

An AI-Assisted Semantic Platform enabling Pesticide Risk Assessment

Ugo Gentile^{1,*†}, Till Goldmann^{1,†}, Francesco Carratta^{2,†} and Julie Moulin^{1,†}

¹*Institute of Food Safety and Analytical Sciences, Nestlé Research 1000 Lausanne, Switzerland*

²*Strategic Research and Operation, Nestlé Research 1000 Lausanne, Switzerland*

Abstract

Pesticide risk assessment requires the integration of heterogeneous regulatory (registration status, maximum residue levels), physico-chemical and analytical (monitoring and alert) data across multiple sources, making the process complex, time-consuming, and difficult to scale. This work proposes a semantic and AI-driven methodology that combines transformer-based Natural Language Processing (NLP) models with a domain-specific ontology to enable automated and consistent risk assessment. NLP techniques extract and normalize relevant entities from unstructured data, while the ontology provides a formal semantic layer to harmonize concepts and model domain-specific relationships. A key contribution is the introduction of a denominator concept, which aggregates chemically related entities, such as salts and metabolites, under a common reference, ensuring alignment between regulatory and analytical perspectives. The approach is currently evaluated in real-world use cases, reducing assessment effort substantially and supporting proactive risk monitoring. This work demonstrates how combining NLP and semantic technologies can enable scalable, explainable decision support in food safety.

Keywords

AI, NLP, Ontology, Chemical Risk Assessment, Pesticide

1. Introduction

Ensuring food quality and safety requires controlling contaminant residues in raw materials across global supply chains. One prominent category of contaminants is pesticides, which can leave behind trace residues that contaminate the ingredients used in food products.

The food industry needs accurate, comprehensive, and up-to-date information on both the pesticides that may have been used within its supply chain and the regulations that apply to them in order to ensure that each product is compliant, safe to consume, and produced using ingredients cultivated in a sustainable manner.

This task is particularly challenging due to (i) the large number of pesticide compounds, (ii) the fragmented and continuously evolving regulatory frameworks across countries, and (iii) the heterogeneity of data sources providing information on pesticides, including regulatory (registration status, maximum residue levels), physico-chemical and analytical (monitoring and alert) data. Traditional approaches to pesticide risk assessment often require significant expert effort and can be challenging to scale when dealing with increasing data volume and diversity. Assessments often require weeks of expert analysis and can, in some cases, become infeasible given the volume and diversity of available data. As a result, these approaches may suffer from limited coverage, as only a subset of compounds can be assessed, as well as from a lack of harmonization across naming conventions and formats. Timely integration of regulatory, physico-chemical and analytical evidence is important to support the identification and assessment of emerging risks.

Posters, Demos, Blue Sky, and Tutorials at SEMANTiCS 2026, Sep 2026, Ghent, Belgium

*Corresponding author.

†These authors contributed equally.

✉ ugo.gentile@rd.nestle.com (U. Gentile); till.goldmann@rdls.nestle.com (T. Goldmann); francesco.carratta1@rd.nestle.com (F. Carratta); julie.moulin1@rdls.nestle.com (J. Moulin)

ORCID 0000-0001-5158-1835 (U. Gentile)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

To address these challenges, recent advances in Natural Language Processing (NLP) provide effective solutions for handling large volumes of heterogeneous and unstructured data. NLP can now enable robust extraction and normalization of entities from noisy textual sources, and is particularly effective at capturing the local semantics of a sentence. However, it remains limited in its ability to consistently link information across multiple documents, sources, or contexts, especially when domain-specific relationships and implicit connections need to be inferred. Addressing such cross-context reasoning and ensuring semantic consistency requires explicit knowledge modeling, typically provided by ontologies. By formally representing entities, relationships, and domain rules, ontologies enable the integration and alignment of information across heterogeneous sources.

The approach presented in this work addresses these limitations by combining NLP-based extraction with ontology-driven semantic harmonization, together with expert-defined rules. The result is an AI-assisted risk assessment platform having at its core a custom pesticide ontology allowing chemically related entities to be grouped according to their functional role in risk assessment. The ontology is not used as a static vocabulary but as an active component of the analytical pipeline, supporting semantic harmonization, knowledge inference, and explainable decision-making at scale.

