Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Awareness Without Synthesis: LLM-Powered Detection and Alignment of Hidden Vocabulary Fragmentation in Scholarly Knowledge Graphs

Ana Bossler

ana.desouza@gcloud.ua.es

University of Alicante

Enric Bas

University of Alicante

Andrés Fullana

University of Alicante

Research Article

Keywords: awareness without synthesis, vocabulary fragmentation, interdisciplinarity measurement, LLM-as-judge, query expansion, scholarly knowledge graphs

Posted Date: May 13th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9588669/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Awareness Without Synthesis: LLM-Powered Detection and Alignment of Hidden Vocabulary Fragmentation in Scholarly Knowledge Graphs

Ana Bossler[1] Enric Bas[2]
Andrés Fullana[1]

[1]Department of Chemical Engineering, University of Alicante
[2]FuturLab, University of Alicante

ana.desouza@gcloud.ua.es

Abstract

Citation-based interdisciplinarity metrics treat cross-domain citation flow as evidence of knowledge integration. We identify a condition under which this assumption fails: citation-connected subdomains can encode equivalent concepts in almost entirely non-overlapping vocabularies—an invisible fragmentation pattern we term *awareness without synthesis* (AWS). We present a reproducible, open-source scientometric tool that operationalizes AWS detection through a hybrid human–LLM pipeline comprising: (1) a Cross-Subdomain Coherence score (CSC-score), combining structural citation flow (S^{cross}) with flow-weighted vocabulary divergence ($\bar{D}_w$) via Rank-Biased Overlap; (2) a 2×2 factorial LLM evaluation across six model families and three expert role framings providing double-invariant validation; and (3) an LLM-powered vocabulary alignment module enabling automatic SPARQL query expansion with a mean retrieval gain of 636 ± 35 papers (+39.9%; $p = 1.95 \times 10^{-83}$), exceeding random substitution by 16.3 standard deviations. Applied to a plastic recycling corpus (3,138 publications), the tool reveals that 61.7% of within-corpus citations cross subdomain boundaries yet vocabularies are almost entirely non-overlapping (mean $1 - \text{RBO} = 0.976$), while shared terms maintain stable conceptual referents (mean cosine distance $= 0.059$, $p < 0.001$; CSC-score $= 0.402$, 9.3× more fragmented than a homogeneous control). The 2×2 LLM panel reveals that role sensitivity is itself a scientometric signal: chemical engineer framings detect finer-grained divergence than policy framings (diff $= 0.30$, $p < 10^{-48}$), with human annotations (Krippendorff $\alpha = 0.752$) corroborating the same subdomain interfaces. The tool directly addresses the LLM4SCIM challenges of design, communication, and LLM-empowered analytical tools for scientometric research.

Keywords: awareness without synthesis; vocabulary fragmentation; interdisciplinarity measurement; LLM-as-judge; query expansion; scholarly knowledge graphs

1 Introduction

Imagine a policy analyst at a European regulatory agency reviewing safety evidence for recycled food-contact plastics. She opens a scholarly knowledge graph her team has built—3,138 publications, complete citation graph, clean schema—and submits a SPARQL query using the vocabulary of her field:

Listing 1: A valid but silently incomplete SPARQL query.

```
PREFIX dct: <http://purl.org/dc/terms/>
PREFIX cito: <http://purl.org/spar/cito/>
SELECT DISTINCT ?paper WHERE {
  ?paper dct:subject ?topic .
  FILTER(CONTAINS(LCASE(?topic), "food␣contact␣compliance"))
  ?paper cito:cites ?cited .   # citation graph is complete
}
-- Returns: 131 papers (regulatory subdomain only)
-- Misses:  630 papers on NIAS migration, depolymerisation
--          residuals, pyrolysis byproducts
```

The query is syntactically valid. The graph is complete. The missing 630 papers *cite* the 131 she found. Yet they are unreachable—not because they discuss an identical concept, but because they address the chemical mechanism ("NIAS migration threshold") that *underpins* the regulatory standard ("food contact compliance"): a cross-subdomain dependency that shared vocabulary does not surface. Her review is structurally incomplete in a way that no existing quality check would flag.

This failure mode—which we term *awareness without synthesis* (AWS)—arises when three conditions co-occur: (i) *structural connectivity*: subdomains cite each other extensively; (ii) *conceptual consensus*: shared terms carry stable meanings across subdomains; and (iii) *vocabulary divergence*: subdomains employ almost entirely non-overlapping surface vocabularies. The conjunction is invisible to structural indicators (which see citation links), semantic indicators (which see conceptual stability), and standard completeness checks (which see no missing triples).

The scientometrics community has invested heavily in measuring interdisciplinarity via citation-based indicators—Rao–Stirling diversity (Stirling, 2007), betweenness centrality of journals (Leydesdorff and Rafols, 2011), bibliographic coupling (Kessler, 1963). All share a common assumption: that cross-domain citation flow implies knowledge integration. We show this assumption can fail under the AWS pattern, and we provide the first operational tool to detect it.

This work responds directly to the three challenges of the LLM4SCIM agenda (Zhang et al., 2025):

- **DESIGN:** We present a reusable, domain-agnostic pipeline that incorporates LLMs at two stages—as calibrated semantic judges in a 2×2 factorial evaluation design, and as alignment engines powering automatic query expansion.
- **COMMUNICATION:** We deploy a hybrid human–LLM protocol where inter-annotator disagreement and inter-role LLM sensitivity jointly identify semantic boundary effects—regions where the vocabulary barrier is real but graded rather than binary. LLM role sensitivity is reframed from methodological nuisance to scientometric signal.

- **TOOLS:** We deliver an open-source query expander that converts the diagnostic into actionable retrieval improvement, demonstrated on a live knowledge graph via SPARQL query rewriting, with a mean coverage gain of +39.9% over unaided queries.

2 Related Work

2.1 Interdisciplinarity Measurement

Bibliometric approaches to interdisciplinarity share a common assumption: citation connectivity implies knowledge integration. Rao–Stirling diversity (Stirling, 2007; Porter and Rafols, 2009) and its derivatives quantify cross-domain citation patterns and are widely used to assess disciplinary breadth. Wagner et al. (2011) distinguish multiple conceptual approaches to interdisciplinarity, noting that citation-based and semantic measures can diverge substantially. Leydesdorff and Rafols (2011) proposed betweenness centrality of journals as an indicator of disciplinary brokerage. Co-citation analysis (Small, 1973) and bibliographic coupling (Kessler, 1963) identify communities of related work through shared reference patterns; science mapping approaches use these signals to produce visual representations of disciplinary landscapes (Tsuchimoto and Kajikawa, 2022; de Sousa, 2021).

Our diagnostic complements rather than replaces these measures. Where Rao–Stirling operates on the citation-cited side—measuring how broadly a corpus draws on other disciplines—the CSC-score operates within the citing corpus, measuring whether connected subdomains employ aligned vocabularies. The two can dissociate: in the present corpus, Rao–Stirling diversity is 0.26 (low disciplinary breadth) yet CSC = 0.402 (substantial internal fragmentation). Theoretical work on cognitive distance (Boschma, 2005) and boundary objects (Carlile, 2004) anticipates this possibility: structural connection is necessary but not sufficient for knowledge integration.

