Exploring the Human–AI Continuum in Terminology Management: Towards Evaluation Criteria for LLM-Assisted Terminology Work *

Natascia Ralli[1,*], Giorgio Maria Di Nunzio[2], Hanne Smaadahl[3], Ana Salgado[4], Federica Vezzani[2], Klaus-Dirk Schmitz[5], Sara Carvalho[6], Klaus Fleischmann[7] and Fred Bane[8]

[1] *Institute for Applied Linguistics, Eurac Research, Bolzano, Italy*
[2] *Università degli Studi di Padova, Padova, Italy*
[3] *SAP, Walldorf, Germany*
[4] *Centre for Linguistics of the University of Porto, Faculty of Arts and Humanities of the University of Porto, Porto, Portugal*
[5] *TH Köln, Cologne, Germany*
[6] *Centre for Languages, Literatures, and Cultures (CLLC), University of Aveiro, Aveiro, Portugal*
[7] *Kaleidoskope, Vienna, Austria*
[8] *TransPerfect, New York, NY, USA*

Abstract

This paper explores two terminology management activities that lie at opposite ends of the automation spectrum within the terminology workflow: term extraction, a largely automatable, repetitive task in which human oversight can be limited, and definition generation, a conceptually demanding activity that requires strong human involvement. Based on initial considerations on the effective and appropriate use of LLM-generated outputs, the paper investigates the key aspects to consider when applying LLMs to both tasks. The outcome of this analysis is a set of evaluation criteria for LLM-assisted term extraction and LLM-assisted definition generation, with particular emphasis on positioning these tasks along the human –AI continuum in order to support quality, reliability, and accountability in terminology work.

Keywords

Human–AI continuum, term extraction, definition generation, terminology management, evaluation criteria

1. Introduction

Managed terminology cannot be an afterthought to other linguistic assets. Serving as the definitive source of information for terms and concepts in an organization, a terminology database (termbase) must be established before, or in parallel with, ontologies, taxonomies, knowledge graphs, and the content itself [1].

Artificial Intelligence (AI) technologies can enhance human expertise and support activities aimed at ensuring consistency, completeness, and reliability of terminological data in a termbase while maintaining human responsibility and oversight [2]. However, human interactions in data collection will remain central as large language model-generated content proliferates online [3]. Particularly with AI assistance, terminological data itself requires human curation to maintain its integrity as a trusted and verifiable resource, as solely relying on probabilistic generation can compromise its reliability [4]. Professional terminology work requires activities such as delimiting

* *5th International Conference on "Multilingual digital terminology today. Design, representation formats and management systems" (MDTT) 2026, June 25-26, 2026, Zadar, Croatia.*

1* Corresponding author.

natascia.ralli@eurac.edu (N. Ralli); giorgiomaria.dinunzio@unipd.it (G. M. Di Nunzio); hanne.smaadahl@sap.com (H. Smaadahl); acsalgado@letras.up.pt (A. Salgado); federica.vezzani@unipd.it (F. Vezzani); klaus.schmitz@th-koeln.de (K.-D. Schmitz); sara.carvalho@ua.pt (S. Carvalho); klaus.fleischmann@kaleidoscope.at (K. Fleischmann); fbane@translations.com (F. Bane)

0000-0002-9663-8169 (N. Ralli); 0000-0001-9709-6392 (G. M. Di Nunzio); 0009-0009-4164-8536 (H. Smaadahl); 0000-0002-6670-3564 (A. Salgado); 0000-0003-2240-6127 (F. Vezzani); 0009-0002-3610-4090 (K.-D. Schmitz); 0000-0002-7501-5405 (S. Carvalho); 0009-0009-7230-0620 (F. Bane)

the domain, distinguishing concepts by essential characteristics, and maintaining systematic relations among concepts and designations. AI applications may lack this domain-specific understanding because many current systems, especially general-purpose large language models (LLMs), produce outputs by optimizing linguistic plausibility based on training data patterns, rather than relying on an explicit, validated model of a domain's concepts, and therefore may replicate patterns found in specialized discourse. Consequently, over-reliance on automated tools in terminology management could overlook deeper conceptual connections and inconsistencies. This may lead to disregarding diverse linguistic expressions, promoting overly rigid standardized terminology, and perpetuating harmful bias or inequities [5].

AI is an umbrella term for technologies that enable machines to perform tasks commonly associated with human intelligence, such as learning, reasoning, and problem-solving [5]. While a range of AI approaches, including Natural Language Processing (NLP), have long been applied in terminology management [6], this study intends to investigate LLMs used in generative artificial intelligence (GenAI) applications. This focus reflects the view that these applications currently present the most significant opportunities and challenges, as well as risks, for enhancing and further automating terminology management tasks [7]. The rapid dissemination and increasing adoption of LLM-based tools are currently reshaping expectations about what can be automated in terminology work, often faster than methodological reflection and quality assurance practices can follow. By giving LLMs center stage, we address the technologies most likely to be adopted (or imposed) in day-to-day terminology management, and we assess their implications for reliability, traceability, and expert accountability. At the same time, we recognize that other AI paradigms, particularly knowledge engineering approaches such as ontologies, may be naturally compatible with concept-oriented terminology principles, and they remain an important point of comparison.

Within this framework, this paper investigates the role of LLMs in terminology management by examining two activities that lie at opposite ends of the automation spectrum within the terminology workflow: term extraction, commonly framed as generating candidate terms from corpora, and often treated as a largely automatable and repetitive task where human oversight can be limited, and definition generation, a conceptually demanding activity requiring strong human oversight. Thus, this paper aims to propose evaluation criteria for LLM-assisted term extraction and definition generation, with particular emphasis on positioning these tasks along the human–AI continuum (Section 1.1).

Throughout this paper, we distinguish between LLM-assisted and LLM-driven workflows. We use LLM-assisted to refer to settings in which an LLM supports human experts by proposing intermediate outputs, such as a list of candidate terms extracted from corpora or draft definitions, while selection, validation, and quality control remain the responsibility of the terminologist. We use LLM-driven for settings in which the LLM is the primary decision-making component and produces the operational output, such as a final ranked term list or a definition intended for direct integration, while human involvement is limited to monitoring or post hoc review. This distinction matters for evaluation, as acceptable quality thresholds, audit requirements, and governance differ depending on whether the model is a drafting aid or the main driver of the workflow.

