Teaching Data Management to Translation Students: From Documentation Practices to Data Literacy

Pilar Sánchez-Gijón
Universitat Autònoma de Barcelona
Spain
pilar.sanchez.gijon@uab.cat

Abstract

This paper proposes an approach to teaching data management in translation programmes, framing it as an extension of established documentation practices. It argues that data governance, sourcing, and processing are core competences for translators working with NMT and LLMs, and outlines a training framework that supports responsible data reuse, quality optimisation, and professional agency.

1 Introduction

The rapid integration of generative AI, neural machine translation (NMT), and large language models (LLMs) into translation workflows has repositioned linguistic data from a secondary by-product of translation projects to a strategic asset that directly affects translation quality, system performance, and ethical compliance. In NMT- and LLM-based workflows, system behaviour and output quality are critically conditioned by the data used for configuration, adaptation, and reuse, making data management a prerequisite for informed and responsible use of machine translation technologies in both professional practice and translator training. As a consequence, data management has emerged as a core competence for future translators, yet it remains underrepresented and insufficiently systematised in translator training (Bowker, 2026; DePalma & Lommel, 2025; Mellinger & Hanson, 2025; Sánchez-Gijón & Torres Simón, forthcoming).

This shift is explicitly reflected in recent competence frameworks. Data management is identified as a transversal competence in the LT-LiDER Language Technology Map, which outlines the skills required for participation in language-centric AI workflows (Secară et al., 2024). This recognition marks a move away from tool-centred training towards competence-based profiles, in which the ability to understand how data are created, described, governed, and reused becomes central to professional agency. While currently in the pilot phase, the proposed framework is robustly grounded in FAIR data principles (Findable, Accessible, Interoperable, and Reusable) and established documentation standards, ensuring its alignment with both academic rigor and industry requirements for data readiness.

2 Industry drivers: data readiness and data spaces

The growing relevance of data management is further reinforced by industry developments. According to the GALA Business Barometer Report (2025), only around 30% of business-generated language data are currently suitable for reuse in AI-driven workflows, revealing a substantial gap between data availability and data readiness (GALA, 2025). In this context, data readiness is defined as the state in which linguistic resources are structurally, qualitatively, and legally prepared for immediate integration into AI-driven workflows. This concept is operationalised through three primary dimensions: technical quality (including cleaning, alignment, and proper formatting), linguistic adequacy (ensuring domain relevance, terminological consistency, and genre conformity), and legal-ethical compliance (encompassing clear licensing, traceable provenance, and data privacy).

At the same time, large-scale initiatives such as the European Language Data Space (https://language-data-space.ec.europa.eu/index_en) foreground data governance, interoperability, licensing, and traceability as foundational elements of the emerging language technology ecosystem. These developments signal a shift towards regulated and value-driven data reuse, in which language professionals are expected to interact critically with datasets rather than merely consume technology outputs (Secară et al., 2024).

3 Teaching data management in translation programmes: a proposal

This paper argues that data management should be explicitly taught in translation programmes, not as a form of technical specialisation, but as a natural extension of documentation practices already familiar to translation students. Translators have long worked with parallel texts, corpora, translation memories, and termbases to ensure adequacy, genre conformity, and functional fitness (Bowker, 2026). From a Translation Studies perspective, data management can therefore be framed as an expansion of documentary and corpus competence adapted to AI-mediated reuse.

In line with Bowker's proposal to teach data through a science-communication approach, abstract notions such as data governance, sourcing, and processing are introduced using familiar concepts, analogies, and visualisation techniques. This approach helps non-technical audiences engage with data-centric concepts while reinforcing the translator's role as an expert decision-maker rather than a passive user of AI (i.e. NMT or LLM) systems.

3.1 Linking data management to Translation Studies theory

The proposed framework explicitly connects data management to established Translation Studies concepts, particularly functionalist approaches and corpus-based translation pedagogy. From this perspective, decisions about data selection, reuse, and annotation are closely aligned with core notions such as *skopos*, communicative purpose, genre conventions, register, and audience expectations (Nord, 2013). Linguistic data are thus understood not as neutral or interchangeable inputs, but as situated resources whose value depends on their adequacy for a specific translation task and context (Sánchez-Gijón & Torres Simón, forthcoming).

This view builds on traditional documentary practices in translator training, where parallel texts, corpora, translation memories, and terminological resources are used to observe how specialised discourse communities communicate and to model target-language production. In AI-mediated translation workflows, these same resources increasingly function as datasets used to customise and optimise NMT systems and LLMs. Framing data management within Translation Studies theory therefore allows students to recognise continuity between long-standing documentation practices and contemporary data-driven technologies, rather than perceiving AI as a conceptual rupture.