2. Related Work

Transformer-based models, such as BERT (Bidirectional Encoder Representations from Transformers) [1], have achieved state-of-the-art performance in tasks such as Named Entity Recognition (NER) and Entity Linking (EL), enabling the extraction and normalization of entities from unstructured data. Recent work has combined NLP with knowledge graphs and ontologies to improve data integration and reasoning capabilities [2]. Ontologies are widely used to structure knowledge and support interoperability across datasets. Resources such as FoodOn [3], FoodEx2 [4], and ChEBI [5] provide comprehensive vocabularies for the food and chemical domains, enabling semantic querying and integration. However, these ontologies are typically designed as static conceptual models and do not address operational requirements such as dynamically linking regulatory definitions with analytical observations in risk assessment pipelines.

In the pesticide domain, applications of ontologies remain limited. Resources such as the Pesticide Properties DataBase (PPDB) [6] provide structured descriptions of physicochemical and toxicological properties, while more recent efforts have explored semantic frameworks and knowledge graphs for data integration and decision support [7]. Nevertheless, these approaches primarily focus on data representation and retrieval, with limited support for reasoning across heterogeneous contexts. In particular, they do not adequately model relationships involving metabolites and salts, or complex regulatory residue definitions, which are essential for the consistent aggregation and assessment of risk. As a result, existing systems struggle to reconcile regulatory, physico-chemical and analytical data that refer to different representations of the same substance.

While ontologies provide a foundation for structuring knowledge, they are rarely operationalized to support active reasoning. This limits their ability to derive the implicit knowledge required for comprehensive risk assessment and to capture domain-specific interpretations, particularly the need for aggregation mechanisms that ensure coherent and consistent decision-making.

3. AI-assisted pesticide risk assessment

The proposed methodology combines NLP models with a pesticide ontology to enable scalable integration and interpretation of heterogeneous data for pesticide risk assessment. NER and EL are used to extract and normalize domain-relevant entities (Pesticides, Foods, and Locations), while the ontology provides structured representation and reasoning capabilities, ensuring consistent and interpretable outputs across datasets. Finally, the data are provided as input to an expert-driven, multi-criteria risk assessment process and are consumed by tailor-made end-user modules designed according to specific user needs, as depicted in Figure 1.