In addition, we distinguish between two analytical levels: (i) the zones of the human–AI continuum (human-driven, collaborative, and AI-driven), and (ii) the role played by LLMs within a workflow (LLM-assisted vs. LLM-driven). The former describes degrees of human and AI involvement, whereas the latter describes how the model functions operationally within the task. LLM-assisted workflows typically fall within the collaborative zone, whereas LLM-driven workflows are situated closer to the AI-driven end of the continuum. Human-driven workflows may include digital support tools, but decision-making and validation remain entirely with the human expert.

The paper is structured as follows: Section 1.1 introduces the human–AI continuum in the context of terminology management; Section 2 reviews the current approaches in what concerns term extraction and definition generation; Section 3 outlines preliminary considerations that must be addressed before using LLMs in terminology workflows to ensure effective and responsible use;

Section 4 constitutes the core of the paper, with Section 4.1 proposing evaluation criteria for LLM-assisted term extraction, while Section 4.2 applies a corresponding framework to LLM-assisted definition generation; Section 5 concludes with final remarks and directions for future work.

1.1. The human–AI continuum in terminology management

Terminology management involves a variety of activities, including terminological data creation and maintenance, as well as terminology systems management [8, 9]. Within terminological data creation and maintenance, activities include extracting terminology to identify and select relevant terms and concepts, conducting research to verify the accuracy of information, and documenting each concept with both optional and mandatory metadata. In a multilingual termbase, terminology work is carried out in all relevant languages. In terms of terminology systems management, activities include maintaining terminology requirements and specifications, which may be organization-specific, regularly reviewing and adapting the data model to support current and future needs, and creating and maintaining workflows and user authorization.

Given the broad range of these activities, terminology management can be situated along a continuum of human-driven vs. AI-driven processes (Figure 1):

- Human-driven: At one end of the continuum, human experts retain full control and accountability. While AI may assist to varying degrees by providing suggestions, all critical judgments and validations are conducted by humans.
- Collaborative: This middle zone represents a partnership between humans and AI, where both contribute to the workflow. AI performs significant portions of the task, while humans supervise, validate, interpret, and intervene as necessary.
- AI-driven: In this zone, AI operates with minimal or no human intervention. It is suitable for low-risk, repetitive, or highly structured tasks, with humans potentially monitoring outcomes. These processes are continually enhanced through strategies such as quality monitoring and improvement, and implementing corrective measures independently of the primary workflow.

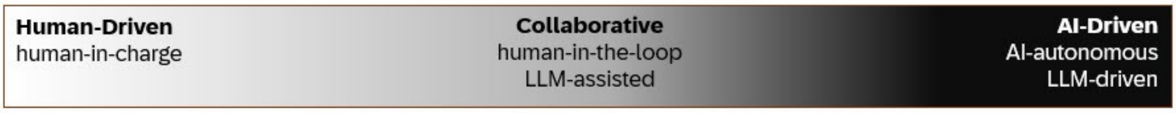

Figure 1: Human–AI continuum.[1]

These zones correspond broadly to distinctions discussed in AI governance literature, but for clarity this paper uses only the labels “human-driven”, “collaborative”, and “AI-driven”. Along this human–AI continuum, human-driven activities include (see [5, 10]):

- Maintaining terminology requirements and specifications.
- Creating and maintaining workflows and user authorization.
- Ensuring inclusion of relevant, essential concepts while actively excluding unnecessary concepts and terms.
- Providing value judgments and determining which terms are preferred, admitted, and deprecated.
- Confirming that definitions are accurate, contexts are authentic, and examples are representative.
- Guarding against plagiarism and intellectual property violations.
- Determining concept models and relationships, systematically organizing concept systems.
- Validating the correctness of metadata.

[1] Hanne Smaadahl. Contribution to SAP internal whitepaper. 2025.

- Confirming terminology is correct and culturally appropriate.
- Ensuring legitimate variations are not mistakenly flagged as anomalies by AI.
- Removing harmful biases and inequities, such as those based on race, gender, or socioeconomic status, to ensure fairness and inclusivity in the data and processes.
- Developing the data model and adapting it to accommodate new use cases.
- Analyzing data to refine and improve the termbase over time.

By contrast, other activities lend themselves to deeper AI involvement, ranging from collaborative forms of human-in-the-loop interaction to more AI-driven modes of operation. These include (see [2, 5, 10]):

- Extracting terms, abbreviations, and definitions from large corpora or databases.
- Identifying and proposing relationships, such as hierarchies and superordinate terms.
- Detecting patterns or logical groupings that might not be immediately obvious.
- Identifying parts of speech for terms and providing linguistic annotations.
- Suggesting synonyms or target language equivalents based on context and usage patterns.
- Validating terms across different documents or databases to ensure consistency.
- Monitoring and analyzing data from various sources to detect emerging trends, identify shifts in terminology usage, and forecast future terminology needs.

Overall, effective terminology management depends on integrating both human expertise and AI capabilities. While AI can automate and enhance many processes, human oversight is required to ensure the quality, cultural sensitivity, and reliability of terminological content. In later sections, this continuum is operationalized through the distinction between LLM-assisted and LLM-driven workflows.

2. Research trends in term extraction and definition generation

Terminology management workflows depend on two upstream operations: identifying candidate terms used in specialized discourse (term extraction) and documenting the conceptual content that makes a terminological entry reusable and verifiable (definition work[2]). Both operations are increasingly presented as suitable for automation; however, they face a structural mismatch between (a) what can be reliably inferred from corpus evidence and (b) what terminology work requires as a trustworthy, concept-oriented, quality-controlled representation that supports consistent specialized communication and downstream tasks such as translation, technical writing, compliance, standardization, and knowledge engineering.

Across both term extraction and definition generation, a recurring methodological issue is that "termhood" [11, 12] and "definitional adequacy" [13] are not purely linguistic properties. Terms are often described as lexical items that verbalize domain concepts, but operationalizing this in text is difficult: corpora contain terminological units, general-language words used technically, named entities, abbreviations, or underspecified expressions, to name just a few. This explains why Automated Term Extraction (ATE) remains challenging despite decades of work and why shared tasks have repeatedly highlighted definitional ambiguity, annotation variability, and cross-domain instability [14, 15].

