DCPAPERS

DCMI 2026

Proceedings of the International Conference on Dublin Core and Metadata Applications 2026

ISSN: 1939-1366

# From Detection to Documentation: A Provenance-Aware Framework for Semantic Evolution

**Hyung Wook Choi** [1] and **Mat Kelly** [1]

[1] Drexel University

## Abstract

Semantic change presents a persistent challenge for knowledge organization, as evolving meanings complicate conceptual clarity, interoperability, and long-term metadata management. Computational approaches detect semantic drift through distributional models and diachronic embeddings. However, they largely treat change as a signal rather than a documented process, offering limited support for tracing how and why meanings evolve. Conversely, ontology and metadata frameworks emphasize structure and provenance but lack mechanisms to capture the dynamic, gradual nature of semantic evolution in unstructured terms. This paper proposes a provenance-aware framework—the Versioned Semantic Shift Ontology (VERSO)—that models semantic evolution as a version-controlled process, in which each semantic state of a term is represented as a versioned entity and changes are expressed through branching and revision operations that capture continuity, divergence, and convergence in meaning. The framework introduces a fuzziness-based measure to represent graded semantic membership, enabling the distinction between incremental shifts and substantial recontextualizations, and integrates this representation with W3C PROV-O provenance standards to document semantic change as a traceable, evidence-based process. A longitudinal case study of the term *network* (1816–2025), grounded in 163 Word2Vec and 50 BERT observations across a merged corpus of COHA, COCA, arXiv, IMSDB, and American Stories data, demonstrates how the framework captures three empirically distinct phases of semantic evolution—stable broadcasting sense (pre-1990), transitional coexistence with computing vocabulary (1990–2006), and sustained multi-sense plurality (post-2007)—as a queryable, version-controlled scholarly record, thereby bridging computational detection and metadata-based documentation while supporting improved interpretability, interoperability, and transparency in evolving knowledge systems.



## 1. Introduction

When new concepts emerge, scholars often borrow or adapt existing terms to introduce them into academic discourse. In the process, those terms accumulate new meanings, while other expressions may arise to capture overlapping ideas. Semantic shift refers to changes in word usage and meaning over time in linguistics [1]. Such shifts result from a variety of factors, including sociocultural influence, linguistic dynamics, communicative practices, and even translation across languages [2], [3]. Although these changes support the spread of ideas and innovation, they complicate the task of maintaining conceptual clarity, obscure the lineage of ideas, and hinder the ability of disciplines to track conceptual evolution accurately [4]. Distributional models, contextual embeddings, and diachronic word representations have been applied to corpora ranging from historical texts to social media, producing valuable insights into how meanings evolve [5], [6], [7], [8], [9]. These methods demonstrate that both gradual and abrupt shifts can be measured quantitatively—yet their focus remains largely on detection. They indicate that a word's neighborhood has changed but do not preserve the lineage of how or why such change occurred. Without contextual annotation, a semantic shift is

**Short Paper**

**Available** Aug 01, 2026

**Correspondence to**
Hyung Wook Choi
hc685@drexel.edu

reduced to vector displacement, obscuring the cultural, disciplinary, or epistemological rationales that drive linguistic innovation.

Tracking semantic evolution has therefore become a primary challenge in both linguistics and knowledge organization. Ontology and knowledge organization research have examined how structured vocabularies evolve [10], [11], [12], and studies on concept drift classify changes as abrupt, incremental, or gradual [13], while provenance models such as W3C PROV-O provide means to record entities, activities, and agents responsible for change. Ontology versioning tools [14] emphasize the necessity of systematically recording the development of knowledge structures in a transparent manner. Yet these studies typically address structured ontological entities rather than unstructured linguistic terms, leaving a gap between computational detection and archival documentation. A small number of efforts attempt to bridge these domains. [15] proposed concept signatures to capture semantic drift within ontologies, and [16] emphasized representing provisional or contested knowledge claims as structured records accommodating epistemic uncertainty. While these studies contributed to integration, they do not offer a systematic framework for treating semantic change as a version-controlled process. There is still a need for a framework that brings together computational techniques for identifying semantic shifts with ontology-based provenance mechanisms for recording them.

This paper addresses this gap by proposing a framework that treats semantic evolution as analogous to software development in a version control system. Each semantic state of a term is represented as a versioned entity, and shifts are represented as branches that can be updated or merged, reflecting continuity, divergence, or convergence in meaning. By embedding fuzziness scores and a provenance-aware membership concept, the framework ensures that each stage of semantic change is explicitly documented. This version-controlled approach transforms semantic shift from a detectable signal into a traceable archival record, enabling the reconstruction of conceptual trajectories across different contexts and times.

