Chemical Identity Lost in Regulation: A Study of Semantic Interoperability in European Chemical Substance Data

Geert Van Haute[1,*], Stijn Goedertier[1], Pieter Fannes[1] and Maxim Van de Wynckel[1]

[1]*Environment Department, Flemish Government, Brussels, Belgium*

Abstract

European regulatory datasets represent chemical substances heterogeneously: CAS numbers, EC numbers, and textual names that may refer to single molecules, mixtures, substance groups, or analytical parameters. We analyse 18 ECHA-based datasets (41,813 records) using a seven-tier linkability taxonomy. Only 62.5% of entries link to a defined molecular structure; 37.5% are non-structure-based. CAS numbers prove unreliable as unique identifiers. Four sentence-embedding models achieve ChemOnt Hit@1 $\leq$ 5.75%; Claude Sonnet 4.6 reaches 39.2% yet still misclassifies the majority. Structure-defined substances are integrated into a knowledge graph with ChemOnt classification and ChEBI biological roles (~48,400 cross-domain links). For non-structure entries, the LLM abstains from direct classification far more often than for structure-defined substances, appropriately reflecting their lack of a defined structure; a separate LLM-based scope-mapping assessment yields 304 validated SKOS triples linking regulatory group entries to ChemOnt nodes. Semantic interoperability requires structural changes at the level of regulatory data modelling and legislation.

Keywords

semantic interoperability, regulatory data, chemical substance identification, knowledge graph, REACH

1. Introduction and Related Work

In chemistry, substance identity is operationalised through structure-based identifiers such as the International Chemical Identifier (InChI) and InChIKey [1], enabling unambiguous linking across databases such as PubChem [2] and the Chemical Entities of Biological Interest database (ChEBI) [3]. In European regulation, defined under the Registration, Evaluation, Authorisation and Restriction of Chemicals framework (REACH) [4] and the Classification, Labelling and Packaging regulation (CLP) [5], a *substance* may refer to a single molecule, a mixture of unknown or variable composition (UVCB) [6], a group of related compounds (e.g. per- and polyfluoroalkyl substances, PFAS, which encompasses individual molecules such as PFOS without explicit set-membership semantics; Figure 2), or an analytical sum parameter, as summarised in Figure 1.

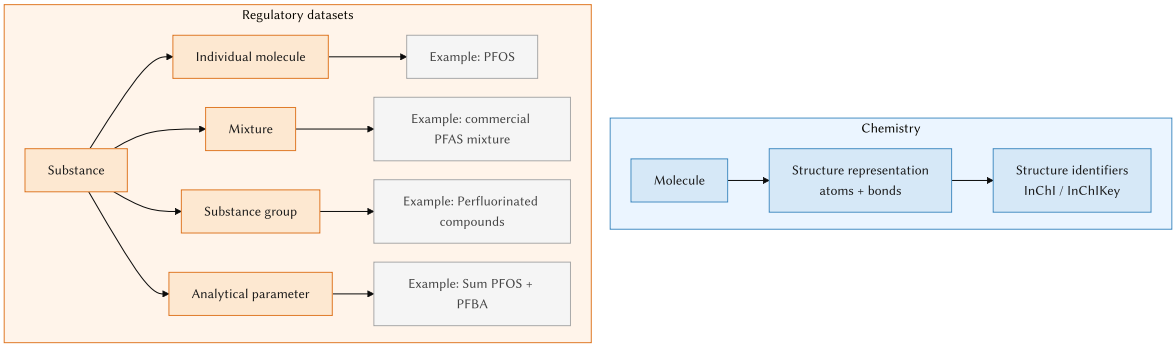

Figure 1: Conceptual contrast between chemistry (structure-based identity via InChI/InChIKey) and regulatory datasets (heterogeneous substance types).