A similar difficulty arises in definition generation. Much of the NLP literature focuses on producing dictionary-style glosses, i.e., short lexicographic definitions designed to capture the meaning of a lexical unit or a particular sense as evidenced in usage and tailored to dictionary

[2] In this regard, a clarification is needed: not all termbases contain definitions. Nevertheless, termbases that are useful in AI scenarios still contain conceptual content that make entries reusable and verifiable (i.e., various metadata such as descriptions, context, notes, examples, part of speech, domain, concept relations, etc.). In this paper, we focus specifically on definitions, as they represent the core of terminology work and, as such, they may be considered a "gold standard" for conceptual documentation.

users (e.g., [16, 17, 18]). These approaches are often framed as "definition modeling" or "definition generation" [19, 20]. This line of work is relevant, but it does not automatically satisfy the requirements of a terminological definition, which aims to define the concept in a domain and situate it within a concept system [21]. Consequently, the central question for terminology management is not whether systems can generate fluent definitions. Rather, it is whether generated (or extracted) definitions are conceptually constrained, domain-appropriate, non-circular, and auditable under professional quality criteria. This distinction matters directly for automation: many generative systems are optimized to use probabilistic methods to produce plausible explanatory text, whereas terminological practice requires controlled conceptual commitments (what the concept is, how it differs from related concepts, and how it relates to the domain model).

2.1. Term extraction

Term extraction (including ATE) is typically conceptualized as producing a ranked list of candidate terms from specialized corpora. Most approaches implement a pipeline, understood as a sequence of steps in text processing [22]. This pipeline generally includes: (i) candidate identification (often via linguistic patterns and POS tagging and chunking), (ii) scoring/ranking (statistical association, domain specificity, contextual features), and (iii) filtering/normalization (variants, nested terms, multiword expressions). Recent surveys emphasize that the field's central challenges are less about extracting strings than about deciding what counts as a term under varying domain boundaries, annotation guidelines, and downstream use cases. Classic ATE methods (e.g., C-value/NC-value) address issues such as multiword termhood and nested terms, combining linguistic constraints with frequency and contextual evidence [23, 24]. Contemporary work increasingly explores neural and transformer-based ranking, and, more recently, LLM-assisted extraction strategies [25]. Shared tasks and curated datasets (e.g., TermEval/ACTER) have contributed to reproducibility and cross-linguistic comparison, but they also expose persistent variability due to different conceptualizations of terms, domain differences, and annotation granularity [15].

2.2. Concept definition

Terminological definition work is concept-oriented: definitions are not primarily explanations of the meaning of a lexical unit (word), but controlled statements that identify a concept within a domain, specify its essential characteristics, and distinguish it from related concepts in a concept system. This approach is central in ISO terminology standards (e.g., ISO 704 [21]), which explicitly link objects, concepts, definitions, and designations and treat definition generation as part of concept formation and concept-system description.

For automation, the literature splits into two partially overlapping streams:

- Definition extraction / knowledge-rich contexts (KRCs): pattern-based and corpus-driven approaches that locate definitional contexts and extract definitional knowledge from specialized texts (e.g., "X is a Y that ..." [26]). This connects directly to terminology work because it leverages authoritative discourse and preserves traceability to sources.
- Definition generation / definition modeling: as commonly addressed in NLP research, generative approaches that produce dictionary-like glosses for words or senses, often evaluated against lexicographic resources. This is valuable as background, but concept-level adequacy in a domain is not guaranteed, especially under polysemy or homonymy and domain shift [27].

While these streams are often treated separately, they partially overlap in practice. Extracted knowledge-rich contexts can serve as grounded inputs for generative models, supporting traceable definition drafting, while generative approaches can in turn be used to normalize, synthesize, or reformulate extracted definitional material into controlled terminological definitions. This interaction is particularly relevant in LLM-assisted workflows, where extraction and generation are increasingly combined.

3. Some preliminary considerations for using LLMs in terminology management

The use of LLMs in terminology management requires prior consideration to ensure their application is effective, appropriate, and justified in terms of time and cost. This consideration involves several key aspects:

- **Task suitability**: automatable and repetitive tasks, such as detecting missing data categories, are generally well-suited to LLM support, as the one-time effort required to automate a recurring task can result in significant gains in efficiency and accuracy. Conversely, conceptual tasks, such as drafting definitions or conceptual notes, require careful assessment of the effort needed for human review and validation.
- **Data quality**: terminological data must be clean, consistent, and well structured [28]. While LLMs can support quality checks by identifying issues such as missing metadata, inconsistent, incomplete, or unstructured data can significantly reduce the usefulness of LLMs and substantially increase human workload during review and validation.
- **Data volume**: LLMs can process large datasets in shorter timeframes compared with fully human-driven approaches [10, 29]. However, for small datasets—particularly in one-time tasks—the use of LLMs may not be justified. In such cases, predominantly human-driven approaches are preferable, as the effort required to export, prepare, and validate both the input data and the model's output can exceed that of manual processing.
- **The impact of errors**: research shows that LLMs are not yet sufficiently reliable for tasks in high-stakes domains such as law or medicine [30, 31]. In terminology management, errors may propagate into domain-specific documentation or translations, potentially causing serious consequences (e.g., bodily harm, legal actions, generation of discriminatory content) [30, 32]. Therefore, strong human oversight is required. By contrast, GenAI is better suited to low-impact tasks—such as preliminary analyses or exploratory checks—where errors are unlikely to have serious consequences and can be addressed through limited human review.
- **Model configuration and usage mode**: Data accuracy depends on two main factors: (i) the type of model used and its degree of adaptation (e.g., base, instruction-tuned, or fine-tuned), and (ii) inference-time control, including prompting strategies, constraints, and retrieval augmentation. The nature of the task determines the appropriate prompting approach. Zero-shot prompting is appropriate when the model can generalize from pretraining/instruction tuning and the task can be specified in the prompt [33], as the model relies entirely on its general training and the explicit instructions provided in the prompt, without the need for illustrative examples (e.g., "Convert this term list into a table, organizing it by the following columns: Entry number, Project, Italian, Definition, German, Comments. Use only terms provided in this list: Don't add or omit any term."). One-shot and few-shot prompting require one or several examples, respectively, and may be used in terminology management, for example, to support the generation of intensional definitions in the sense of "immediate superordinate concept + delimiting characteristic(s)" [33]. Fine-tuned models may deliver stronger performance and more reliable outputs in terms of domain specificity and terminological correctness [32, 34]. For example, they can be used to assess terminology compliance in high-stakes domains [32]. However, they require task-relevant, high-quality data (often substantial, depending on the adaptation method and target quality) [34, 35] and careful dataset curation, which can be particularly challenging for low-resource languages [32, 35]. As a result, fine-tuning may not be suitable for all terminology tasks.