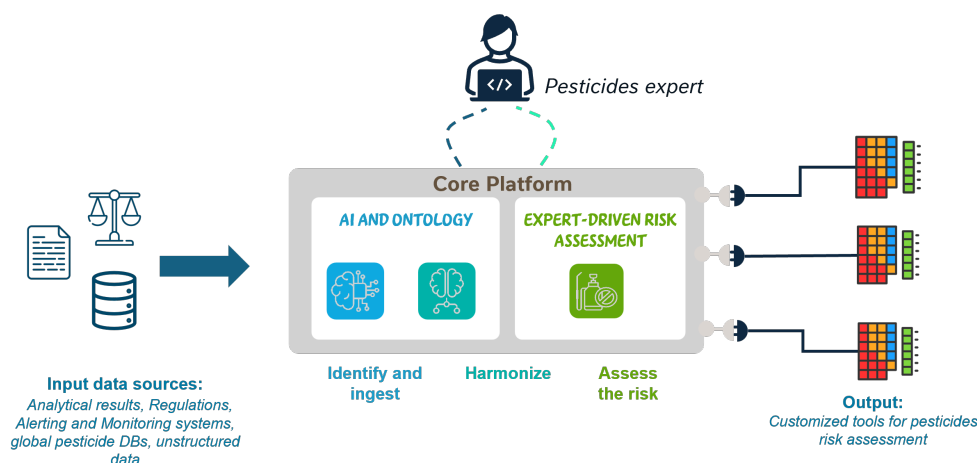


Figure 1: AI-Assisted Pesticide Risk Assessment platform

The core of the framework is a dedicated pesticide ontology developed to address the intrinsic complexity of the domain. Pesticide-related information is fragmented across regulatory (registration status, maximum residue levels), analytical (monitoring and alert) and physico-chemical transformations, where a single substance may exist in multiple forms, such as salts, metabolites, and esters, and may be represented differently depending on the information source considered. For example, assessing the potential risk of a compound may require evaluating both its regulatory registration status and its detection in analytical results obtained from a specific food material. However, an active substance may be registered at the level of a salt (e.g., a sodium form), while analytical results detect only its metabolites, and regulatory limits may be associated with complex residue definitions that consider the sum of multiple chemicals.

Without an explicit semantic link between these representations, information on pesticides cannot be reconciled. For this reason, the pesticide ontology has been designed through the consolidation of multiple data sources and expert curation, with a strong emphasis on consistent terminology and domain-specific relationships. It introduces an aggregation mechanism named *denominator* to group chemically related entities according to their functional role, enabling alignment between data perspectives mentioned above. By embedding these representations within the methodology, the ontology acts not only as a harmonization layer but also as a foundation for semantic reasoning, supporting more consistent, scalable, and explainable decision support.

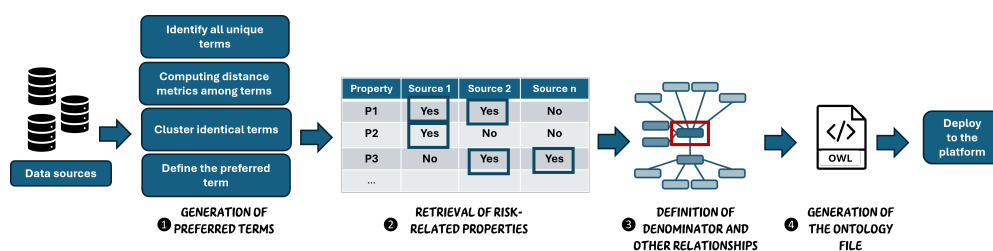


Figure 2: Building process of the pesticide ontology

The construction of the ontology follows a structured multi-step workflow illustrated in Figure 2. The process begins with the consolidation of chemical entities from various sources, together with their key descriptors (e.g., CAS, InChIKey, ...) as well as source-specific properties that will later be used in the risk assessment. Assessing the similarity between chemical entities from the various sources using their descriptors allows identical terms to be clustered under the same umbrella, noting that specific terms potentially missing from the sources may still be added manually. A preferred term is then defined for each cluster (1).

Once aligned, relevant properties are combined through data cleaning, harmonization, and consolidation steps (2). This enables the integration of complementary information (e.g., compound toxicity) into unified entity representations. At this stage, relationships between entities, such as salts, metabolites, and parent compounds, are explicitly established through domain rules validated by experts.

The denominator is then derived as an aggregation mechanism by identifying the core chemical entity that semantically links related preferred terms across contexts (3). Concretely, this is achieved by analyzing transformation relationships (e.g., "is metabolite of", "is salt of") and regulatory definitions, and selecting the reference compound that acts as the common entity linking reporting and interpretation. For example, a parent compound may be selected as the denominator when it serves as the basis for residue definitions and connects both registered forms (e.g., salts) and analytically detected forms (e.g., metabolites).

Finally, the ontology is enriched with axioms encoding these relationships and serialized into an OWL representation (4), enabling reasoning over the constructed graph. An example of such an axiom is reported in Equation (1): if a compound belongs to a denominator, and this denominator is included in a specific list, then the compound itself is also treated as part of that list.

$$\text{Parameter}(p) \wedge \text{hasDenominator}(p, d) \wedge \text{SpecificList}(sl) \wedge \text{hasReference}(sl, d) \rightarrow \text{InSpecificList}(p, \text{true}) \quad (1)$$

The resulting structure allows the system to consistently link regulatory information with analytical findings, providing a semantically grounded aggregation mechanism essential for risk assessment.

