Chemical Identity Lost in Regulation

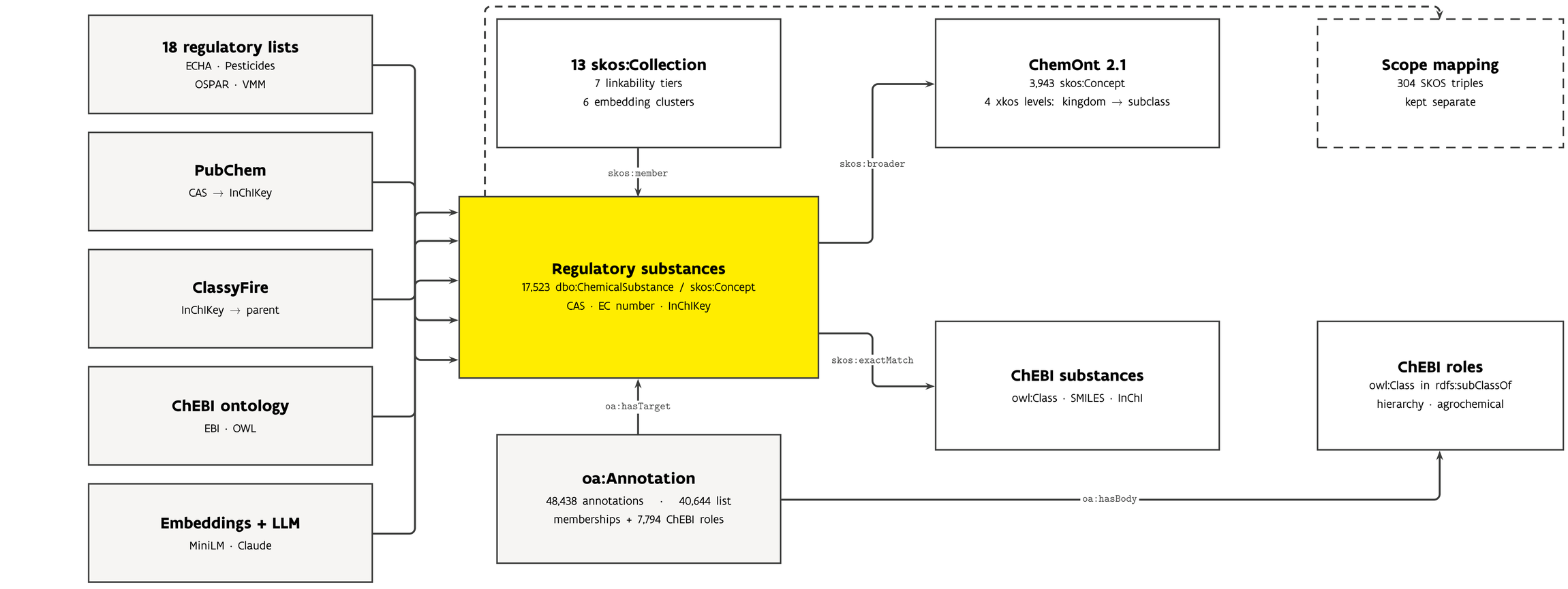
A Study of Semantic Interoperability in European Chemical Substance Data

Geert Van Haute **Stijn Goedertier** **Pieter Fannes** **Maxim Van de Wynckel**

Flanders State of the Art
Environment Department

One semantic model over 18 regulatory sources

solid arrows: asserted RDF relations · dashed: deliberately not merged

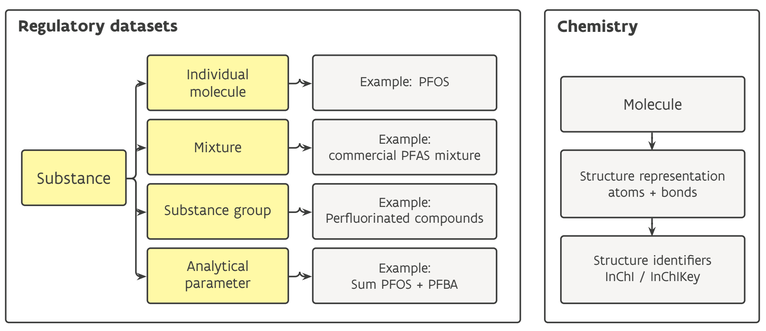


Regulatory substances form one SKOS concept scheme; ChemOnt supplies chemical classification over four XKOS levels, ChEBI an orthogonal axis of biological roles, and an annotation layer keeps regulatory provenance attached to every claim.