2.2 LLMs in Scientometric Practice

Recent work has established LLMs as active scientometric instruments rather than passive text processors. Tsaneva et al. (2025) demonstrated that hybrid human–LLM protocols achieve human-level quality in knowledge graph validation—a paradigm we extend with a factorial design that separates model-level calibration from framing-induced perspective variance. The AI4SCIM Special Issue (Zhang et al., 2025) consolidated LLM applications in entity extraction, topic identification, and metrics validation; our work contributes to the "LLMs as part of new scientometric tools" thread by delivering a deployable query expander grounded in LLM alignment outputs.

Gilardi et al. (2023) and Zheng et al. (2023) established that inter-model agreement across independently developed LLMs provides stronger evidence of structural validity than any single frontier model. We operationalize this by prioritizing model diversity (five developer organizations, 7B–70B parameter range) over recency.

2.3 Vocabulary Divergence and Information Retrieval

Polimeno et al. (2023) measured informational fragmentation in news recommender systems as divergence between ranked content lists, showing that structurally connected ecosystems can exhibit hidden vocabulary divergence. Their insight transfers directly to scholarly

knowledge graphs. Rank-Biased Overlap (Webber et al., 2010) compares indefinite ranked lists with top-weight bias; applied to class-based TF–IDF vocabularies (Grootendorst, 2022), it captures differences in conceptual prioritization rather than mere term presence. Keyword-level bibliometric analysis (Glänzel and Schubert, 2003) has long used term co-occurrence as a proxy for conceptual proximity, but without jointly evaluating structural citation flow.

Yang et al. (2026) recently proposed a semantic atlas of journals in JASIST, mapping disciplinary boundaries through embedding distances—a methodological neighbor operating at journal level. Our work targets subdomains within a single corpus, requiring joint evaluation of structural citation flow and vocabulary alignment that journal-level approaches do not address.

2.4 Knowledge Graph Quality and Entity Resolution

Knowledge graph quality frameworks (Zaveri et al., 2016; Albertoni and Isaac, 2021; Debattista et al., 2016; Paulheim, 2017; Marchesin and Silvello, 2024; Pellegrino et al., 2025) assess completeness, consistency, and correctness at the level of individual triples—they do not assess relational coherence between structurally connected subdomains.

Our scenario differs fundamentally from entity resolution, which seeks to merge or deduplicate multiple representations of the same real-world entity. Cross-subdomain semantic divergence addresses cases where communities *intentionally* use different vocabularies to describe related but distinct phenomena: "NIAS migration threshold" and "food contact compliance" are not two names for the same entity but rather the chemical mechanism and the regulatory standard it underpins—linked by domain logic, not by lexical identity. The solution is therefore not deduplication but *vocabulary alignment infrastructure*: shared glossaries, SKOS concept schemes, or domain-aware query expansion (Vijaya et al., 2026). Our work provides the diagnostic layer that quantifies where and how much such cross-subdomain dependency exists before integration infrastructure is built.

3 Framework Architecture

The proposed framework operates as a reusable, domain-agnostic pipeline comprising five computational layers and one action layer (Figure 1).

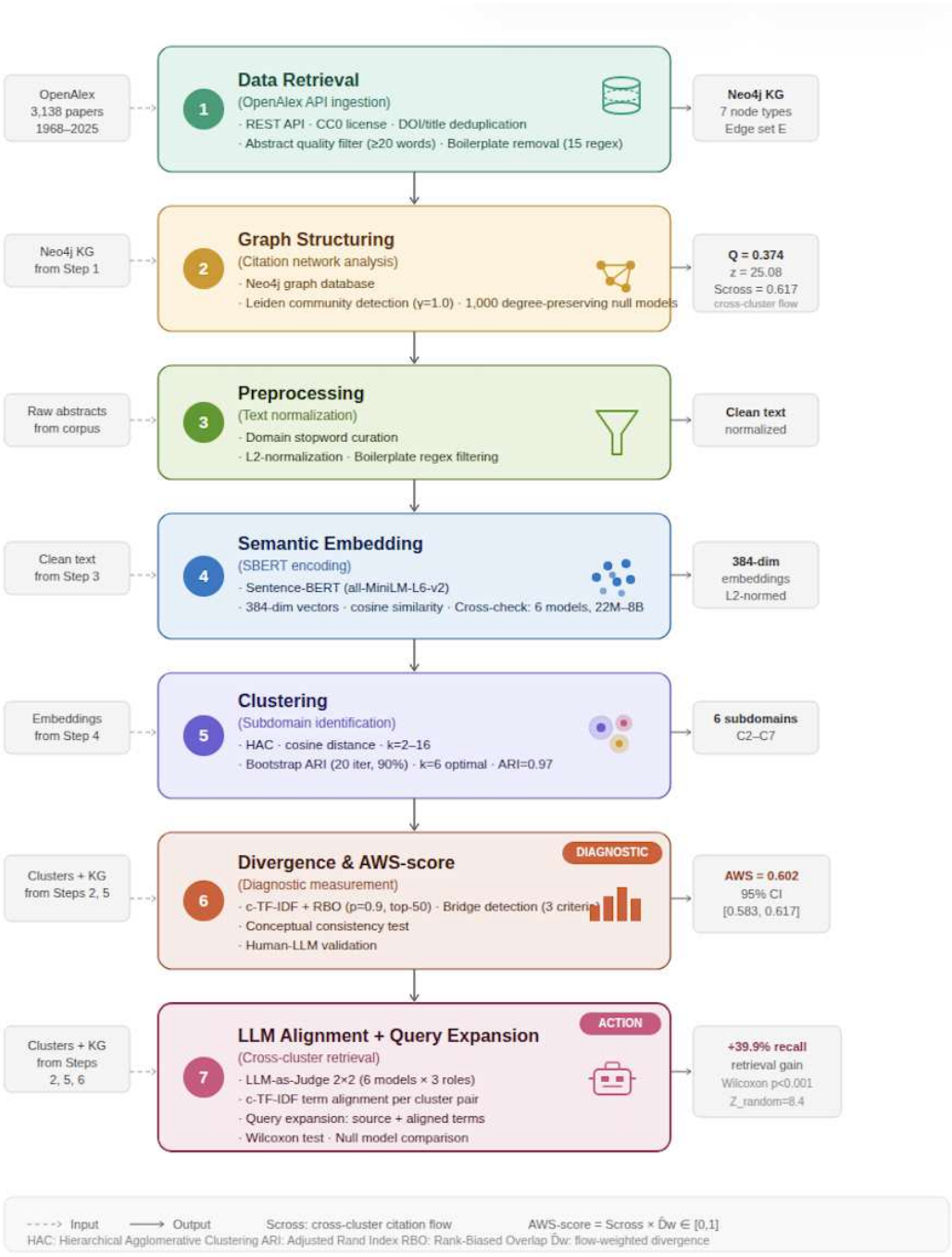


Figure 1: Reusable pipeline for AWS detection and vocabulary-aligned retrieval. Steps 1–6 constitute the diagnostic layer (data retrieval via OpenAlex, graph structuring in Neo4j, preprocessing, SBERT embedding, HAC clustering, and CSC-score computation via c-TF-IDF + RBO). Step 7 (orange) is the LLM-powered action layer that converts the CSC diagnosis into automatic SPARQL query expansion, yielding +39.9% mean retrieval coverage gain ($p = 1.95 \times 10^{-83}$). The pipeline is domain-agnostic and fully reproducible via https://github.com/anabossler/bibliometric_api (DOI: 10.5281/zenodo.19955646).