- **Provenance**: AI-generated content should be explicitly identifiable [36]. AI-generated definitions should be clearly marked as such in a termbase, for example, by indicating 'AI-Generated' as the source. This would support transparency and enable a targeted review process.

4. Description of the methodology for identifying evaluation criteria

Within a terminological framework, termbase entries are treated as concept-oriented, quality-controlled, and auditable artifacts. ISO 704 [21] frames terminology work in terms of the relationships among objects, concepts, definitions, and designations, providing guidance for definition writing. Evaluation, therefore, goes beyond the "linguistic plausibility" of extracted strings or generated text to consider their fitness for integration into a terminology resource under defined quality constraints.

Against this backdrop, the evaluation criteria proposed in this paper were derived through a procedure that connects (i) terminology work requirements as formalized in standards for concept-oriented terminological description and termbase quality, with (ii) operational evaluation practices emerging from recent shared tasks and benchmarks targeting term extraction and, more recently, terminology definition generation. On the evaluation-practice side, we performed a focused scoping review of recent and established evaluation initiatives that operationalize terminology-related tasks in reproducible settings. As regards term extraction, we included (a) long-standing ATE evaluation campaigns that shaped common metrics and datasets (e.g., TermEval 2020 using ACTER) and (b) newer domain- and language-specific evaluation efforts (e.g., ATE-IT at EVALITA 2026) [14, 37]. For definition generation in terminology work, we included DETECH (co-located with MDTT 2026) as a task explicitly coupling term extraction and medical concept definition generation, and we used Graphine [38] as a terminology-specific benchmark and dataset for large-scale biomedical terminology definition generation. For tasks that sit at the boundary between terminology and explanation for accessibility, we considered SimpleText @ CLEF 2024 Task 2 ("difficult concept identification and explanation"), because it makes explicit the requirement to decide which domain notions require explanation and to produce usable explanatory outputs [39, 40, 41].

4.1. Criteria for evaluating LLM-assisted term extraction

In LLM-assisted ATE, the LLM serves as the primary engine generating ranked outputs. Evaluation must distinguish between (i) whether the system identifies plausible candidate strings and (ii) whether those candidates are suitable for use in terminology work. Shared tasks such as TermEval 2020 (https://termeval.ugent.be/) operationalize the first aspect through gold annotations and conventional performance measures; ATE-IT extends this logic in a language- and scenario-specific setting (Italian, institutional waste-management discourse), which is particularly relevant for terminology management in public-sector contexts.

Accordingly, we propose an evaluation organized around four dimensions:

- **Extraction correctness** addresses the quality of the candidate list: boundary accuracy for multiword terms (for example, "automatic term extraction"), correct handling of nested terms (for example, "automatic term extraction" vs "term extraction"), and standard evaluation performance metrics (precision/recall/F1, with ranking-aware metrics where outputs are prioritized lists). Term extraction shared tasks repeatedly show that boundary decisions and nested structures are non-trivial, making explicit evaluation of span correctness indispensable.
- **Termhood and domain specificity** capture the terminological requirement that extracted candidates should correspond to domain-relevant designations rather than merely salient words, named entities, or generic vocabulary. Here, the evaluation focus shifts from "is this a well-formed phrase?" to "is this a stable designation for a domain concept or knowledge

unit?”. This criterion is central when outputs are intended for termbase curation rather than for exploratory keywording [12]. This requirement becomes especially critical in LLM-assisted settings, where outputs are intended for direct integration into a termbase rather than used only as drafting aids.

- **LLM-specific operational reliability** becomes a primary evaluation criterion because the dominant failure modes differ from traditional ATE. In particular, when LLMs are used in instruction-following mode, evaluation must check closed-world compliance (“do not add terms not present in the source”), sensitivity to prompt phrasing, and reproducibility under controlled experimental settings. These properties are rarely foregrounded in older ATE evaluations, but become decisive when LLMs are integrated into production terminology pipelines.
- **Auditability for human review** requires that extracted candidates be accompanied by evidence that supports curator decisions. From a terminological perspective, the most valuable “explanation” is often not a model rationale, but a trace to corpus contexts (attestations) that justify inclusion, preferred form selection, and variant consolidation. This emphasis on evidence aligns with recent tasks such as SimpleText Task 2, where identification is coupled with the need to provide explanations of term difficulty and clarification strategies. Such tasks highlight that making the basis of a decision explicit—whether for end users or expert curators—is central to the usability and evaluability of system outputs.

4.2. Criteria for evaluating LLM-assisted definition generation

In LLM-assisted definition generation, the model may support drafting or reformulation, but conceptual validation remains a human responsibility. As mentioned in Section 2.2, in terminology management, definitions play a central role since they allow us to recognize a concept and differentiate it from other concepts, explaining the conceptual content that a given term conveys [21, 42].

Definitions generated by or with the help of LLM applications may appear linguistically plausible yet may fail to meet terminological quality requirements. To avoid the uncritical acceptance of AI-generated content, criteria for evaluating definitions must be defined when LLMs are used. Building on previous research on quality control ([12, 43, 44, 45, 46, 47]) and considering that definitions should comply with relevant literature and ISO standards, we can distinguish three different types of criteria: content-related, linguistic, and formal.

These criteria are illustrated with concrete example tasks to clarify the required interventions. Their ordering reflects the human–AI continuum in terms of increasing automatability: content-related criteria require strong human oversight because they demand domain-specific knowledge, critical domain judgment, and explicit human accountability for the final definition; linguistic criteria allow for intermediate levels of human oversight or human–AI collaboration, and formal criteria can be assessed with limited human oversight and a higher degree of automation.

4.2.1. Content-related criteria

Content-related criteria are the most critical in evaluating definitions, as they focus on what is defined rather than how it is phrased (linguistic) or how it is stored (formal). They assess conceptual accuracy and domain validity, requiring strong human oversight along the human–AI continuum. Key aspects include:

- **Conceptual accuracy**: a definition must correctly and appropriately describe the concept with respect to its domain or, in the case of legal terminology, the relevant legal system. This is due to the system-boundness of legal language [48, 49] and applies even in pluricentric languages (e.g., English, German). For example, the British legal concept of “parental responsibility”, defined as “all the rights, duties, powers, responsibilities and authority which by law a parent of a child has in relation to the child and his [or her] property" [50], corresponds to distinct designations in German depending on the legal system: *Obsorge* in Austria, *elterliche Sorge* in Germany and Switzerland, and *elterliche Verantwortung* in South Tyrol [51, 52].