4. Results and Discussion

The proposed approach demonstrated measurable operational benefits and qualitative improvements in the pesticide risk assessment, with the ontology acting as a central enabler. In the deployed environment, assessment effort was substantially reduced and coverage is now reaching thousands of pesticide-related entities and multiple heterogeneous sources integrated into a unified framework.

The added value of the approach has been confirmed by target business users thanks to its deployment in a real-world setting. Compared with previous manual workflows, users reported that the system supported the identification and review of relevant risk factors and greater consistency in the interpretation of evidence originating from multiple sources. In some cases, the system identified additional relevant signals compared with prior assessment approaches or allowed to reassess existing risks based on a more complete body of evidence. Consequently, experts perceived the resulting evaluations as more comprehensive and better aligned with the available regulatory and analytical evidence. A more quantitative evaluation of the impact on assessment quality is currently ongoing, based on the adoption of the tool by target business users in their daily work.

4.1. Performance evaluation of NER and EL

The NLP components were evaluated on independently curated expert datasets representative of real-world pesticide risk assessment activities. The datasets include multilingual regulatory documents (registration status, maximum residue levels), scientific reports and analytical (monitoring and alert) records. To ensure representative evaluation conditions, each dataset was partitioned using a stratified 70/30 train-test split, preserving label frequencies across both subsets. All reported metrics were computed exclusively on the held-out test set.

Named Entity Recognition (NER) was evaluated on 4,218 manually annotated samples containing contaminant and measurement-unit entities. As shown in Table 1, the model achieved precision, recall and F1 score consistent with strong NER systems described in the most recent benchmarks [8] for comparable domains.

Entity Linking (EL) was evaluated separately for contaminant, food, and location entities. The contaminant-linking task represents the most challenging setting due to the ontology search space of

approximately 5,000 candidate concepts. Nevertheless in all the cases its performance was consistent with ranges commonly reported for state-of-the-art chemical entity normalization approaches on benchmark datasets [9].

Table 1

NER evaluation dataset and performance.

Entity Types	Samples	Threshold	Precision	Recall	F1
Contaminant, Unit	4,218	≥ 0.80	89%	91%	90%

Table 2

Entity Linking evaluation datasets and performance.

EL Type	Samples	Taxonomy Size	Threshold	Precision	Recall	F1
Contaminant	11,230	~5,000	≥ 0.95	83%	79%	81%
Food	8,314	175	≥ 0.80	91%	89%	90%
Location	2,894	294	≥ 0.80	91%	85%	87%

4.2. Ontology evaluation

The ontology plays a central role in the proposed framework by enabling the harmonization of heterogeneous chemical representations originating from the different sources considered. Overall, 34,698 source-specific chemical representations collected from seven independent datasets were consolidated into 5,875 harmonized concepts connected through 22,871 semantic relationships. This semantic normalization allows the automatic reconciliation of regulatory, toxicological, and analytical information referring to the same underlying substance despite variations in naming conventions, identifiers, salts, metabolites, or residue definitions. As a result, fragmented information can be integrated into a coherent and explainable knowledge structure suitable for large-scale risk assessment. From an operational perspective, the ontology also streamlines the onboarding of new compounds through automated identifier alignment, concept classification, and relationship generation, reducing the integration effort to approximately three minutes per compound.

Evaluating completeness is inherently challenging in a domain as continuously evolving as pesticides. In the context in which the ontology has been deployed, a more meaningful measure is operational coverage and continuous improvement.

The system has been designed such that, when the NLP pipeline encounters an entity that cannot be linked with sufficient confidence, the case is analyzed to determine whether the concept is missing from the ontology, in which case it is added through the curation workflow, or whether the issue originates from the entity-linking model, in which case the example is incorporated into the training set. Furthermore, the ontology is regularly enriched using key regulatory and reference sources. It is important to note that the ontology contains substantially more terms than the preferred-term lists used for harmonization, as it also captures synonyms, commercial names, analytical denominations, metabolites, salts, and source-specific representations. Therefore, practical coverage is continuously assessed through operational use and ongoing ontology enrichment rather than through a fixed completeness measure.