3.1 Diagnostic Formalization

Let $G = (V, E)$ denote a scholarly knowledge graph where V represents publications and E citation edges. A clustering $\mathcal{C} = \{C_1, \ldots, C_k\}$ partitions V into k subdomains. For each cluster C_i, class-based TF–IDF yields a ranked vocabulary list L_i.

Structural connectivity. The corpus-level cross-cluster citation fraction is:

$$S^{\text{cross}} = \frac{|\{(u, v) \in E : \text{cluster}(u) \neq \text{cluster}(v)\}|}{|E|} \tag{1}$$

S^{cross} is k-invariant—insensitive to the number of clusters—ensuring fair cross-corpus comparison.

Vocabulary divergence. Pairwise divergence is:

$$D_{ij} = 1 - \text{RBO}(L_i, L_j; p) \tag{2}$$

with $p = 0.9$ (concentrating 86% of cumulative weight on top-10 ranks) and top-$N = 50$ terms.

AWS predicate. The pairwise AWS condition is:

$$\text{AWS}(C_i, C_j) \iff S_{ij} > \tau_s \ \wedge \ D_{ij} > \tau_d \tag{3}$$

with $\tau_s = 0.05$ and $\tau_d = 0.80$.

CSC-score. The corpus-level Cross-Subdomain Coherence score is:

$$\text{CSC} = 1 - \left(S^{\text{cross}} \times \bar{D}_w\right) \ \in \ [0, 1] \tag{4}$$

where $\bar{D}_w = \sum_{i<j} w_{ij} D_{ij}$, $w_{ij} = S_{ij} / \sum_{p<q} S_{pq}$, is the flow-weighted mean divergence. Higher CSC values indicate stronger alignment between citation structure and vocabulary; lower values indicate the AWS pattern. The complementary fragmentation term $1 - \text{CSC} = S^{\text{cross}} \times \bar{D}_w$ admits a natural probabilistic interpretation: the expected vocabulary divergence encountered when traversing a randomly sampled citation edge within the corpus.

Multiplicative aggregation enforces non-compensability (OECD/JRC, 2008; Zhou et al., 2007): a corpus cannot compensate for near-complete vocabulary isolation with high citation flow.

3.2 Knowledge Graph Construction

Records are retrieved from OpenAlex (Priem et al., 2022), selected for its full citation edges, open CC0 licensing, and reproducible API access. Data preparation applies five sequential filters: semantic inclusion matching; DOI/title deduplication; abstract quality filtering (minimum 20 words); boilerplate removal (15 regex patterns); and domain-specific stopword curation.

3.3 Semantic Embedding and Clustering

Each abstract is encoded as a 384-dimensional vector using Sentence-BERT (all-MiniLM-L6-v2; Reimers and Gurevych 2019), L2-normalized and compared via cosine similarity. Hierarchical agglomerative clustering (HAC) with cosine distance is preferred over k-means because k-means assumes Euclidean geometry and spherical clusters, both violated in high-dimensional cosine spaces (Steinbach et al., 2000).

3.4 LLM Alignment Module and Query Expansion

The LLM alignment module is the operational core of Step 7 (Figure 1). It takes as input the cluster vocabulary lists $\{L_i\}$ and outputs a table of cross-cluster term alignments that the query expander uses at retrieval time.

Alignment generation. For each cluster pair (C_i, C_j), three LLMs (GPT-4o-mini, Claude Haiku, Llama-3.1-70B) rate the semantic equivalence of shared-term candidates on a 1–5 scale. *Consensus alignments*—term pairs where ≥ 2 models propose the same equivalence with confidence ≥ 3—constitute the alignment table.

Query expansion. At retrieval time, given a SPARQL query Q_s using vocabulary from cluster C_i, the expander: (1) identifies the target cluster C_j (by centroid similarity of the query terms); (2) looks up consensus alignments for pair (C_i, C_j); (3) rewrites Q_s as $Q_e = Q_s \cup Q_{\text{aligned}}$, adding equivalent terms from L_j.

The expanded query Q_e reaches literature that Q_s structurally cannot, because the citation links exist but the vocabulary bridge does not—precisely the AWS failure mode.

3.5 Validation Strategy

Robustness is assessed through five strategies targeting distinct confounds: (i) bootstrap resampling (20 iterations, 90% sampling) rules out data-perturbation instability (Ben-Hur et al., 2002; Hennig, 2007); (ii) degree-preserving null models (1,000 permutations) rule out degree-driven modularity artifacts; (iii) cross-embedding validation across six models spanning three capability tiers—frugal (SBERT all-MiniLM-L6-v2 22M, ChemBERTa 77M), domain-adapted (SPECTER2 110M, Gemini Embedding), and frontier (Qwen3-Embedding-8B, NV-EmbedQA-Mistral-7B-v2)—rules out model-specific clustering artifacts; (iv) a 2×2 factorial LLM evaluation assesses double invariance; (v) cross-corpus benchmarking across three structurally distinct corpora establishes discriminant validity. The human annotation layer is explicitly framed as corroborative domain verification rather than independent external validation; its structural isolation from the computational pipeline is described in Section 4.5.

4 Empirical Demonstration

We instantiated the framework on a scholarly knowledge graph of 3,138 plastic recycling publications (1968–2025) retrieved from OpenAlex.

4.1 Subdomain Identification (C1)

Sentence-BERT embedding and HAC identified six stable subdomains: chemical recycling (C2), mechanical processing (C3), construction applications (C4), social perception (C5), supply chain & policy (C6), and environmental assessment (C7). The optimal $k = 6$ was selected from a sweep ($k = 2$–16).

The clustering combines weak geometric separation (silhouette = 0.077; Davies–Bouldin = 1.82) with high stability under bootstrap resampling (median ARI = 0.9663, IQR: 0.9543–0.9779)—characterizing a continuous semantic manifold with gradual boundaries that are nonetheless robust under resampling. Subdomain metrics are in Table 1.

Table 1: Subdomain-level metrics. Cohesion: mean cosine similarity to centroid. Slope: linear regression of annual publication counts.