## 2. A Framework for Tracing Semantic Evolution with Fuzziness and Provenance

This framework adapts version control principles to the documentation of semantic evolution. Each semantic state of a term is a distinct versioned entity; updates mark incremental adjustments, and branches capture divergent developments across contexts. Rather than capturing only final outcomes, the model preserves the processes and interpretive decisions that shape semantic change, creating a lineage that reflects both continuity and divergence.

The foundation is the *fuzziness score*, which quantifies how strongly a term continues to belong to its original semantic category. Calculated from neighborhood overlap, positional change, and similarity reduction, it classifies shifts as stable, transitional, or provisional. High fuzziness values are documented as incremental revisions; low values warrant new branches that record distinct conceptual trajectories. This graded interpretation is consistent with theories of typicality and fuzzy inclusion [17], [18]. When scores fall near decision thresholds, changes are stored as provisional entities —"branches under review"—that remain visible until corroborating evidence warrants confirmation, revision, or merging, ensuring that ambiguity is documented rather than discarded.

Documentation is formalized using W3C PROV-O. Entities correspond to semantic states, activities to drift detection processes, and agents to the models or curators responsible. Lineage is captured through `prov:wasDerivedFrom` and `prov:wasRevisionOf`, with qualified derivations storing fuzziness scores, anchor sets, and neighbor turnover evidence. This ensures every update is archived with its evidential grounding and interpretive context.

To operationalize these principles, we introduce the **Versioned Semantic Shift Ontology (VERSO)** as a normalization layer between computational detection and provenance documentation. Rather than applying PROV-O directly to raw pipeline outputs, VERSO first abstracts detected shifts into stable domain-level objects: a `verso:LexicalItem` as a persistent term identifier, a `verso:SenseAnchor` encoding the historically attested definition and reference year, and a `verso:ShiftProfile` aggregating year-indexed observations into a coherent trajectory. This abstraction ensures that records from different pipelines or embedding architectures contribute to the same versioned lineage without conflation, and that the evidence of fuzziness underlying each versioning decision travels with the archival record.



## 3. Case Study: network (1816–2025)

The *network* case study demonstrates not only what the framework can detect, but how raw computational outputs are systematically transformed into version-controlled provenance records. The analysis pipeline produces two JSON files—one per embedding model—each containing a `provenance` block recording the model configuration and thresholds, a `context_anchor` block encoding the reference definition and anchor neighbors, and a sequence of `shift_moments`, each with a full metric set per year. VERSO maps these structures onto domain-level objects: the `provenance` block becomes a `verso:ModelRun` activity; the `context_anchor` becomes a `verso:SenseAnchor` linked to the OED broadcasting definition at the 1962 reference year; and each `shift_moment` becomes a `verso:YearObservation` entity carrying neighborhood overlap, positional change, fuzziness score, membership category, and uncertainty as explicit, queryable attributes. Critically, the raw fuzziness value is preserved as a primary metric rather than collapsed into categorical labels, so that future investigators can re-classify observations under different threshold conventions without losing the underlying evidence.

The PROV-O layer then connects these VERSO entities into a versioned lineage. The Word2Vec model run (`embedding_model: Word2Vec, alignment: Local Procrustes, threshold: 0.08/0.18`) and the BERT model run (`embedding_model:`

`BERT, alignment: centroid-based, threshold: 0.40/0.70`) are instantiated as separate `prov:Activity` instances, each linked via `prov:wasInformedBy` to a `verso:Finalization` activity that consolidates their outputs into a shared `verso:ShiftProfile`. The divergence in model-specific thresholds—which is necessary because W2V and BERT operate on different fuzziness scales—is preserved as provenance metadata rather than normalized away, making the modeling decision auditable.

Three versioning events emerge from this mapping. In the stable phase (pre-1990), all Word2Vec observations record overlap at or near 1.000 with fuzziness below 0.08; these are asserted as `prov:wasRevisionOf` extensions of the main trunk `ShiftProfile`, documenting incremental continuity with the 1962 anchor. In 1990, overlap drops to 0.533 and fuzziness rises to 0.173, crossing the transitional threshold: VERSO instantiates Branch A as a new `ShiftProfile` linked via `prov:wasRevisionOf` to the trunk, recording the emergence of computing collocates as a documented divergence rather than an overwrite. In 2007, overlap collapses from 0.533 to 0.050 in a single year while fuzziness crosses the provisional threshold (0.162→0.210); BERT independently corroborates the signal (overlap: 0.067). This convergent cross-model evidence triggers Branch B instantiation—a second `prov:wasRevisionOf` link recording the biological and neural sense divergence. By 2025, overlap reaches 0.017 and positional change 0.448, the largest semantic reorganization in the term's documented history, yet the full lineage from the 1962 anchor through Branches A and B remains intact and traversable in the provenance graph.