5. Conclusion

This work shows that combining Artificial Intelligence with domain-specific semantic technologies can support the digitalization of pesticide risk assessment. While NLP supports large-scale extraction from unstructured data, the ontology ensures semantic consistency, enables cross-context reasoning, and transforms data into structured knowledge. By explicitly modeling domain relationships and

embedding reasoning into the analytical pipeline, the approach provides explainable and consistent decision support.

The results highlight that the semantic layer is not only a data harmonization tool but also a core component of scalable and reliable risk assessment, bridging regulatory, scientific and analytical domains that are typically disconnected. The approach illustrates the value of combining cutting-edge AI technologies with widely adopted semantic technologies to support proactive, data-driven decision-making.

Ongoing work focuses on improving ontology scalability through data-driven discovery of relevant concepts, enabling semi-automatic suggestions for new ontology branches. This aims to support continuous enrichment while reducing expert effort and keeping the model aligned with evolving knowledge and regulations.

Declaration on Generative AI

During the preparation of this work, the authors used GPT 5.6 for grammar and spelling check. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

- [1] J. Burstein, C. Doran, T. Solorio (Eds.), BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, Association for Computational Linguistics, Minneapolis, Minnesota, 2019. URL: <https://aclanthology.org/N19-1423/>. doi:10.18653/v1/N19-1423.
- [2] A. Hogan, E. Blomqvist, M. Cochez, C. D'amato, G. D. Melo, C. Gutierrez, S. Kirrane, J. E. L. Gayo, R. Navigli, S. Neumaier, A.-C. N. Ngomo, A. Polleres, S. M. Rashid, A. Rula, L. Schmelzeisen, J. Sequeda, S. Staab, A. Zimmermann, Knowledge graphs, *ACM Comput. Surv.* 54 (2021). URL: <https://doi.org/10.1145/3447772>. doi:10.1145/3447772.
- [3] D. M. Dooley, E. J. Griffiths, G. S. Gosal, P. L. Buttigieg, R. Hoehndorf, M. C. Lange, L. M. Schriml, F. S. L. Brinkman, W. W. L. Hsiao, Foodon: a harmonized food ontology to increase global food traceability, quality control and data integration, *npj Science of Food* 2 (2018). URL: <https://doi.org/10.1038/s41538-018-0032-6>. doi:10.1038/s41538-018-0032-6.
- [4] EFSA (European Food Safety Authority), The food classification and description system FoodEx2, Technical Report EN-804, EFSA Supporting Publications, 2015. URL: <https://doi.org/10.2903/sp.efsa.2015.EN-804>. doi:10.2903/sp.efsa.2015.EN-804.
- [5] J. Hastings, G. Owen, A. Dekker, M. Ennis, N. Kale, V. Muthukrishnan, S. Turner, N. Swainston, P. Mendes, C. Steinbeck, Chebi in 2016: Improved services and an expanding collection of metabolites, *Nucleic Acids Research* 44 (2016) D1214–D1219. URL: <https://doi.org/10.1093/nar/gkv1031>. doi:10.1093/nar/gkv1031.
- [6] University of Hertfordshire, The pesticide properties database (ppdb), <https://sitem.herts.ac.uk/aeru/ppdb/>, n.d. Agriculture & Environment Research Unit, accessed 2026.
- [7] C. Caracciolo, A. Stellato, A. Morshed, G. Johannsen, S. Rajbhandari, Y. Jaques, J. Keizer, The agrovoc linked dataset, *Semant. Web* 4 (2013) 341–348.
- [8] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, J. Kang, Biobert: a pre-trained biomedical language representation model for biomedical text mining, *Bioinformatics* 36 (2020) 1234–1240. URL: <https://doi.org/10.1093/bioinformatics/btz682>. doi:10.1093/bioinformatics/btz682.
- [9] H. Kim, M. Sung, W. Yoon, S. Park, J. Kang, Improving tagging consistency and entity coverage for chemical identification in full-text articles, *CoRR abs/2111.10584* (2021). URL: <https://arxiv.org/abs/2111.10584>. arXiv:2111.10584.