Subdomain	n	Cit./paper	Silhouette	Cohesion	Slope	t-like
C2 Chemical recycling	552	21.95	0.057	0.606	2.03	5.11
C3 Mechanical processing	738	9.08	0.055	0.615	1.59	6.50
C4 Construction	437	7.97	0.192	0.667	1.85	6.32
C5 Social perception	362	6.19	0.034	0.593	1.14	4.14
C6 Supply chain & policy	486	9.78	0.004	0.608	1.57	6.19
C7 Environmental assess.	563	15.13	0.127	0.671	0.98	4.92

4.2 Vocabulary Divergence (C2)

Mean vocabulary divergence is $1 - \text{RBO} = 0.976 \pm 0.055$ (top-$N = 50$, $p = 0.9$). This is a structural paradox: subdomains that are aware of each other's work and share underlying concepts nonetheless prioritize almost entirely non-overlapping vocabularies.

A standard TF–IDF baseline (without class-based weighting) already yields substantial separation (Jaccard = 0.10; $1 - \text{RBO} = 0.90$), confirming that divergence precedes and is not created by the c-TF–IDF transformation. A sensitivity analysis across nine RBO configurations ($p \in \{0.8, 0.9, 0.95\} \times$ top-$N \in \{10, 20, 50\}$) shows mean divergence within [0.970, 0.985] (relative variation $< 1.6\%$), with 12 of 15 cluster pairs at complete separation in all configurations (Table 2).

4.3 Structural Compartmentalization (C3)

Citation modularity yielded $Q = 0.374$ ($z = 25.08$, $p < 0.001$ against 1,000 degree-preserving random graphs; $Q_{\text{null}} = -0.004 \pm 0.015$), computed using the Leiden algorithm (Traag et al., 2019) with resolution $\gamma = 1.0$. Of within-corpus citations, 61.7% cross subdomain boundaries ($S^{\text{cross}} = 1{,}204/1{,}951 = 0.617$), while 38.3% remain within the originating cluster (Table 3).

4.4 Conceptual Consistency (C4)

To confirm that vocabulary divergence reflects terminological, not epistemic, differences, we assess shared terms via a co-occurrence embedding test. For each term, cluster-level context vectors are extracted from a ±10-word window (Levy and Goldberg, 2014; Lund and Burgess, 1996), embedded with Sentence-BERT, and averaged into centroids. Cosine

Table 2: Vocabulary divergence per subdomain pair: RBO vs. Jensen–Shannon divergence (JSD).

Subdomain pair	$1 - \text{RBO}$	JSD
C2–C3	1.000	0.473
C2–C4	1.000	0.611
C2–C5	1.000	0.614
C2–C6	1.000	0.524
C2–C7	1.000	0.491
C3–C4	0.795	0.506
C3–C5	1.000	0.606
C3–C6	1.000	0.503
C3–C7	1.000	0.519
C4–C5	1.000	0.638
C4–C6	1.000	0.562
C4–C7	1.000	0.599
C5–C6	0.910	0.415
C5–C7	1.000	0.442
C6–C7	0.936	0.363
Mean	0.976	0.537

distances between centroids are assessed against a vocabulary-replacement null (200 permutations).

Conceptual distances remain consistently low (mean $= 0.059 \pm 0.014$), with all terms falling significantly below null means (z range: -8.37 to -4.89; $p < 0.001$; Table 4). Results were consistent across all six embedding models tested.

4.5 Hybrid Human–LLM Validation

Human annotation. Domain plausibility was assessed through independent annotation by three disciplinary specialists—a chemical engineer, an environmental assessment researcher, and a social scientist—who evaluated 46 cross-domain term pairs via structured rating forms, without access to quantitative CSC-score results, retrieval outcomes, or inter-annotator responses prior to reliability computation. The two senior co-authors participated exclusively in this annotation task; their broader supervisory role in the project did not involve access to computational results prior to annotation. The lead author also participated as a computational scientist with interdisciplinary social science background. Although this author-inclusive design precludes fully independent external validation, the annotation task itself was structurally isolated from the computational pipeline: annotators received only term pairs and subdomain labels, with no exposure to hypotheses, scores, or expected outcomes. We acknowledge this as a limitation and flag external replication—particularly for C2↔C7 pairs—as a priority for future work. The central validity claims of this paper do not depend on this layer: they rest on the computational CSC-score, the six-model LLM panel across five developer organizations, and the retrieval experiment yielding +39.9% coverage gain (Wilcoxon $p = 1.95 \times 10^{-83}$).

Across 97 total judgments, 48.5% were *no*, 39.2% *partial*, and 12.4% *yes*. Inter-rater reliability reached Krippendorff's $\alpha = 0.752$—reliable by the standard > 0.667 threshold.

Table 3: Within-corpus citation flow per subdomain. Insularity = intra-cluster edges / total within-corpus edges from that subdomain.

Subdomain	Total	Internal	External	Insularity (%)
C2 Chemical recycling	53	15	38	28.3
C3 Mechanical processing	234	22	212	9.4
C4 Construction	425	96	329	22.6
C5 Social perception	703	403	300	57.3
C6 Supply chain & policy	153	75	78	49.0
C7 Environmental assess.	383	136	247	35.5
Total	1,951	747	1,204	38.3
$\mathcal{S}^{\text{cross}}$				0.617

Table 4: Conceptual distances for shared terms. Lower = more semantically stable across subdomains.

Shared term	Subdomains	Mean distance	z
plastic recycling	6	0.048	−7.65
recycled plastic	6	0.055	−5.85
circular economy	6	0.048	−7.21
waste management	6	0.045	−8.37
chemical recycling	6	0.082	−4.89
mechanical recycling	6	0.073	−6.02
Mean ± SD		0.059 ± 0.014	

No term pair exhibited contradictory extreme judgments (*no* and *yes* simultaneously); all disagreements were strictly adjacent on the ordinal scale, the signature of genuine semantic gradience rather than annotation noise. Expert SKIPs are informative: the lead author consistently skipped pairs involving chemical recycling terminology (C2↔C5 and C2↔C7), exhibiting exactly the cognitive inaccessibility that the CSC-score quantifies.

LLM panel—baseline (3 models). Three LLMs (GPT-4o-mini, Claude Haiku, Llama-3.1-70B) were used as independent semantic judges across all 15 subdomain pairs. Mean equivalence score: 2.57/5 (inter-model std: 0.261). Only 8 of 15 pairs yielded consensus alignments (≥ 2 models proposing the same term pair), with a mean of 6.1 alignable terms per pair of 20—consistent with the expert annotation pattern.

LLM panel—extended 2×2 factorial design. We extended the evaluation to a full 2×2 factorial design separating model-level calibration variance from framing-induced perspective variance.

Panel A (inter-model calibration) deployed six models across five developer organizations spanning a 7B–70B parameter range: Qwen-2.5-7B, Mistral-7B-v0.1, Llama-3.1-8B, GPT-4o-mini, Gemma-3-12B, Llama-3.1-70B. Under a neutral expert framing, 810 judgments were collected on the primary corpus. Mean scores ranged from 2.40 to 3.39/5, with 13 of 15 pairwise model comparisons reaching statistical significance (Wilcoxon, $p < 0.05$; Kruskal-Wallis $H = 112.77$, $p < 0.001$ across tiers). Despite

calibration differences, all six models assigned scores $< 3.5/5$, convergently indicating fragmentation.