In line with Bowker's work (2026) on data literacy for translators, this framework also draws on a science-communication approach to make data-centric concepts accessible to non-technical audiences. By linking abstract notions such as governance, sourcing, and processing to familiar translational decisions, students are encouraged to reflect critically on how theoretical principles inform practical data management choices. This integration reinforces the translator's role as an informed decision-maker who applies theoretical knowledge to the management and reuse of linguistic data in AI-supported translation environments (Bowker, ibid.).

3.2 Core aspects to be taught

This proposal conceptualises data management in translation education as a structured set of three core dimensions that together support the responsible, effective, and sustainable use of linguistic data throughout the translation process. These dimensions reflect key decision-making areas that translators already engage with implicitly, and which require explicit articulation in AI-mediated workflows:

Data governance refers to how linguistic resources are documented, versioned, traced, and made reusable over time. It includes decisions about transparency, responsibility, access conditions, and long-term maintenance of translation data. By addressing transparency and responsibility, this dimension directly operationalises ethical compliance, ensuring that the management of linguistic assets meets both legal standards and professional moral obligations.

Data sourcing concerns the identification and selection of appropriate linguistic resources, taking

into account provenance, relevance, licensing conditions, representativeness, and alignment with the translation brief.

Data processing involves the preparation of selected resources for use, including cleaning, filtering, normalisation, annotation, and other transformations required to ensure that data are fit for a specific translation purpose or for optimising NMT and LLM performance.

From a pedagogical perspective, these aspects can be introduced progressively within translation or specialised translation courses. They can be aligned with familiar stages of the translation process: sourcing decisions are addressed when analysing the brief and selecting documentation; processing decisions are explored during the preparation of translation resources; and governance issues are foregrounded at delivery and closure stages.

Crucially, these dimensions must be considered both at the start and at the end of a translation project. Initial decisions shape how data are assembled and optimised to support the translation task, while closing-phase decisions ensure traceability, ethical compliance, and future reuse. Teaching this lifecycle-oriented perspective reinforces data management as an integral component of professional project preparation and closure.

3.3 Data processing for translation purposes

In the context of AI-mediated translation workflows, data processing constitutes a pivotal set of activities that directly mediate between the translation brief and the behaviour of NMT systems and LLMs. Drawing on Krüger's characterisation of data processing (2024), this section conceptualises data processing not as a neutral technical pipeline, but as a series of context-dependent decisions embedded in professional translation work.

Datasets serve as 'instructional material' to fine-tune NMT and LLM systems. By curating high-quality examples, students 'teach' models to follow specific terminology or styles. This is essentially a documentary task where students select and clean resources to optimize system performance for a given task. From a translation perspective, data processing tasks take place primarily during the project preparation phase, when linguistic resources are identified, assessed, and transformed to ensure alignment with the communicative purpose of the assignment. Typical processing activities include cleaning and filtering translation memories, selecting or discarding segments based on relevance and quality, normalising terminology and style, deduplicating heterogeneous content, and annotating data with minimal but meaningful metadata. Basic annotation entails adding essential metadata (such as domain, genre, context or client labels) to help NMT and LLM systems prioritize relevant linguistic patterns and adapt to specific contextual requirements. These decisions directly influence how NMT systems and LLMs prioritise patterns, handle terminology, and adapt to genre, register, or client-specific constraints once deployed.

Importantly, data processing practices in translation mirror long-standing quality-oriented activities such as revision, validation, and documentary selection, but acquire renewed significance in AI-driven contexts due to the increased sensitivity of models to noisy, heterogeneous, or poorly documented data. Framing processing as a translation-relevant activity enables students to understand how preparatory data work conditions translation outcomes before generation takes place, rather than treating quality issues solely as post-editing problems.

Pedagogically, data processing can be addressed within integrated translation and specialised translation courses that cover the full project lifecycle. By embedding processing decisions alongside brief analysis, documentation, and delivery tasks, students are encouraged to reflect on how theory-informed choices about data preparation contribute to fit-for-purpose MT and LLM performance. This approach reinforces the translator's role as a strategic agent who actively configures linguistic resources to achieve adequacy, consistency, and contextual appropriateness.

3.4 Proposal for a data management training framework

Rather than isolating data-related skills in stand-alone technology courses, the framework is conceived as a transversal structure that can be embedded across translation and specialised translation modules, mirroring the full lifecycle of translation projects. This orientation is also consistent with current discussions around the revision of the European Master's in Translation (EMT) Competence Framework, which increasingly foregrounds data-related competences, including the management, documentation, and responsible reuse of linguistic data in technology-enhanced translation workflows.

The framework is organised around three complementary moments. At the project initiation stage, students learn to assess the translation brief and to identify, source, and preliminarily evaluate linguistic resources with explicit attention to provenance, relevance, and licensing. During the project preparation and execution stage, students engage in data processing activities, such as filtering translation memories, compiling or refining small corpora, and performing basic annotation, to support fit-for-purpose NMT or LLM use. Finally, at the project closure stage, students are explicitly trained to consolidate, document, and govern the linguistic data that have been used or produced, ensuring traceability, ethical compliance, and conditions for future reuse.