- **Identification of *genus proximum* and *differentiae specificae***: according to ISO 704 [21], intensional definitions should begin with the immediate superordinate concept (*genus proximum*) followed by the delimiting characteristics (*differentiae specificae*). This definitional format is inherently concise, as it requires selecting only those characteristics that are essential to situate the concept within the concept system (e.g., a table is a type of furniture). In high-stakes domains such as law, this requirement for conciseness makes the careful identification of delimiting characteristics particularly critical. However, in such domains, conciseness may need to be subordinated to the completeness of legally binding characteristics of the definition. This process, if conciseness is applied too rigidly, risks becoming arbitrary and may result in incomplete or misleading definitions. Consider for example the definition of "personal data" according to the EU regulation 2016/679 (GDPR) [53]: "any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person". A concise rewording of this definition could be "information that identifies a person or allows a person to be identified". However, as can be seen, such a reformulation would exclude several characteristics that the legislator considers essential. Notably, it does not explicitly address indirect identification and therefore risks being interpreted as limited to direct identification only. Moreover, the legal definition refers to "any information relating to an identified or identifiable natural person". From this perspective, the formulation "information that identifies a person" is restrictive and does not fully reflect the scope of the legal concept.
- **Domain specificity**: concepts must be defined according to their specific domain. Homonymy may occur even within the same domain, notably across different subdomains, making precise domain identification necessary. For instance, the Italian term *atto* may refer to a document or to an action, depending on whether it is used in civil and administrative law or in criminal law [52]. Correct identification of the relevant subdomain and the corresponding definition is also crucial for selecting appropriate equivalents in other languages, as *atto* corresponds to different terms in German or English depending on the legal subdomain.
- **Context sensitivity**: the context must be considered in the drafting of definitions. In the legal domain, for example, the degree of specificity of a definition may vary depending on the category of the legal source (legislation, legal doctrine, case law). Consequently, definitions should be evaluated in relation to their intended context to ensure an appropriate level of granularity.
- **Non-circularity**: a definition must avoid explaining a concept by directly or indirectly referring to the same concept. Circular definitions such as "a teacher is a person who teaches" fail to provide explanatory value and undermine terminological clarity.
- **Currency**: this concerns both the content of the definition and the designations used to define the concept. In domains such as law, concepts may change over time. Designations may also change—even when the underlying concept remains stable. For example, in 2004, an Italian legislative decree (Legislative Decree No. 59/2004, [54]) changed the designation from *scuola elementare* to *scuola primaria*, while the concept itself remained the same ("elementary school"). Definitions and designations must therefore be verified to reflect current legislation, standards, usage, or technology.

- **Sources**: definitions should be documented by reliable, authentic, and traceable sources. However, as Heinisch (2025) [10] points out, LLMs may draw on information that does not originate from the relevant domain, may be outdated, artificially generated, or difficult to trace, and may therefore fail to fit the intended purpose or context.

4.2.2. Linguistic criteria

These criteria concern the linguistic aspects of definitions (cf. [43]), and are focused on surface formulation (e.g., avoiding homonyms, colloquialisms, stigmatizing adjectives) rather than the concept level. They address, for example, spelling checks, the identification of typos and grammatical errors, and the verification of linguistic quality (including the use of unambiguous terms, a neutral and concise tone), and the absence of unnecessary examples, extraneous information, or biased elements. Linguistic criteria also cover style and consistency so that definitions remain easy to compare, validate, translate, and reuse across the termbase and other applications. This includes using a consistent definition pattern across entries (e.g., “X is a type of Y that…”), avoiding hedging—such as the use of modal markers like *generally* or *typically*—when a strict definitional statement is required, and keeping definitions concise by removing verbosity and redundancy. In addition, they encompass compliance with the definition format conventions at the level of linguistic realization: an intensional definition should begin with the immediate generic concept and avoid introductory formulas such as “According to the Criminal Code, a crime is…” [21].

LLMs can aid linguistic checks by detecting errors, ambiguities, and suggesting revisions. Human oversight ranges from minimal for automatable tasks to strong for ensuring accuracy and terminological compliance.

4.2.3. Formal criteria

These criteria concern ensuring that all formal rules have been respected [43], such as the correct placement of definitions within the termbase, namely assignment to the appropriate data category and level (e.g., concept level or language level, depending on the termbase structure). They also include verifying that the definition is present and the sources are recorded correctly and, where applicable, checking that hyperlinks are still active. In addition, formal criteria cover the detection of technical issues such as embedded hard-line breaks in definition fields, which may result from copy-and-paste operations from web sources [47].

The evaluation of formal criteria can be largely supported by LLMs, as these applications can help verify the presence, placement, and basic formatting of definitions in accordance with predefined rules (e.g., internal guidelines). Such tasks are typically repetitive and routine, making them particularly suitable for LLM-assisted processing. Human oversight is therefore required but remains limited. Terminologists, however, are responsible for defining the formal requirements, such as where definitions should be placed, during the termbase design and data modeling phase.

5. Discussion and future work

This paper examined how term extraction and concept definition, two core activities in terminology management, should be evaluated when supported by GenAI, particularly LLMs. The underlying principle is that automation in terminology work cannot be evaluated solely by speed, fluency, or generic performance, but must also be judged on its ability to produce reliable, concept-oriented, quality-controlled resources. The proposed criteria draw on terminology standards, practices, and recent evaluations of extraction, explanation, and definition generation in specialized domains.

One of our main conclusions is that term extraction and definition generation occupy structurally distinct positions along the human–AI continuum. This reflects not only their cognitive demands, but also the nature of their failure modes and the associated costs of correction.