Posters, Demos, Blue Sky, and Tutorials at SEMANTiCS 2026, Sep 2026, Ghent, Belgium
geert.vanhaute@vlaanderen.be (G. Van Haute)
0009-0002-9836-8139 (G. Van Haute); 0009-0003-4602-6601 (S. Goedertier); 0009-0001-7859-1977 (P. Fannes); 0000-0003-0314-7107 (M. Van de Wynckel)

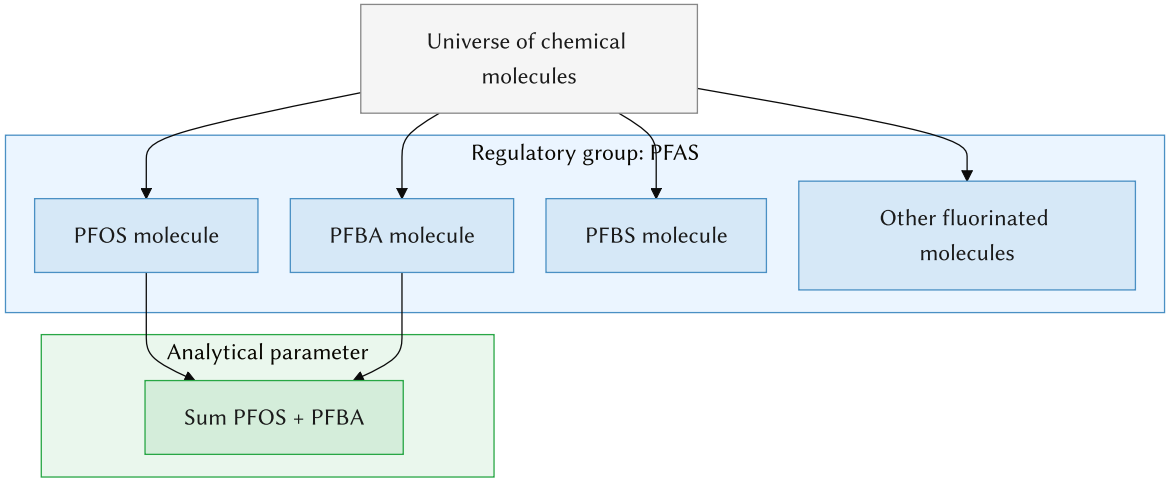

Figure 2: PFAS as a regulatory group without explicit set-membership semantics; *Sum PFOS + PFBA* as a derived analytical parameter.

Such regulatory entities (whether single molecules, UVCBs, substance groups, or sum parameters) are identified by Chemical Abstracts Service (CAS) numbers [7], textual names, or both—identifiers that carry no structural information. Much of this information remains confidential or inaccessible in machine-readable form: REACH restricts public disclosure of the full Substance Identity Profile and UVCB composition [8], leaving only non-machine-readable safety data sheets and classifications publicly available [9, 10]. This semantic mismatch limits automated integration [11] and FAIR (Findable, Accessible, Interoperable, Reusable) data reuse [12]. Prior work shows that structure-based identifiers and ontology-driven knowledge graphs enable interoperability in chemistry and in single-domain regulatory contexts such as cosmetics risk assessment [13], while legal-data integration across jurisdictions remains fragmented [14]. However, the cross-dataset mismatch between structure-based chemical identity and regulatory substance definitions has not previously been systematically characterised.

We address three research questions: (RQ1) To what extent can regulatory substances be linked to a unique chemical structure? (RQ2) How reliable are CAS numbers as cross-dataset identifiers? (RQ3) Can embedding-based or LLM methods compensate for absent structure identifiers? Contributions: (C1) a seven-tier linkability taxonomy; (C2) empirical demonstration of CAS inconsistency; (C3) a benchmark showing the failure of embedding- and LLM-based approaches; (C4) an integrated knowledge graph linking 18 sources with ChemOnt [15] and ChEBI; (C5) 304 validated SKOS (Simple Knowledge Organization System) [16] mapping triples as a legislative–ChemOnt crosswalk.

2. Methods

Data integration and linkability. We integrated 18 sources: six European Chemicals Agency (ECHA) obligation lists (Candidate [17], Authorisation, Restriction, POPs, EU Positive, CLH), nine ECHA activity lists, REACH registrations [18], the EU Pesticides Database, OSPAR priority chemicals, and a Flemish administrative substance list. Per entry, CAS, EC number [19], and substance name were collected; substances were linked to InChIKey via PubChem [2]. Each entry was then assigned to one of seven tiers: tier 1 (direct InChIKey link), tiers 2–3 (partially resolvable), tier 4 (unresolved CAS), tier 5 (mixture or UVCB [6]), tier 6 (administrative entry), tier 7 (no usable identifier). InChI extensions such as MInChI [1] support representation of some mixtures in principle, and computational methods exist to select representative constituent structures for UVCBs [6], but neither is reflected in the regulatory datasets studied here; tier 5 entries therefore resist structure-based linking as currently published. Bidirectional CAS–InChIKey mappings were analysed to assess identifier consistency [20].