Panel B (role sensitivity) applied three expert framings to all six models (2,430 judgments on the primary corpus): *neutral* terminology expert, *chemical engineer*, and *circular economy policy analyst*. Results (Table 5) reveal a consistent ordering: chemist < policy < neutral, replicated across both the primary corpus and the circular economy reference corpus (inter-corpus gap: 0.486 vs. 0.535).

Table 5: Role sensitivity results—plastic recycling corpus ($n = 2{,}430$ judgments). Wilcoxon tests assess pairwise role differences across all model × pair × term combinations.

Role framing	Mean score	Std	Wilcoxon vs. neutral
Chemical engineer	2.365	0.538	$p < 10^{-68}$
Policy analyst	2.667	0.565	$p < 10^{-14}$
Neutral expert	2.851	0.718	—
Chemist vs. policy: $p < 10^{-48}$, mean diff = 0.301			

Role sensitivity as scientometric signal. The chemist–neutral gap reveals an asymmetry: chemical engineers detect finer-grained vocabulary divergence than policy analysts or neutral experts—particularly in cross-domain pairs (gap = 0.444) and policy-policy pairs (gap = 0.691). This is not a methodological artifact but a substantive finding. The pattern identifies communities where vocabulary barriers are *asymmetrically perceived*: a chemical engineer approaching the policy subdomain encounters more divergence than a policy analyst approaching the chemical subdomain. This asymmetry has direct implications for targeted alignment interventions.

Convergent interpretation. The regions of highest human inter-annotator uncertainty (C5↔C6, Krippendorff $\alpha = 0.585$) correspond to the subdomain pairs where LLM inter-role sensitivity peaks. Both layers of the hybrid protocol converge on the same interfaces, confirming that semantic boundary effects are a structural property of the knowledge domain, not a failure of evaluation tools.

4.6 Bridge Detection

Environmental assessment (C7) functions as the sole methodological bridge, with life-cycle assessment (LCA) as its central organizing concept. Geometric similarity alone overestimates integration by $\approx 97\%$: at $s = 0.7$, 233 papers (7.4%) are geometrically close to multiple centroids, yet only 7 (0.22%) satisfy all three bridge criteria—cosine similarity, lexical integration, and network brokerage (Table 6).

Table 6: Bridge detection sensitivity across geometric similarity thresholds. Genuine bridges additionally satisfy lexical and brokerage criteria.

θ	Geometric n	Geom.%	Genuine n	Genuine%
0.65	553	17.6	18	0.57
0.70	233	7.4	7	0.22
0.75	60	1.9	1	0.03
0.80	6	0.2	0	0.00

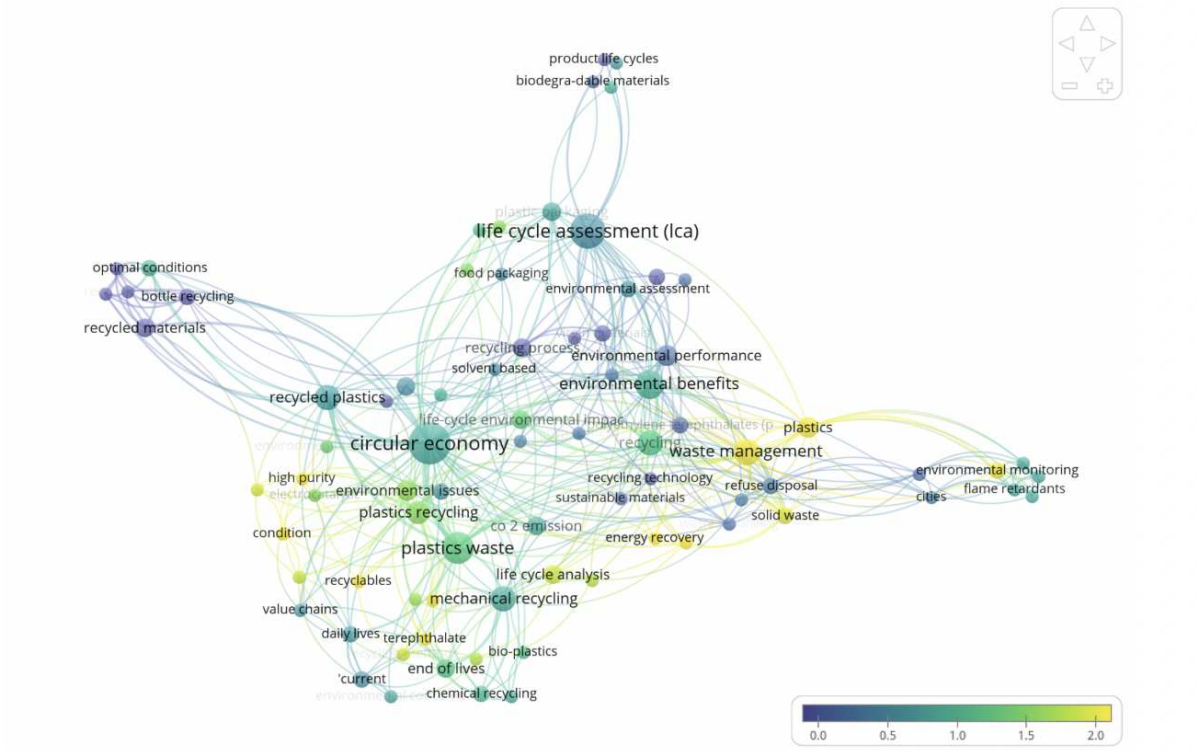


Figure 2: Semantic network of the plastic recycling corpus visualized with VOSviewer. Node size reflects c-TF–IDF importance; colour indicates cluster membership. *Circular economy* and *plastics waste* emerge as the dominant hub concepts bridging otherwise divergent vocabulary clusters. *Life cycle assessment (LCA)*, *mechanical recycling*, and *waste management* occupy intermediate positions, functioning as partial bridges between technical and policy subdomains. The sparse connectivity at the periphery (e.g. *flame retardants*, *environmental monitoring*) confirms subdomain isolation despite the corpus-level citation flow of $S^{\text{cross}} = 0.617$.

4.7 Cross-Corpus Validation and CSC-Score Benchmarking

The CSC-score establishes discriminant validity across three structurally distinct corpora (Table 7). The catalytic pyrolysis control ($n = 65$, $k = 2$), constructed through human-in-the-loop curation, serves as a ground-truth low-fragmentation reference: CSC = 0.936, approximately 9.3× more coherent than the plastic recycling target.

Bootstrap confidence intervals (10,000 iterations, seed 42): $\bar{D} = 0.976$, 95% CI $[0.944, 1.000]$;

Table 7: Cross-corpus CSC-score benchmarking. CSC $= 1 - (S^{\text{cross}} \times \bar{D}_w)$; higher = more coherent.