Within this framework, two examples of viable pedagogical activities illustrate how data management competences can be operationalised and will be presented and discussed during the workshop presentation:

1. **From project initiation to delivery**: students carry out a complete translation project, from analysing the brief and sourcing data to processing selected resources and delivering the translation with NMT or LLM support. Throughout the project, students document data-related decisions and link them systematically to observed effects on output quality, adequacy, and consistency.
2. **Project closure and documentation reuse**: after project completion, students focus explicitly on the closing phase. They consolidate validated resources, enrich them with minimal metadata (e.g. domain, genre, client, revision status), define reuse or restriction conditions, and reflect on ethical implications.

This strategic approach also integrates broader critical concerns, including the ethical and social implications of data provenance and the ecological sustainability of data-intensive translation workflows. The aim of this framework is not to train translators as data engineers, but to enable them to conceptualise, justify, and document data management decisions as an integral part of translation expertise and professional practice in AI-mediated environments. This activity foregrounds sustainability, efficiency, and responsibility in data reuse (Moorkens, 2022; Weber, 2021).

4 Final remarks

Training in data management represents a significant professional opportunity for translators in an increasingly data-driven language industry, particularly in contexts where NMT and LLMs are integrated into everyday translation workflows. As the performance, reliability, and limitations of these technologies are largely conditioned by the data on which they rely, data management competences enable translators not only to use MT systems, but also to understand, configure, and critically assess them. By developing competences in the governance, processing, and reuse of linguistic data, translators can move beyond the role of end-users of machine translation technologies and position themselves as knowledgeable managers of linguistic assets who actively shape AI-mediated workflows. More broadly, data management training empowers translators to exert greater control over their own production and over the linguistic resources generated through translation work. This competence supports more efficient project preparation, facilitates the responsible reuse of validated documentation, and enhances long-term value creation from translation data. In this sense, data management is not merely a technical add-on, but a strategic skill that strengthens the translator's professional agency and relevance in AI-mediated translation environments.

5 Acknowledgements

The LT-LiDER (ref. 2023-1-ES01-KA220-HED-000161531) has been funded with support from the European Commission. This publication reflects the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

References

Bowker, Lynne. 2026. Teaching translation students about data in the age of generative AI. In Penet, J. C., Joss Moorkens, and Masaru Yamada, editors, *Teaching Translation in the Age of Generative AI: New Paradigm, New Learning?*, pages 67–86. Language Science Press, Berlin.

DePalma, Donald A. and Arle Lommel, 2025. Transforming translation: The evolution and impact of AI on language transfer and communication. In Sun, Sanjun, Kanglong Liu, and Riccardo Moratto, editors, *Translation Studies in the Age of Artificial Intelligence*, pages 18–41. Routledge, London.

GALA. 2025. *Business Barometer Report – Technology, AI, and Automation. Globalization and Localization Association*.

Krüger, Ralph. 2024. Some reflections on the interface between professional machine translation literacy and data literacy. *Journal of Data Mining and Digital Humanities*. Special Issue: Towards a Robotics of Translation, 1–11.

Mellinger, Christopher D. and Thomas A. Hanson. 2025. Data sources and issues of optimisation in translation and interpreting technologies. In Baumgarten, Stefan and Michael Tieber, editors, *The Routledge Handbook of Translation Technology and Society*, pages 300–312. Routledge, London.

Moorkens, Joss. 2022. Ethics and machine translation. In Kenny, Dorothy, editor, *Machine Translation for Everyone. Empowering Users in the Age of Artificial Intelligence*, pages 121–138. Language Science Press, Berlin.

Nord, Christiane. 2013. Functionalism in translation studies. In Millán, Carmen and Francesca Bartrina, editors, *The Routledge Handbook of Translation Studies*, pages 201–212. Routledge, London.

Secară, Alina, María Isabel Rivas Ginel, Antonio Toral, Ana Guerberof Arenas, Dragoș Ciobanu, Justus Brockmann, Claudia Plieseis, Raluca-Maria Chereji, and Caroline Rossi. 2024. *LT-LiDER Language Technology Map – Technologies in Translation Practice and their Impact on the Skills Needed*.

Sánchez-Gijón, Pilar and Ester Torres-Simon, forthcoming. Data management planning. In Castilho, Sheila, Dorothy Kenny, Joss Moorkens, and María Isabel Rivas Ginel, editors, *Developing AI Literacy for Translation*. Language Science Press, Berlin.

Weber, Tobias. 2021. The curation of language data as a distinct academic activity: A call to action for researchers, educators, funders, and policymakers. *Journal of Open Humanities Data*, 7:1–10.