Embedding- and LLM-based matching. We test the hypothesis that embedding similarity can serve as a proxy for chemical taxonomy assignment, expecting it to fail because substance names for groups and mixtures carry no structural information. Substance names and ~3,800 ChemOnt class labels [15] were encoded using four sentence-transformer models [21] (MiniLM, MPNet, SciBERT, BioBERT). Performance was measured using Hit@k against 28,204 ClassyFire-assigned, tier-1 (structure-linked) substances as ground truth, since only these carry a verifiable ChemOnt class. Given the embedding models' near-complete failure on this benchmark (Table 1), they were not carried forward to the harder, tier 4/5/7 non-structure population. Claude Sonnet 4.6 was applied to the same tier-1 benchmark under zero-shot prompting, then, for the tier 4/5/7 non-structure entries, further used to (a) classify ChemOnt membership directly and (b) assess how the entry's name relates to a ChemOnt node (exact, subset, or broader scope), using SKOS mapping types [16]. Structure-defined substances were integrated into an RDF knowledge graph with ChEBI biological roles [3], and a ToxPi-inspired composite priority score [22] was computed to illustrate the analytical value.

3. Results

Linkability (RQ1). Integration of 18 sources yielded 41,813 records. Applying the seven-tier taxonomy to 33,452 unique entries (Figure 3) shows that 62.5% (20,909) link directly via InChIKey (tier 1), while 37.5% cannot be mapped to a defined structure: 11.6% carry unresolved CAS numbers (tier 4), 10.8% are mixtures or UVCBs (tier 5), and 14.9% lack any usable identifier (tier 7). Tier composition differs substantially between regulatory instruments. Tiers 2–3 and 6 together cover under 0.3% of entries here; the taxonomy retains them to generalise to sources where these categories are populated.