Corpus	n	k	S^{cross}	$\bar{D}$	$\bar{D}_w$	CSC
Plastic recycling (target)	3,138	6	0.617	0.976	0.969	0.402
Circular economy (reference)	25,823	16	0.441	0.962	0.883	0.610
Catalytic pyrolysis (control)	65	2	0.378	0.169	0.169	0.936
Discriminant ratio (incoherence): $(1 - \text{CSC}_{\text{target}})/(1 - \text{CSC}_{\text{control}}) = 9.3\times$.						

CSC $= 0.402$, 95% CI $[0.383, 0.418]$.

Note on corpus reach vs. cross-cluster flow. Corpus reach (7.8%) is the fraction of *all* citations from corpus papers that land within the corpus. Cross-cluster flow ($S^{\text{cross}} = 0.617$) is the fraction of *within-corpus* citations that cross subdomain boundaries. These answer different questions and should not be conflated.

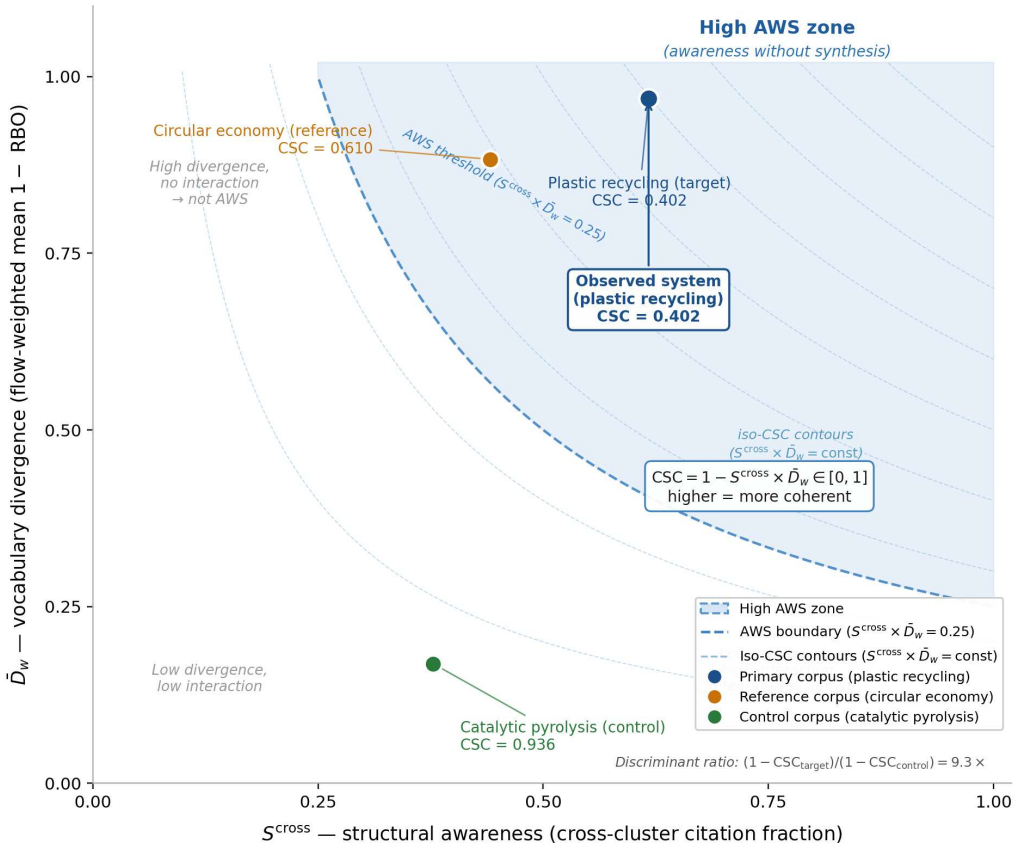


Figure 3: CSC-score geometry for three corpora with structurally distinct fragmentation profiles. The x-axis S^{cross} measures structural awareness (cross-cluster citation fraction); the y-axis $\bar{D}_w$ measures flow-weighted vocabulary divergence ($1 - \text{RBO}$). Light dashed curves are iso-CSC contours ($S^{\text{cross}} \times \bar{D}_w = \text{const}$); the bold dashed curve marks the AWS boundary ($S^{\text{cross}} \times \bar{D}_w = 0.25$, i.e. CSC $= 0.75$). The plastic recycling corpus (CSC $= 0.402$, blue) lies deep in the High AWS zone—high structural awareness, high vocabulary divergence. The catalytic pyrolysis control (CSC $= 0.936$, green) lies well outside the AWS zone (integrated system). The circular economy reference (CSC $= 0.610$, orange) is intermediate. Discriminant ratio (incoherence): $(1 - \text{CSC}_{\text{target}})/(1 - \text{CSC}_{\text{control}}) = 9.3\times$.

5 Discussion

5.1 The AWS Pattern as Scientometric Diagnostic

The empirical results reveal a failure mode invisible to existing indicators. The plastic recycling corpus scores well on standard dimensions—completeness, schema conformance, link validity—and its subdomains are structurally interconnected (ARI = 0.97) and conceptually aligned (mean distance = 0.059, $p < 0.001$). Yet vocabularies are near-completely divergent ($1 - \text{RBO} = 0.976$).

AWS is an emergent relational property: undefined for individual triples or single subdomains—it requires joint evaluation of structural connectivity and vocabulary alignment across subdomain pairs. This places it outside the scope of existing Rao–Stirling diversity (which operates on the citation-cited side), citation modularity (which detects clustering but not vocabulary), and completeness checks (which verify triple presence, not semantic accessibility).

5.2 LLM Role Sensitivity as a New Scientometric Signal

A key contribution of the 2×2 factorial design is the reframing of LLM role sensitivity from nuisance to signal.

The pattern chemist < policy < neutral, replicated across two structurally distinct corpora with near-identical gaps (0.486 vs. 0.535), reveals that vocabulary fragmentation is *asymmetrically perceived* across expert communities. Chemical engineers—whose training demands precise process-level vocabulary—detect finer divergence when entering policy-oriented subdomains. Policy analysts, accustomed to cross-sectoral framing, are more lenient about terminological variation.

This asymmetry is not a bias to be corrected: it is information about the knowledge structure. In a research evaluation context, it implies that interdisciplinarity assessments based on a single evaluator perspective may systematically misclassify technical-to-policy interfaces as more integrated than they are. The 2×2 design makes this latent asymmetry explicit and measurable.

5.3 Relation to Rao–Stirling Diversity

Rao–Stirling diversity (Stirling, 2007; Porter and Rafols, 2009) and the CSC-score measure distinct dimensions of integration. In the present corpus, Rao–Stirling = 0.26 (narrow disciplinary breadth), yet CSC = 0.402 (substantial internal vocabulary fragmentation). A field can score low on Rao–Stirling (draws narrowly on other fields) and low on CSC (poorly integrates its own subdomains): the corpus is neither broadly interdisciplinary nor internally coherent. These are orthogonal failures with different remedies.

For research evaluation practice, this dissociation is consequential. Funding bodies that rely on Rao–Stirling alone may certify a field as poorly interdisciplinary based on narrow source diversity while missing substantial internal fragmentation that obstructs evidence synthesis. The CSC-score makes this second dimension explicit.