The result is a scholarly record that is meaningfully different from a detection report: each versioning decision is traceable to specific metric thresholds, each threshold is linked to a documented model configuration, and each model configuration is associated with a named agent and timestamp. A future investigator querying the SPARQL endpoint can reconstruct not only what the term *network* meant at any point between 1816 and 2025, but why each branch was instantiated, which model provided the warrant, and what the uncertainty was at the moment of the decision.



## 4. Conclusion

This paper proposed a version-controlled framework for documenting semantic evolution, bridging computational detection with ontology-based provenance through the VERSO normalization layer. By treating each semantic state as a versioned entity with explicit fuzziness scores, the model preserves continuity, ambiguity, and divergence as auditable scholarly records. The *network* case study demonstrates how three empirically distinct phases—stable sense consolidation, transitional coexistence, and polysemous plurality—are documented as a traceable lineage rather than discarded as analytical outputs.

The framework carries implications across information science and knowledge organization practices. For controlled vocabulary management, VERSO records provide structured, time-stamped evidence for terminology revision decisions—documenting not only that a term's meaning has shifted but when, at what rate, and with what degree of certainty. For ontology engineering, the PROV-O integration supplies the missing evidential grounding for structural updates, raising the epistemic standard for versioning decisions. More broadly, by aligning with W3C semantic web standards and producing SPARQL-queryable records, the framework supports the FAIR Principles for long-term scholarly infrastructure.

## References


[1] M. Vanhove, Ed., *From Polysemy to Semantic Change: Towards a Typology of Lexical Semantic Associations*. John Benjamins, 2008.

[2] J. Grzega and M. Schöner, "English and General Historical Lexicology: Materials for Onomasiology Seminars," Onomasiology Online, 2007.

[3] L. Vandevoorde, *Semantic Differences in Translation: Exploring the Field of Inchoativity*. Language Science Press, 2020.

[4] A. Fokkens, S. ter Braake, I. Maks, and D. Ceolin, *On the semantics of concept drift: Towards formal definitions of semantic change*. in Drift-a-LOD@EKAW. 2016.

[5] D. Fišer and N. Ljubešić, "Distributional modelling for semantic shift detection," *International Journal of Lexicography*, vol. 32, pp. 163–183, 2019.

[6] M. Martinc, P. K. Novak, and S. Pollak, "Leveraging contextual embeddings for detecting diachronic semantic shift," 2019.

[7] W. L. Hamilton, J. Leskovec, and D. Jurafsky, *Cultural shift or linguistic drift? Comparing two computational measures of semantic change*. in Proceedings of the Conference on Empirical Methods in Natural Language Processing. NIH Public Access, 2016, pp. 2116–2121.

[8] W. L. Hamilton, J. Leskovec, and D. Jurafsky, *Diachronic word embeddings reveal statistical laws of semantic change*. in Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. 2016, pp. 1489–1501.

[9] A. Kutuzov, L. Øvrelid, T. Szymanski, and E. Velldal, "Diachronic word embeddings and semantic shifts: A survey," 2018.

[10] N. F. Noy and M. Klein, "Ontology evolution: Not the same as schema evolution," *Knowledge and Information Systems*, vol. 6, pp. 428–440, 2004.

[11] E. Kontopoulos *et al.*, *Ontology-based representation of context of use in digital preservation*. in WHiSe@ESWC. 2016, pp. 65–72.

[12] T. G. Stavropoulos, S. Andreadis, M. Riga, E. Kontopoulos, P. Mitzias, and I. Kompatsiaris, *A framework for measuring semantic drift in ontologies*. in SEMANTiCS (Posters, Demos, SuCCESS). 2016.

[13] S. Wang, S. Schlobach, and M. Klein, "Concept drift and how to identify it," *Journal of Web Semantics*, vol. 9, pp. 247–265, 2011.

[14] P. Mutton and J. Golbeck, *Visualization of semantic metadata and ontologies*. in Proceedings of the Seventh International Conference on Information Visualization (IV 2003). IEEE, 2003, pp. 300–305.

[15] J. A. Gulla, G. Solskinnsbakk, P. Myrseth, V. Haderlein, and O. Cerrato, *Concept signatures and semantic drift*. in WEBIST 2010. Springer, 2010, pp. 101–113.

[16] S. Cristofaro, E. M. Sanfilippo, P. Sichera, and D. Spampinato, *Towards the representation of claims in ontologies for the digital humanities*. in SWODCH. 2021.

[17] S. Calegari and E. Sanchez, *A fuzzy ontology-approach to improve semantic information retrieval*, vol. 327. in URSW, vol. 327. 2007, pp. 1–6.

[18] J. A. Hampton, "Typicality, graded membership, and vagueness," *Cognitive Science*, vol. 31, pp. 355–384, 2007.