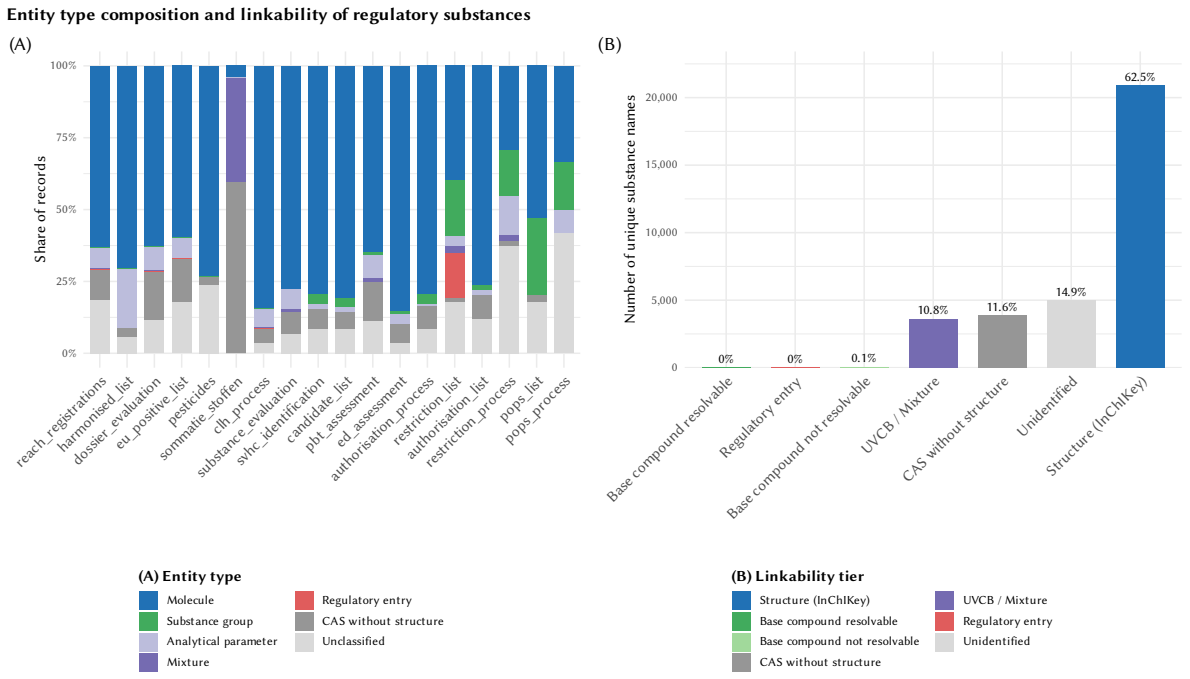

Figure 3: (A) Entity type composition across 18 sources. (B) Seven-tier linkability taxonomy: 62.5% of 33,452 unique entries link directly via InChIKey (tier 1); 14.9% lack any usable identifier (tier 7).

CAS number reliability (RQ2). 70.0% of structure-resolved InChIKeys appear in only a single source, making CAS-based cross-dataset linking fundamentally insufficient as a shared key. Ambiguity compounds this: 195 InChIKeys are each associated with more than one CAS number (up to seven), while 4 CAS numbers resolve to two distinct InChIKeys [20, 23]; for example, CAS 26471-62-5 (toluene diisocyanate) maps to two positional isomers with different InChIKeys across sources. Even for structure-

defined substances, then, a flat CAS number cannot substitute for a structure-based identifier: the same semantic gap that leaves non-structure entities unlinked entirely (RQ1).

Failure of embedding- and LLM-based approaches (RQ3). Table 1 shows that, on the tier-1 (structure-linked) benchmark, embedding-based matching fails: the best model (`all-MiniLM-L6-v2`) achieves Hit@1 of 5.75%, meaning the correct ChemOnt class is retrieved in fewer than 1 in 17 attempts; all others score below 1%. Claude Sonnet 4.6 reaches Hit@1 $= 39.2\%$ (seven-fold improvement); high-confidence assignments reach 53.9% accuracy while low-confidence cases score 0%. The LLM still misclassifies the majority of substances. Applied to the tier 4/5/7 non-structure entries that make up the 37.5% unlinked population (RQ1), the LLM abstains from classification far more often than for structure-defined substances, confirming that textual descriptions cannot substitute for structural information. The same embedding pipeline, applied directly to this non-structure population rather than the tier-1 benchmark, confirms the same failure from the other direction: of 12,522 candidate substance names, only 132 (1.1%) pass even the permissive $\text{score}_1 > 0.80$ threshold, so the benchmark result is not an artefact of the evaluation set.