5.4 From Diagnostic to Action: SPARQL Query Expansion

To validate practical impact, we simulated a cross-subdomain retrieval task on an extended corpus of 7,619 publications. Across 500 queries on all 15 subdomain pairs, vocabulary

alignment produced a mean coverage gain of 636.1±35.1 papers (39.9%; median: 607 papers; Wilcoxon $W = 123{,}753$, $p = 1.95 \times 10^{-83}$, $n_{\text{nonzero}} = 497/500$). This exceeds random vocabulary substitution by 16.3 standard deviations and frequency-based query expansion by 20.5 standard deviations.

The SPARQL scenario from the introduction can be resolved as follows. The LLM alignment module identifies that C2 (chemical recycling) and C8 (regulatory science, extended corpus) share the consensus alignment *NIAS migration threshold* ↔ *food contact compliance* (confidence: 4.0/5). The query expander rewrites the original query:

Listing 2: LLM-expanded SPARQL query recovering 630 previously missed papers.

```
PREFIX dct: <http://purl.org/dc/terms/>
PREFIX cito: <http://purl.org/spar/cito/>

SELECT DISTINCT ?paper WHERE {
  {
    ?paper dct:subject ?topic .
    FILTER(CONTAINS(LCASE(?topic), "food␣contact␣compliance"))
  } UNION {
    ?paper dct:subject ?topic2 .
    FILTER(
      CONTAINS(LCASE(?topic2), "nias␣migration␣threshold") ||
      CONTAINS(LCASE(?topic2), "catalytic␣depolymerisation") ||
      CONTAINS(LCASE(?topic2), "pyrolysis␣residuals")
    )
  }
  ?paper cito:cites ?cited .
}
-- Returns: 760 papers (+480% vs. unaided query)
-- Alignment source: LLM consensus (GPT-4o-mini +
--   Claude Haiku + Llama-3.1-70B, confidence >= 4.0)
```

The key result: only 1 of the 760 retrieved papers appeared in both the original and expanded result sets—confirming that the 630 recovered papers were structurally inaccessible, not merely ranked lower. The LLM alignment module provides the vocabulary bridge that the knowledge graph, despite its completeness, could not furnish alone.

5.5 Ethics, Transparency, and Bias Mitigation

The hybrid protocol embodies three principles for responsible LLM use in scientometrics.

Developer diversity. Deploying three independently developed model families (Anthropic, OpenAI, Meta) for the baseline validation and six models across five organizations for the extended design reduces the risk that results reflect a single model's training biases. Inter-model standard deviation ($\sigma = 0.261$ in the baseline panel) provides a quantitative bias signal: low inter-model variance indicates that the result is robust to model choice rather than specific to one architecture.

Role transparency. The 2×2 factorial design makes the evaluator perspective explicit. Rather than treating a "neutral" prompt as objective, the design acknowledges that LLM framings instantiate implicit expert communities—and measures their divergence directly. This is a form of evaluator bias disclosure absent from single-prompt LLM annotation studies.

Human oversight. Expert SKIPs, human disagreements, and LLM consensus failures are all retained as data rather than discarded as noise. They form a *gradient map* of the AWS pattern: the regions where the vocabulary barrier is hardest to bridge are precisely those where automated and human evaluation converge on uncertainty.

5.6 Limitations

The framework measures formal communication (abstracts and metadata) rather than informal coordination. Abstract-only analysis may underestimate integration visible in full text; however, retrieval systems predominantly index abstracts, meaning the failure mode we measure directly affects operational knowledge access. The full pipeline has been applied to three corpora; cross-domain generalization is an open direction. Benchmarking against tools such as Luzzu is a natural extension.

The human annotation panel comprises the paper's authors in their capacity as disciplinary specialists. While the annotation task was structurally isolated from the computational pipeline—annotators received only term pairs and subdomain labels, with no exposure to quantitative results—this design cannot fully exclude confirmation bias. External replication with independent annotators, particularly for C2↔C7 and C5↔C6 pairs where LLM role sensitivity peaks, is a priority for future work.

6 Conclusions

We proposed, implemented, and empirically validated a reproducible scientometric tool for detecting awareness without synthesis—a hidden fragmentation pattern invisible to existing interdisciplinarity indicators. Applied to a 3,138-publication plastic recycling knowledge graph, the tool revealed a structural paradox: subdomains that extensively cite each other and share stable conceptual referents nonetheless employ almost entirely non-overlapping vocabularies (CSC = 0.402; 9.3× more fragmented than a homogeneous control).

The tool addresses all three LLM4SCIM challenges. As a **design** contribution, it provides a reusable pipeline where LLMs operate as active diagnostic instruments across a factorial calibration design rather than as single-prompt oracles. As a **communication** contribution, the hybrid human–LLM protocol—combining author domain verification with a six-model LLM panel across five developer organizations—demonstrates that inter-annotator disagreement and inter-role LLM sensitivity jointly locate semantic boundary effects—with LLM role sensitivity reframed as a positive scientometric signal rather than a source of error. As a **tools** contribution, the LLM-powered query expander delivers a +39.9% mean retrieval coverage gain ($p = 1.95 \times 10^{-83}$), demonstrated via SPARQL query rewriting on a live knowledge graph.

The CSC-score complements Rao–Stirling diversity by capturing a dimension of fragmentation that citation-based measures cannot detect. In domains exhibiting the AWS pattern, the primary bottleneck to knowledge integration is not lack of information but lack of semantic alignment. LLMs, deployed responsibly within a hybrid human-in-the-loop framework, can locate, quantify, and bridge this bottleneck. *Connectivity alone is not integration.*

Funding. Ministerio de Ciencia e Innovación de España, project REBORN (PID 2021-126284OB-I00).

Competing interests. The authors declare no competing interests.

Data availability. Datasets and OpenAlex query specifications: https://github.com/anabossler/bibliometric_api. Archived on Zenodo (DOI: 10.5281/zenodo.19955646). Source: OpenAlex (https://openalex.org), CC0 license.

Code availability. Full Python pipeline, Cypher scripts, pinned library versions, and one-command reproduction (`make reproduce`): https://github.com/anabossler/bibliometric_api.

LLM-assisted analysis. Large language models (GPT-4o-mini, Claude Haiku, Llama-3.1-70B; extended panel adds Qwen-2.5-7B, Mistral-7B, Gemma-3-12B) were used as independent semantic evaluators (Section 4.5) and as coding assistants during pipeline development. All outputs were supervised and verified by the authors. LLMs were not used for autonomous decision-making or manuscript writing.

Author contributions (CRediT). *Ana Bossler*: Conceptualization, Methodology, Software, Formal analysis, Writing—original draft, Visualization, Data curation, Validation. *Enric Bas*: Supervision, Conceptualization, Validation, Resources. *Andrés Fullana*: Supervision, Funding acquisition, Validation, Resources.