Term extraction—particularly when framed as candidate identification and ranking from a bounded corpus—exhibits a relatively high degree of automatable regularity. Its outputs can be evaluated using established techniques such as boundary accuracy, precision/recall, ranking quality, and evidence-based review support. In this setting, LLMs can be useful as term extractors when integrated under strict constraints, including closed-world compliance and reproducibility via corpus attestations. When these conditions are met, human oversight can reasonably be limited to validation and curation decisions, consistent with a collaborative workflow or, in some low-risk scenarios, a more AI-driven governance model. However, we also highlight that LLM-mediated extraction introduces issues like prompt sensitivity, instability across runs, and the risk of invented candidates, that are less prominent in traditional ATE pipelines and therefore must be treated as specific evaluation dimensions rather than as secondary implementation details.

By contrast, definition generation is not simply more difficult, but different in kind, because it aims at producing a terminological definition: a controlled statement that defines a concept in a domain, ideally situating it within a concept system through superordinate concepts and delimiting characteristics. This raises requirements that cannot be satisfied by fluency alone. A definition that reads well may still be inadequate if it is ungrounded, conceptually misaligned, circular, or inconsistent with neighboring concepts. Therefore, we argue that conceptual accuracy, domain correctness, grounding and traceability, as well as terminological usability, should be fundamental criteria in the evaluation. In practice, this positions definition generation closer to the human-driven end of the continuum that is, to settings in which humans retain responsibility for key decisions and oversight: even when LLMs are helpful for drafting, paraphrasing, or proposing intensional formulations, the role of the human terminologist is not reduced to surface validation but remains central to conceptual decisions, quality assurance, and accountability. If term extraction can, under controlled conditions, approximate a repetitive, structure-preserving automation problem, definition generation remains a concept-sensitive knowledge task, in which the system's persuasive natural language output can mask errors that are costly to detect and correct.

A second conclusion concerns how "evaluation" must be understood in LLM-assisted and LLM-driven terminology workflows. Evaluation criteria should be explicitly linked to usage mode and governance: the same system behavior may be acceptable in a suggestion workflow with strong human oversight, but unacceptable in autonomous termbase population. Consequently, criteria such as traceability, reproducibility, and closed-world adherence are not optional refinements; they determine whether an output can enter a termbase as a trusted and verifiable resource or must remain a simple drafting aid. This governance-aware perspective also provides a practical response to the current LLM adoption wave: rather than framing the question as whether LLMs should or should not be used in terminology management, we suggest that LLMs should be components whose admissible role varies by task, risk level, and the degree to which outputs can be grounded.

The proposed evaluation criteria are intended to be actionable. Although they have not yet been empirically validated, they are conceived as a theoretically grounded framework to be operationalized and tested in future pilot studies and user-centered evaluations, particularly to assess usability and integration into expert workflows. They translate the conceptual commitments of terminology work into operational checks that can be applied in evaluation protocols or embedded in tools as quality controls. At the same time, the field lacks sufficiently mature, terminology-native evaluation resources for concept definition generation, and the gap between existing NLP definition-generation benchmarks and terminological definition requirements remains substantial. This motivates a final, forward-looking implication: future work should prioritize evaluation designs that reflect terminology management as a lifecycle practice, not as isolated NLP tasks, and that operationalize concept orientation through structured constraints, evidence linking, and curator-centered measures such as correction effort and decision confidence. In this sense, the human–AI continuum is not just a framing but an evaluative instrument: it forces the community to articulate what can be safely delegated to automation, under what controls, and with what accountability, while preserving the integrity of terminology resources as dependable infrastructures for specialized communication.

Acknowledgments

We want to express our sincere thanks to Ariel Kahana (SAP) for proofreading the paper and for providing valuable linguistic input. We also thank the International Organization for Standardization (ISO) for providing a collaborative framework that fosters the development of standards in terminology and language resources. Notably, we wish to thank the colleagues of Working Group 6 "Terminology management and artificial intelligence" of ISO/TC 37/SC 3 "Management of terminology resources" for the valuable discussions, insights, and shared expertise, which have significantly contributed to and enhanced this work. Additionally, we gratefully acknowledge the support of the authors' respective institutions, companies and national standardization bodies for the professional environment and ongoing collaboration they provide.

Declaration on Generative AI

In preparing this contribution, generative AI tools (Copilot, ChatGPT-5.2) and an AI-powered writing assistant (Grammarly) were used to improve grammar, language, and style. The content was then reviewed and edited with assistance from a native English speaker. The authors take full responsibility for the content of this publication.

References

[1] M. Iantosca, The Semantic Content Maturity Model, 2022. Accessed 12 January, 2026. URL: https://thinkingdocumentation.com/blog/f/the-semantic-content-maturity-model.

[2] A. San Martín, Toward Human-Centered AI-Assisted Terminology Work, arXiv, arXiv:2512.18859, 2025. doi:10.48550/arXiv.2512.18859.

[3] I. Shumailov, Z. Shumaylov, Y. Zhao, N. Papernot, R. Anderson, Y. Gal, AI models collapse when trained on recursively generated data, Nature 631 (2024) 755–759. doi:10.1038/s41586-024-07566-y.

[4] M. Iantosca, A few things newcomers often miss about Knowledge Graphs (KGs), comment on LinkedIn post, 2025. Accessed 12 January, 2026. URL: https://www.linkedin.com/feed/update/urn:li:ugcPost:7331063813904965638.

[5] ISO/AWI TR 25896, Management of Terminology Resources — Artificial Intelligence for terminology management, International Organization for Standardization, Geneva, Switzerland, 2025. To appear.

[6] K. Warburton, Terminology in the Age of AI, Multilingual (March 2025). Accessed 12 January, 2026. URL: https://multilingual.com/magazine/march-2025/terminology-in-the-age-of-ai/.

[7] M. Khemakhem, C. Valentini, N. Ralli, S. Barros, G. Löckinger, F. Vezzani, A. Salgado, Z. Zhang, S. Mahr, S. Carvalho, K. Fleischmann, R. Costa, Terminology management meets AI: The ISO/TC 37/SC 3/WG 6 initiative, in: Proceedings of the 5th Conference on Language, Data and Knowledge: TermTrends 2025. Unior Press, Naples, Italy, 2025, pp. 16–24. URL: https://aclanthology.org/2025.termtrends-1.2/.

[8] P. Drewer, K.-D. Schmitz, Terminologiemanagement: Grundlagen – Methoden – Werkzeuge, Springer Vieweg, Wiesbaden, Germany, 2017.

[9] R. Arntz, H. Picht, K.-D. Schmitz, Einführung in die Terminologiearbeit, 8th. ed., Georg Olms Verlag, Hildesheim, Germany, 2021.