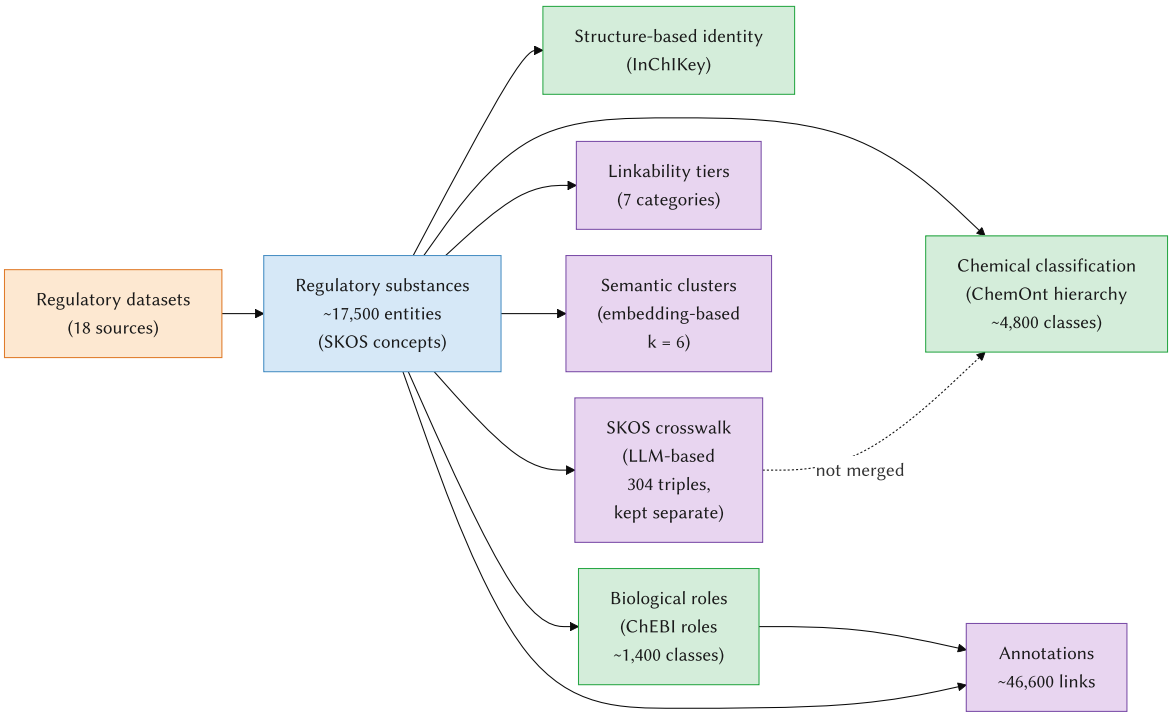

Figure 4: Overview of the integrated knowledge graph: regulatory substances (SKOS concepts) are linked to structure-based identity (InChIKey), chemical classification (ChemOnt), and biological roles (ChEBI), with derived linkability tiers, embedding clusters, and an LLM-based SKOS crosswalk kept separate from (not merged into) the ChemOnt hierarchy.

Knowledge graph and SKOS crosswalk (C4, C5). Figure 4 summarises the integrated knowledge graph: ~17,500 regulatory substances (SKOS concepts) are linked to structure-based identity, chemical classification, and biological roles. This is fewer than the 20,909 tier-1-linked entries (RQ1) because names sharing one InChIKey collapse into a single concept. Structure-defined substances were integrated across lists, ChemOnt, and ChEBI, yielding ~48,400 cross-domain links (40,644 list memberships, 7,794 ChEBI roles) and enabling composite priority scoring [22] infeasible on fragmented datasets.

LLM-based scope-mapping identifies 31.7% of sampled non-structure entries as mappable to a ChemOnt node (vs. 0.8% by pattern matching), yielding 304 high-confidence SKOS triples (285 `broadMatch`, 18 `exactMatch`, 1 `narrowMatch`). The dominance of `broadMatch` reflects that

Model	Hit@1	Hit@3	Hit@5	Hit@10
all-MiniLM-L6-v2	0.0575	0.0986	0.1234	0.1621
all-mpnet-base-v2	0.0093	0.0152	0.0184	0.0247
SciBERT	0.0040	0.0071	0.0102	0.0131
BioBERT	0.0047	0.0077	0.0099	0.0150
Claude Sonnet 4.6[†]	0.392	—	—	—

[†]LLM: single label; only Hit@1 is comparable across rows.

Table 1
ChemOnt classification performance (direct-parent level). Ground truth: 28,204 ClassyFire-assigned substances [15].

regulatory group entries typically define constrained subsets of ChemOnt classes (e.g. *fatty acids, C16–22 ⊂ Fatty acids and conjugates*), making scope machine-interpretable. The remaining 68% of non-structure entries could not be reliably mapped.

4. Discussion and Conclusion

Over one-third of regulatory substance entries cannot be linked to a molecular structure; CAS numbers introduce ambiguity rather than resolving it; and neither embeddings nor LLMs can substitute for structural information. Full manual reconciliation of non-structure entries (an estimated 3,150 person-years for REACH alone[1]) is infeasible.