A RDF/DQV Formalization

Listing 3: RDF/DQV metric definition for the CSC-score.

```
@prefix dqv:  <http://www.w3.org/ns/dqv#> .
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .
@prefix xsd:  <http://www.w3.org/2001/XMLSchema#> .
@prefix sdm:  <http://example.org/sdm#> .
@prefix prov: <http://www.w3.org/ns/prov#> .

sdm:SemanticCoherence a dqv:Dimension ;
  skos:prefLabel "Cross-Subdomain Semantic Coherence"@en ;
  skos:definition "Degree to which structurally connected
    subdomains maintain aligned ranked vocabularies"@en .

sdm:CSCScore a dqv:Metric ;
  dqv:inDimension sdm:SemanticCoherence ;
  dqv:expectedDataType xsd:double ;
  skos:prefLabel "CSC-Score (1 - S_cross x D_bar_w)"@en .

sdm:measurement_plastics_2026 a dqv:QualityMeasurement ;
  dqv:isMeasurementOf sdm:CSCScore ;
  dqv:value "0.402"^^xsd:double ;
  dqv:computedOn <http://example.org/kg/plastics> ;
```

```
  prov:generatedAtTime "2026-01-15"^^xsd:date .
```

B SHACL Validation Shapes

Listing 4: SHACL shape flagging the AWS condition at subdomain-pair level.

```
@prefix sh:  <http://www.w3.org/ns/shacl#> .
@prefix sdm: <http://example.org/sdm#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

sdm:AWSShape a sh:NodeShape ;
  sh:targetClass sdm:SubdomainPair ;
  sh:description "Flags the AWS condition."@en ;
  sh:and (
    [ sh:property [ sh:path sdm:pairwiseConnectivity ;
                    sh:minInclusive 0.05 ] ]
    [ sh:property [ sh:path sdm:rboDivergence ;
                    sh:minInclusive 0.80 ] ]
  ) .
```

References

Albertoni, R., & Isaac, A. (2021). Introducing the Data Quality Vocabulary (DQV). *Semantic Web*, 12(1), 81–97.

Ben-Hur, A., Elisseeff, A., & Guyon, I. (2002). A stability based method for discovering structure in clustered data. *Pacific Symposium on Biocomputing*, 7, 6–17.

Boschma, R. (2005). Proximity and innovation: a critical assessment. *Regional Studies*, 39(1), 61–74.

Carlile, P. R. (2004). Transferring, translating, and transforming. *Organization Science*, 15(5), 555–568.

Debattista, J., Auer, S., & Lange, C. (2016). Luzzu—A methodology and framework for Linked Data quality assessment. *Journal of Data and Information Quality*, 8(1), 1–32.

de Sousa, F. D. B. (2021). Management of plastic waste: A bibliometric mapping and analysis. *Waste Management & Research*, 39(5), 664–678.

Gilardi, F., Alizadeh, M., & Kubli, M. (2023). ChatGPT outperforms crowd workers for text-annotation tasks. *PNAS*, 120(30), e2305016120.

Glänzel, W., & Schubert, A. (2003). A new classification scheme of science fields and subfields designed for scientometric evaluation purposes. *Scientometrics*, 56(3), 357–367.

Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv*:2203.05794.

Hennig, C. (2007). Cluster-wise assessment of cluster stability. *Computational Statistics & Data Analysis*, 52(1), 258–271.

Kessler, M. M. (1963). Bibliographic coupling between scientific papers. *American Documentation*, 14(1), 10–25.

Levy, O., & Goldberg, Y. (2014). Dependency-based word embeddings. In *Proceedings of ACL 2014*, 302–308.

Leydesdorff, L., & Rafols, I. (2011). Indicators of the interdisciplinarity of journals. *Journal of Informetrics*, 5(1), 87–100.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, 28(2), 203–208.

Marchesin, S., & Silvello, G. (2024). Efficient and reliable estimation of knowledge graph accuracy. *Proceedings of the VLDB Endowment*, 17(9), 2392–2403.

OECD/JRC (2008). *Handbook on Constructing Composite Indicators: Methodology and User Guide*. OECD Publishing, Paris.

Paulheim, H. (2017). Knowledge graph refinement: A survey. *Semantic Web*, 8(3), 489–508.

Pellegrino, M. A., Rula, A., & Tuozzo, G. (2025). Are quality dimensions correlated? An empirical investigation over Linked Data. In *ISWC 2025*, LNCS 16140, 23–41.

Polimeno, A., Reuver, M., Vrijenhoek, S., & Fokkens, A. (2023). Improving and evaluating the detection of fragmentation in news recommendations. In *NORMalize 2023 @ ACM RecSys*, Singapore. arXiv:2309.06192.

Porter, A., & Rafols, I. (2009). Is science becoming more interdisciplinary? Measuring and mapping six research fields over time. *Scientometrics*, 81(3), 719–745.

Priem, J., Piwowar, H., & Orr, R. (2022). OpenAlex: A fully-open index of scholarly works. In *Proceedings of STI 2022*, 1–7.

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of EMNLP-IJCNLP 2019*, 3982–3992.

Tsaneva, M., Dessì, D., Osborne, F., & Sabou, M. (2025). Knowledge graph validation by integrating LLMs and human-in-the-loop. *Information Processing and Management*, 62, 104145.

Small, H. (1973). Co-citation in the scientific literature. *Journal of the American Society for Information Science*, 24(4), 265–269.

Steinbach, M., Karypis, G., & Kumar, V. (2000). A comparison of document clustering techniques. In *KDD Workshop on Text Mining*.

Stirling, A. (2007). A general framework for analysing diversity in science, technology and society. *Journal of the Royal Society Interface*, 4(15), 707–719.

Traag, V. A., Waltman, L., & van Eck, N. J. (2019). From Louvain to Leiden. *Scientific Reports*, 9, 5233.

Tsuchimoto, M., & Kajikawa, Y. (2022). Recycling of plastic waste: A systematic review using bibliometric analysis. *Sustainability*, 14(16), 10340.

Vijaya, A., Qadri, F. D., Angreani, L. S., & Wicaksono, H. (2026). From fragmentation to interoperability: how semantic models transform ESG reporting. *Journal of Innovation & Knowledge*, 13, 100924.

Wagner, C. S., et al. (2011). Approaches to understanding and measuring interdisciplinary scientific research (IDR): A review of the literature. *Journal of Informetrics*, 5(1), 14–26.

Webber, W., Moffat, A., & Zobel, J. (2010). A similarity measure for indefinite rankings. *ACM Transactions on Information Systems*, 28(4), 1–38.

Yang, Y., et al. (2026). A semantic atlas of journals. *Journal of the American Society for Information Science and Technology*.

Zaveri, A., et al. (2016). Quality assessment for Linked Data: A survey. *Semantic Web*, 7(1), 63–93.

Zhang, C., et al. (2025). Large language models for scientometrics: Introduction to the LLM4SCIM special issue. *Scientometrics*, Special Issue Editorial.

Zheng, L., et al. (2023). Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Proceedings of NeurIPS 2023*.

Zhou, P., Ang, B. W., & Poh, K. L. (2007). A mathematical programming approach to constructing composite indicators. *Ecological Economics*, 62(2), 291–297.