[10] B. Heinisch, Next-gen terminology: Transforming terminology work with large language models, Across Languages and Cultures 26 (2025) 64–80. doi:10.1556/084.2025.01061.

[11] K. Warburton, The corporate terminologist, in: M.-C. L'Homme, K. Kageura (Eds.), Terminology and Lexicography Research and Practice, volume 21, John Benjamins Amsterdam/Philadelphia, 2021. doi:10.1075/tlrp.21.
[12] ISO 26162-3, Management of terminology resources — Terminology databases — Part 3: Content, International Organization for Standardization, Geneva, Switzerland, 2023. URL: https://www.iso.org/standard/80464.html.
[13] M. T. Cabré Castellví, Terminology: Theory, methods and applications, John Benjamins Amsterdam/Philadelphia, 1999. doi:10.1075/tlrp.1.
[14] A. Rigouts Terryn, V. Hoste, P. Drouin, E. Lefever, TermEval 2020: Shared task on automatic term extraction using the annotated corpora for term extraction research (ACTER) dataset, in: Proceedings of the 6th International Workshop on Computational Terminology, European Language Resources Association, Paris, France, 2020, pp. 85–94. URL: https://aclanthology.org/2020.computerm-1.12/.
[15] G. M. Di Nunzio, S. Marchesin, G. Silvello, A systematic review of automatic term extraction: What happened in 2022?, Digital Scholarship in the Humanities 38 (2023) i41–i47. doi:10.1093/llc/fqad030.
[16] B. Svensén, A handbook of lexicography: The theory and practice of dictionary-making, Cambridge University Press, Cambridge, UK, 2009.
[17] I. Mel'čuk, A. Polguère, Theory and practice of lexicographic definition, Journal of Cognitive Science 19 (2018) 417-470. doi:10.17791/JCS.2018.19.4.417.
[18] G.-M. De Schryver, Generative AI and lexicography: The current state of the art using ChatGPT, International Journal of Lexicography 36 (2023) 355-387. doi:10.1093/ijl/ecad021.
[19] T. Mickus, D. Paperno, M. Constant, Mark my word: A sequence-to-sequence approach to definition modeling, in: Proceedings of the First NLPL Workshop on Deep Learning for Natural Language Processing, Linköping University Electronic Press, Linköping, Sweden, 2019, pp. 1–11. URL: https://aclanthology.org/W19-6201/.
[20] V. Segonne, T. Mickus, Definition modeling: To model definitions. Generating definitions with little to no semantics, in: Proceedings of the 15th International Conference on Computational Semantics, Association for Computational Linguistics, Nancy, France, 2023, pp. 258–266. URL: https://aclanthology.org/2023.iwcs-1.27/.
[21] ISO 704, Terminology Work — Principles and methods, International Organization for Standardization, Geneva, Switzerland, 2022. URL: https://www.iso.org/standard/79077.html.
[22] B. B. Elov, Sh. M. Khamroeva, Z. Y. Xusainova, The pipeline processing of NLP, E3S Web of Conferences 413 (2023). doi:10.1051/e3sconf/202341303011.
[23] K. Frantzi, S. Ananiadou, H. Mima, Automatic recognition of multi-word terms: The C-value/NC-value method, International Journal on Digital Libraries, 3 (2000) 115–130. doi:10.1007/s007999900023.
[24] H. Smaadahl, F. Fouvry, Terminology extraction: Legacy terminology in content authoring and translation, in: Proceedings of the Corpus Linguistics Conference (CL2007), University of Birmingham, Birmingham, England, 2007. URL: https://ucrel.lancs.ac.uk/publications/CL2007/paper/115_Paper.pdf.
[25] Y. Chun, M. Kim, D. Kim, C. Park, H. Lim, Enhancing automatic term extraction with large language models via syntactic retrieval, in: W. Che, J. Nabende, E. Shutova, M. T. Pilehvar (Eds.), Findings of the Association for Computational Linguistics: ACL 2025, Association for Computational Linguistics, Vienna, Austria, 2025, pp. 9916–9926. doi:10.18653/v1/2025.findings-acl.516.
[26] I. Meyer, Extracting knowledge-rich contexts for terminography: A conceptual and methodological framework, in: D. Bourigault, C. Jacquemin, M.-C. L'Homme (Eds.), Recent Advances in Computational Terminology, John Benjamins Publishing Company, Amsterdam/Philadelphia, 2001, pp. 279-302. doi:10.1075/nlp.2.15mey.
[27] A. Gadetsky, I. Yakubovskiy, D. Vetrov, Conditional generators of words definitions, in: I. Gurevych, Y. Miyao (Eds.), Proceedings of the 56th Annual Meeting of the Association for

Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 266–271. doi:10.18653/v1/P18-2043.
[28] F. Holdermann, Terminologie und Terminologiemanagement im Zeitalter der Digitalisierung, in: P. Drewer, F. Mayer, D. Pulitano (Eds.), Terminologie in der KI – KI in der Terminologie. Akten des Symposions, Worms, 27–29 März 2025, Deutscher Terminologie-Tag e.V., Munich, Germany, 2025, pp. 23–38.
[29] J. Giguere, A. Iankovskaia, Leveraging large language models to extract terminology, in: Proceedings of the First Workshop on NLP Tools and Resources for Translation and Interpreting Applications, INCOMA Ltd., Shoumen, Bulgaria, 2023, pp. 57–60. URL: https://aclanthology.org/2023.nlp4tia-1.9/.
[30] C.-C. Hung, W. Ben Rim, L. Frost, L. Bruckner, C. Lawrence, Walking a tightrope—evaluating large language models in high-risk domains, in: Proceedings of the 1st GenBench Workshop on (Benchmarking) Generalisation in NLP, Association for Computational Linguistics, Singapore, 2023, pp. 99–111. URL: https://aclanthology.org/2023.genbench-1.8/.
[31] M. Sallam, ChatGPT utility in healthcare education, research, and practice: Systematic review on the promising perspectives and valid concerns, Healthcare 11 (2023). doi:10.3390/healthcare11060887.
[32] P. Di Natale, E. W. Stemle, E. Chiocchetti, M. Alber, N. Ralli, I. Stanizzi, E. Benini, The LegISTyr test set: Investigating off-the-shelf instruction-tuned LLMs for terminology-constrained translation in a low-resource language variety, in: Proceedings of the Workshop TermTrends25: Bridging the Gap between Terminological Resources and Large Language Models, Unior Press, Naples, Italy, 2025, pp. 1–15. URL: https://aclanthology.org/2025.termtrends-1.1/.
[33] R. Herwartz, KI in der Lehre. Prompting-Methoden und Glossar-Einbindung in NMÜ und LLMs, in: P. Drewer, F. Mayer, D. Pulitano (Eds.), Terminologie in der KI – KI in der Terminologie. Akten des Symposions, Worms, 27–29 März 2025, Deutscher Terminologie-Tag e.V., Munich, Germany, 2025, pp. 135–50.
[34] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, et al., Language models are few-shot learners, in: Advances in Neural Information Processing Systems 33, Curran Associates, Red Hook, NY, 2020. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.
[35] A. Edwards, J. Camacho-Collados, Language models for text classification: Is in-context learning enough?, in: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), ELRA and ICCL, Torino, Italy, 2024, pp. 10058–10072. URL: https://aclanthology.org/2024.lrec-main.879/.
[36] M. Mora-Cantallops, S. Sánchez-Alonso, E. García-Barriocanal, M. A. Sicilia, Traceability for trustworthy AI: A review of models and tools, Big Data and Cognitive Computing 5 (2021). doi:10.3390/bdcc5020020.
[37] ATE-IT, ATE-IT (Automatic Term Extraction) project website, n.d. Accessed January 9, 2026. URL: https://nicolacirillo.github.io/ate-it/.
[38] Z. Liu, S. Wang, Y. Gu, R. Zhang, M. Zhang, S. Wang, Graphine: A dataset for graph-aware terminology definition generation, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 3453–3463. doi:10.48550/arXiv.2109.04018.
[39] G. M. Di Nunzio, F. Vezzani, V. Bonato, H. Azarbonyad, J. Kamps, L. Ermakova, Overview of the CLEF 2024 SimpleText task 2: Identify and explain difficult concepts, in: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024), Volume 3740 of CEUR Workshop Proceedings, 2024. URL: https://ceur-ws.org/Vol-3740/#paper-306.
[40] L. Ermakova, E. SanJuan, S. Huet, H. Azarbonyad, G. M. Di Nunzio, F. Vezzani, J. D'Souza, J. Kamps, Overview of the CLEF 2024 SimpleText track, in: L. Goeuriot, et al. (Eds.),

Experimental IR Meets Multilinguality, Multimodality, and Interaction, Springer Nature Switzerland, Cham, 2024. doi:10.1007/978-3-031-71908-0_13.
[41] SimpleText Project, CLEF SimpleText track, n.d. Accessed January 9, 2026. URL: https://simpletext-project.com/2024/en/.
[42] A. San Martín, What generative artificial intelligence means for terminological definitions, in: Proceedings of the 3rd International Conference on Multilingual digital terminology today (MDTT 2024), Volume 3703 of CEUR Workshop Proceedings, 2024. URL: https://ceur-ws.org/Vol-3703/paper1.pdf.
[43] E. Chiocchetti, B. Heinisch-Obermoser, G. Löckinger, V. Lušicky, N. Ralli, I. Stanizzi, T. Wissik, Guidelines for Collaborative Legal/Administrative Terminology Work, EURAC Research, Bolzano, Italy, 2013. URL: https://cordis.europa.eu/docs/projects/cnect/7/270917/080/deliverables/001-D33Guidelinesforcollaborativelegaladministrativeterminologywork.pdf.
[44] C. Valentini, Quality control in terminology management, in: H. E. Thomsen, A. Pareja-Lora, B. Nistrup Madsen (Eds.), Term bases and linguistic linked open data: Proceedings of TKE 2016 the 12th International conference on Terminology and Knowledge Engineering, Copenhagen Business School, 2016, pp. 34–43. URL: http://openarchive.cbs.dk/handle/10398/9323.
[45] E. Chiocchetti, V. Lušicky, T. Wissik, Multilingual legal terminology databases: Workflows and roles, in: Ł. Biel, H. J. Kockaert (Eds.), Handbook of Terminology. Volume 3: Legal Terminology, John Benjamins Publishing Company, Amsterdam/Philadelphia, 2023, pp. 458–484. doi:10.1075/hot.3.mul1.
[46] B. Heinisch, Terminological usability – adapting terminological databases to different user groups according to usability principles: the case of UniVieTerm, Terminology Science & Research 26 (2023) 24–44. URL: https://journal-eaft-aet.net/index.php/tsr/issue/archive.
[47] N. Ralli, D. M. Tamás, How to create and manage terminology resources: A practical guide from two termbases, Journal of Digital Terminology and Lexicography, 1 (2025) 61–82. doi:10.25430/pupj.jdtl.1751631260.
[48] G. R. de Groot, Das Übersetzen juristischer Terminologie, in: G.-R. de Groot, R. Schulze (Eds.), Recht und Übersetzen, Nomos, Baden-Baden, Germany, 1999.
[49] A. Gambaro, R. Sacco, Sistemi giuridici comparati, 4th. ed., UTET, Milano, Italy, 2024.
[50] Children Act 1989, Section 3, 1989. Accessed 12 January, 2026. URL: https://www.legislation.gov.uk/ukpga/1989/41/section/3.
[51] R. Muhr, M. Peinhopf, Wörterbuch rechtsterminologischer Unterschiede Österreich–Deutschland, volume 16 of Österreichisches Deutsch – Sprache der Gegenwart, Peter Lang, Frankfurt am Main, Berlin, Bern, Bruxelles, New York, Oxford, Wien, 2015. doi:10.3726/978-3-653-06050-8.
[52] bistro, Informationssystem für Rechtsterminologie, n.d. Accessed 12 January, 2026. URL: https://bistro.eurac.edu/.
[53] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 (General Data Protection Regulation), 2016. Accessed 12 January, 2026. URL: https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng.
[54] Decreto Legislativo 19 febbraio 2004, n. 59: Definizione delle norme generali relative alla scuola dell'infanzia e al primo ciclo dell'istruzione, 2004. Accessed 12 January, 2026. URL: https://www.istruzione.it/esame_di_stato/Primo_Ciclo/normativa/allegati/Decreto%20legislativo%2059%20del%202004,%20Definizione%20delle%20norme%20generali%20relative%20alla%20scuola%20dell'infanzia%20e%20al%20primo%20ciclo%20dell'istruzione.pdf.