These findings show that downstream harmonisation cannot substitute for change at the source: the different regulatory instruments that define and collect substance data (REACH, CLP, and the sector-specific lists built on them) must encode structure-based and group-level semantics consistently, not only the systems that integrate them afterwards. Three recommendations follow: (1) require structure-based identifiers (InChIKey; MInChI [1] for mixtures, or representative-constituent structures [6] for UVCBs where a single structure does not apply) across regulatory instruments; (2) mandate explicit representation of substance groups using formal ontologies such as ChemOnt, so that group membership is computed automatically from structure via ClassyFire rather than maintained as a static, manually curated list — shown feasible here by the 304 SKOS crosswalk triples (C5); (3) publish regulatory data as linked data [24] with persistent URIs and SKOS taxonomy links, enabling FAIR-compliant [12] integration under the Interoperable Europe Act [11]. Future work should assess generality beyond ECHA/EU data (e.g. US EPA, UK HSE) and develop curated resolution workflows for the 37.5% of entries that resist automated linking, guided by the taxonomy introduced here.

Limitations: Scoped to 18 ECHA-based datasets; ClassyFire ground truth is itself algorithmically assigned, not expert-validated, and a manual sample shows most apparent LLM errors select a chemically valid alternative label rather than a hallucinated one; LLM evaluation uses a single commercial model (Claude Sonnet 4.6); 68% of non-structure entries resist automated alignment. *Code and data:* https://doi.org/10.5281/zenodo.21512303 [25].

Declaration on the Use of Generative AI

Research subject. Claude Sonnet 4.6 (`claude-sonnet-4-6`, Anthropic) was used as the *object of study* in three analyses reported in this paper: LLM-based ChemOnt classification benchmarking (to explore whether an LLM can compensate for the absence of structure-based identifiers) and LLM-based equivalence assessment for SKOS mapping of legislative group entries to ChemOnt nodes (to explore whether a legislative–ChemOnt crosswalk is constructible). The model was queried via the Anthropic API with prompt caching enabled at an API cost of approximately $100.

[1]Worst-case pairwise reconciliation (comparing every pair; real reconciliation would cluster) of $n \approx 11{,}250$ non-structure entries needs $n(n-1)/2 \approx 6.3 \times 10^7$ comparisons; at ~100/day and 200 working days/year, that is $\approx 3{,}150$ person-years.

Development tool. Claude Code (Anthropic, model: `claude-sonnet-4-6`) was used during the preparation of this work for: (1) writing, debugging, and documenting analysis code in R, Bash, and Python; and (2) assisting with system configuration to enable local execution of transformer-based embedding models. These locally executed embedding models consumed negligible additional energy relative to the LLM API experiments. After using this tool, the author reviewed and edited all content as needed and takes full responsibility for the publication's content.

References

[1] S. R. Heller, A. McNaught, I. Pletnev, S. Stein, D. Tchekhovskoi, InChI, the IUPAC international chemical identifier, Journal of Cheminformatics 7 (2015) 23. doi:10.1186/s13321-015-0068-4.

[2] S. Kim, P. A. Thiessen, E. E. Bolton, S. H. Bryant, Pug-soap and pug-rest: web services for programmatic access to chemical information in pubchem, Nucleic Acids Research 43 (2015) W605–W611. doi:10.1093/nar/gkv396.

[3] K. Degtyarenko, et al., ChEBI: a database and ontology for chemical entities of biological interest, Nucleic Acids Research 36 (2008) D344–D350. doi:10.1093/nar/gkm791.

[4] European Parliament, Council of the European Union, Regulation (EC) No 1907/2006 concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH), Regulation L 396, Official Journal of the European Union, 2006. URL: http://data.europa.eu/eli/reg/2006/1907/2022-03-01.

[5] European Parliament, Council of the European Union, Regulation (EC) No 1272/2008 on classification, labelling and packaging of substances and mixtures (CLP), Regulation L 353, Official Journal of the European Union, 2008. URL: http://data.europa.eu/eli/reg/2008/1272/oj.

[6] S. S. Kutsarova, et al., UVCB substances II: Development of an endpoint-nonspecific procedure for selection of computationally generated representative constituents, Environmental Toxicology and Chemistry 38 (2019) 682–694. doi:10.1002/etc.4358.

[7] Chemical Abstracts Service, Cas registry, https://www.cas.org, 2024. Accessed: 2026-04-08.

[8] E. Parliament, C. of the European Union, Regulation (ec) no 1907/2006 reach, article 118: Confidentiality, 2006. URL: http://data.europa.eu/eli/reg/2006/1907/2025-10-23.

