pollution” in no more than 300 characters. List the source(s) used for drafting this definition.		
MODEL	ChatGPT-4-turbo	DeepSeek-V3	Perplexity
OUTPUT	Data pollution refers to the presence of inaccurate, misleading, or irrelevant data that contaminates datasets, leading to poor analysis, faulty decisions, or biased outcomes in	Data pollution refers to the contamination or degradation of data quality due to errors, inconsistencies, or misleading information, often	Data pollution refers to the presence of inaccurate, duplicate, unauthentic, or damaging data within datasets, which degrades data quality and can impair analysis, AI training, and

	<p>data-driven processes.</p> <p>The definition I provided is based on general knowledge synthesized from reputable sources on data quality and data management, including:</p> <ul style="list-style-type: none"> - Academic literature on data integrity and quality (e.g., research from journals like <i>Information Systems Research</i>). - Industry best practices from organizations like the Data Management Association (DAMA). - Common definitions and discussions from authoritative sources such as IBM, Gartner, and data science communities. <p>No single direct source was quoted; instead, the definition is a concise summary informed by a broad understanding of the term as used in data science and information systems. Would you like references to specific academic papers or articles on the topic?</p>	<p>leading to flawed analysis and decision-making. (299 characters)</p> <p>I didn't use any specific external sources to draft the definition of “data pollution.” Instead, I relied on my general knowledge of data science and technology terminology.</p> <p>If you'd like an academically or professionally cited definition, I can look up authoritative sources (e.g., research papers, industry reports, or standards organizations). Let me know how you'd like to proceed!</p>	<p>decision-making processes. It includes errors, inconsistencies, and misleading information contaminating otherwise clean data.</p> <p>The definition of “data pollution” provided is primarily based on the conceptual framework and detailed discussion from Omri Ben-Shahar's 2019 article <i>Data Pollution</i> in the <i>Journal of Legal Analysis</i> (Oxford Academic), which explores the harms caused by data activities and the analogy to environmental pollution. Additional insights were drawn from a 2023 article on LinkedIn explaining data pollution as degradation and contamination of data affecting quality and trustworthiness. These sources together inform the broad understanding of data pollution as contamination or harm caused by inaccurate, excessive, or misused data.</p>
COMMENT	the original meaning of the term, source unprovided		
RELEVANCE	low		

Figure 25. Unsuccessful prompt for defining a term that acquired a new meaning