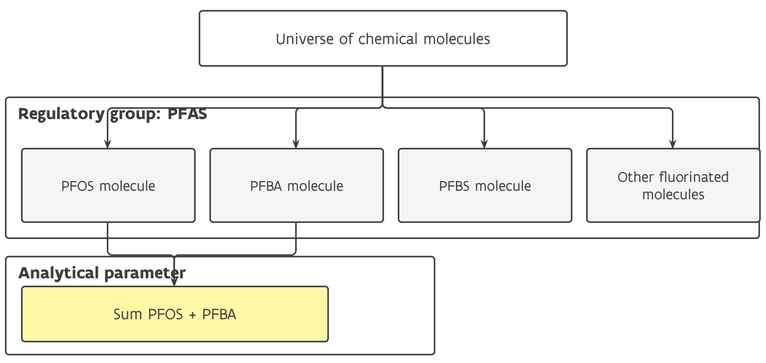
Why chemical identity breaks down

In chemistry, identity is exact: **InChI / InChIKey** link a molecule unambiguously across PubChem and ChEBI.

In EU regulation (**REACH**, **CLP**) a substance may be a molecule, a mixture or **UVCB**, a **group** such as PFAS with no set membership, or an analytical sum parameter – carrying only **CAS numbers** and names, which encode no structure.



Structure-based chemical identity versus the heterogeneous substance types found in regulatory datasets.



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

62.5% link to a defined structure

37.5% no structural identity

39.2% best LLM Hit@1

RQ1 Can substances be linked to a unique structure?
RQ2 How reliable is CAS across datasets?
RQ3 Can embeddings or LLMs compensate?

What the data shows

Linkability (RQ1). Of 33,452 unique entries, **62.5% (20,909)** link directly via InChIKey. The other **37.5%** resist structure-based linking: 11.6% unresolved CAS (tier 4), 10.8% mixtures or UVCBs (tier 5), 14.9% no usable identifier (tier 7).

CAS reliability (RQ2). **70.0%** of structure-resolved InChIKeys occur in a single source, so CAS cannot serve as a shared key. Ambiguity compounds this: 195 InChIKeys carry multiple CAS numbers (up to seven) and 4 CAS numbers resolve to two distinct InChIKeys – for example toluene diisocyanate, which maps to two positional isomers.

Embeddings & LLMs cannot fill the gap

Embeddings & LLM (RQ3). Embedding-based matching collapses: the best model reaches **Hit@1 = 5.75%**, the rest score below 1%. Claude Sonnet 4.6 reaches **39.2%** – a seven-fold gain, yet it still misclassifies the majority. Text is not a substitute for structure.

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**	–	–	–

ChemOnt classification, direct-parent level. Only Hit@1 is comparable across rows; ground truth is 28,204 ClassyFire assignments (structure-linked entries).

37.5%

of regulatory substance entries have no machine-readable structural identity.

What the model makes possible

Two axes, not one. ChemOnt classifies substances chemically (`skos:broader`); ChEBI says what they *do*. Both reach the same concept, so "every substance on an obligation list acting as a carcinogenic agent" is one query.

Provenance survives integration. **48,438** `oa:Annotation` nodes carry the cross-domain links (40,644 list memberships, 7,794 ChEBI roles) so 18 sources can disagree without overwriting one another.

A classification, not just a tree. Four `xkos:ClassificationLevel` steps (kingdom → subclass) make ChemOnt navigable at a chosen granularity.

Scope mapping, kept separate

For entries without a structure, an LLM assesses how a regulatory group relates to a ChemOnt node, yielding **304 validated SKOS triples** (285 `broadMatch`, 18 `exactMatch`, 1 `narrowMatch`). Regulatory groups are constrained *subsets* of a ChemOnt class, not equivalents.

It is published as a **separate graph**: machine-generated mappings do not enter the core until validated by experts.

Conclusions & recommendations

Over one-third of entries cannot be linked to a structure, CAS adds ambiguity, and neither embeddings nor LLMs replace structural information. Manual reconciliation (∼3,150 person-years for REACH) is infeasible. **Harmonisation must happen at the source.**

1 Require structure-based identifiers (InChIKey; MInChI for mixtures) across instruments.

2 Mandate explicit substance-group representation via ontologies such as ChemOnt.

3 Publish regulatory data as linked data with persistent URIs and SKOS links.

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)

doi.org/10.5281/zenodo.21512303