[9] E. Parliament, C. of the European Union, Regulation (ec) no 1907/2006 reach, article 31: Requirements for safety data sheets, 2006. URL: http://data.europa.eu/eli/reg/2006/1907/2025-10-23.

[10] E. Parliament, C. of the European Union, Regulation (ec) no 1907/2006 (reach), article 119: Information on substances to be made publicly available, 2006. URL: http://data.europa.eu/eli/reg/2006/1907/2025-10-23.

[11] European Parliament, Council of the European Union, Regulation (EU) 2024/903 (Interoperable Europe Act), Regulation L 2024/903, Official Journal of the European Union, 2024. URL: http://data.europa.eu/eli/reg/2024/903/oj.

[12] M. D. Wilkinson, et al., The FAIR guiding principles for scientific data management and stewardship, Scientific Data 3 (2016) 160018. doi:10.1038/sdata.2016.18.

[13] S. Sepehri, A. Heymans, D. De Win, J. Maushagen, A. Sanctorum, C. Debruyne, R. M. Rodrigues, J. De Kock, V. Rogiers, O. De Troyer, T. Vanhaecke, The TOXIN knowledge graph: Supporting animal-free risk assessment of cosmetics, Database 2025 (2025). doi:10.1093/database/baae121.

[14] E. Filtz, S. Kirrane, A. Polleres, The linked legal data landscape: linking legal data across different countries, Artificial Intelligence and Law 29 (2021) 485–539. doi:10.1007/s10506-021-09282-8.

[15] Y. Djoumbou Feunang, et al., ClassyFire: automated chemical classification with a comprehensive, computable taxonomy, Journal of Cheminformatics 8 (2016) 61. doi:10.1186/s13321-016-0174-y.

[16] A. Miles, S. Bechhofer, SKOS Simple Knowledge Organization System Reference, W3C Recommendation, World Wide Web Consortium, 2009. URL: https://www.w3.org/TR/skos-reference/.

[17] European Chemicals Agency (ECHA), Candidate list of substances of very high concern for authorisation, https://chem.echa.europa.eu/obligation-lists/candidateList, 2025. Substances meeting SVHC criteria under REACH Article 57; updated periodically. Accessed: 15 April 2026.
[18] European Chemicals Agency (ECHA), Reach registered substances database, https://echa.europa.eu/information-on-chemicals/registered-substances, 2025. Database of substances registered under REACH, including identity information and hazard data. Accessed: 15 April 2026.
[19] European Chemicals Agency (ECHA), About EC and list numbers, https://echa.europa.eu/en/about-ec-and-list-numbers, n.d. Explanation of the EC number system (EINECS, ELINCS, NLP) used to identify chemical substances in EU legislation. Accessed: 15 April 2026.
[20] S. A. Akhondi, J. A. Kors, S. Muresan, Consistency of systematic chemical identifiers within and between small-molecule databases, Journal of Cheminformatics 4 (2012) 35. doi:10.1186/1758-2946-4-35.
[21] N. Reimers, I. Gurevych, Sentence-BERT: Sentence embeddings using Siamese BERT-networks, in: Proceedings of EMNLP 2019, Association for Computational Linguistics, 2019, pp. 3982–3992. doi:10.18653/v1/D19-1410.
[22] D. M. Reif, et al., Endocrine profiling and prioritization of environmental chemicals using ToxCast data, Environmental Health Perspectives 118 (2010) 1714–1720. doi:10.1289/ehp.1002180.
[23] A. J. Williams, S. Ekins, A quality alert and call for improved curation of public chemistry databases, Drug Discovery Today 16 (2011) 747–750. URL: https://www.sciencedirect.com/science/article/pii/S1359644611002406. doi:10.1016/j.drudis.2011.07.007.
[24] T. Berners-Lee, Linked data, https://www.w3.org/DesignIssues/LinkedData.html, 2006. W3C Design Issues.
[25] G. Van Haute, S. Goedertier, P. Fannes, M. Van de Wynckel, Chemical identity lost in regulation: code and data, https://doi.org/10.5281/zenodo.21512303, 2026. doi:10.5281/zenodo.21512303.