Making Low-Visibility Policies Discoverable: A Document-Embedding Approach to Inclusive Cross-Boundary Policy Comparison

Extended abstract for ICSC 2026 | Topic: Digital inclusion in the Global South

Abstract. Global policy knowledge is unevenly visible. Some policy texts are widely cited, indexed, and discussed internationally; others address similar governance problems but remain weakly represented, particularly in the Global South. This proof-of-concept study frames the issue as digital inclusion in policy knowledge systems: when comparison depends on citation visibility, relevant documents may be excluded from cross-boundary analysis. We test a document-embedding workflow that places policy texts from different visibility positions in a shared analytical space, making low-visibility documents more discoverable across boundaries. A corpus of 3,628 Chinese information technology policies is used as a methodological testbed, recovering stable thematic structure. We then examine a 24-document cross-boundary corpus on government information openness spanning high-visibility texts, regional frameworks, and low-visibility national practices. Low-visibility national practices show a within-tier mean semantic similarity of 0.53, matching high-visibility texts (0.53), while the high-to-low cross-tier mean reaches 0.48. This suggests that limited comparative visibility is not fully explained by content dissimilarity, and that structurally under-visible documents can be surfaced through document-embedding comparison. These findings show how document-embedding approaches can complement citation-based methods for more inclusive cross-boundary policy discovery.
Keywords: digital inclusion; document embeddings; policy discoverability; cross-boundary policy comparison

1 Introduction

Comparative policy research often depends on visible traces: formal citations, international organization references, indexed repositories, and repeated discussion in English-language literature. These signals shape which policy documents become easy to find and compare. Policies from countries or systems with weaker international indexing may remain outside the comparative frame even when they address widely shared governance problems. Visibility thus becomes a gatekeeper for comparison, creating an inclusion problem in global policy knowledge systems.

This extended abstract shifts the focus from policy influence to policy discoverability. As a proof of concept, we ask whether document-embedding representation can surface low-visibility policy texts substantively comparable to high-visibility ones. Digital inclusion here concerns not only access to digital infrastructure, but also the representation of under-visible policy documents in computational knowledge systems. We use embedding-based semantic similarity as a discovery device for identifying comparable documents that citation-based systems would not surface.

2 From Visibility to Discoverability

We distinguish three concepts. Citation visibility refers to the extent to which a policy text is represented through formal citations, international policy channels, indexing systems, or comparative academic discussion. Low-visibility policy documents are official policies that are substantively relevant but weakly represented in these visible channels. Embedding-based discoverability refers to locating such documents through vector-based similarity in goals, instruments, and governance problems rather than through citation links alone.

This framing builds on knowledge organization research arguing that what becomes findable depends on how knowledge is represented and organized [1], and on information equity scholarship showing that global knowledge circulation is structurally uneven [2]. Document embeddings offer a complementary representational strategy: they transform full policy texts into semantic vectors and construct a shared comparison space in which under-visible documents can be retrieved and examined alongside more visible texts [3, 4].

3 Data and Workflow

The workflow has two stages. First, we use a China information technology policy corpus of 3,628 documents from 1949 to 2024 as a methodological testbed. This corpus was assembled from iPolicy and Pkulaw through seed keywords, synonym expansion across fifteen domain-specific terms, title retrieval, deduplication, and double-blind expert screening for policy relevance and substantive information-technology content. This stage uses the GTE-based text-embedding-v3 model to test whether document-embedding representation can recover interpretable thematic structure within a large and coherent policy system.

Second, we extend one recovered theme, government information openness, into a 24-document cross-boundary corpus. The corpus includes seven high-visibility texts, four regional framework texts, nine low-visibility national-practice texts, and four China anchor texts. Visibility was coded before embedding based on observable traces of circulation, including appearance in English-language comparative policy literature, presence in international organization or regional policy channels, and inclusion in policy-indexing or policy-learning repositories. Because tier labels were assigned before similarity computation, visibility is not defined from the embedding results themselves. For

this cross-boundary corpus, documents were encoded with DashScope text-embedding-v4 as 1,024-dimensional semantic vectors, and pairwise cosine similarity was used to compare within-tier and cross-tier proximity and identify candidate high-low visibility policy pairs.

4 Preliminary Findings

The China corpus produces interpretable thematic structure across eight modules, including telecommunications regulation, ICT development, network security, industrial internet, and government information openness. We select government information openness, the largest module with 98 documents, as the bridge topic because it appears across institutional settings and connects public information access, open data reuse, administrative transparency, and public-sector data governance. Fig. 1 summarizes the workflow and preliminary cross-boundary results.

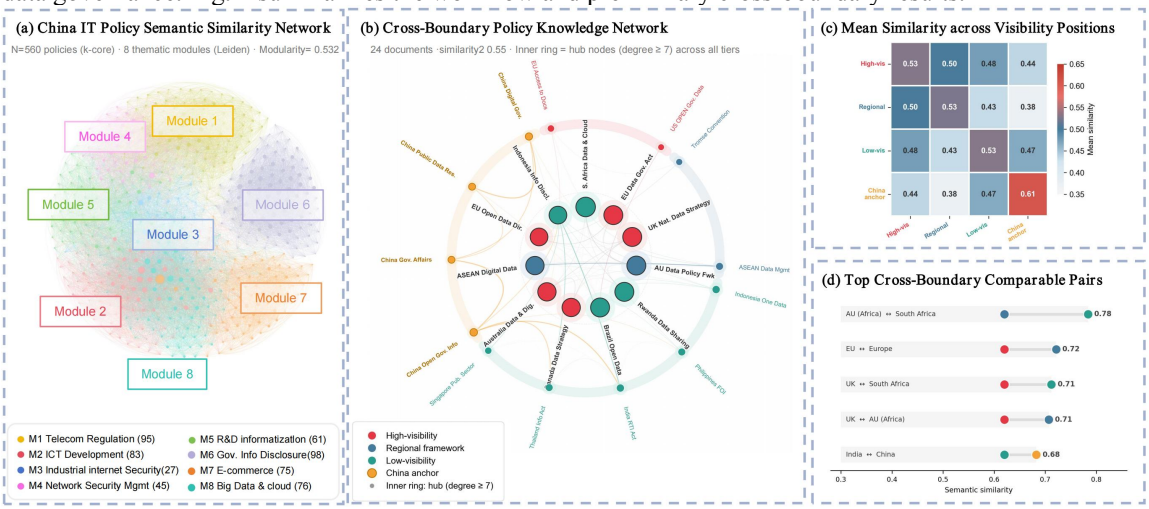

Fig. 1. Document-embedding workflow and cross-boundary similarity results.

As shown in Fig. 1, the 24 cross-boundary policies form a shared document-embedding comparison space across visibility positions, with high-visibility hub texts such as the EU Open Data Directive and UK National Data Strategy connected to low-visibility national practices including Rwanda Data Sharing and Brazil Open Data. The tier-level matrix shows that low-visibility national practices achieve a within-tier mean of 0.53, equal to the 0.53 observed among high-visibility texts, while the high-to-low cross-tier mean reaches 0.48. The top five comparable pairs range from 0.68 to 0.78, spanning Global South national practices and high-visibility frameworks around public information access and data reuse. The equal within-tier means suggest that low visibility is not simply a lack of semantic coherence. These scores identify low-visibility texts as substantively comparable cases for expert review and future qualitative validation.

5 Contribution and Next Steps

This proof-of-concept document-embedding approach reduces dependence on citation visibility and supports more inclusive cross-boundary policy discovery. It complements citation-based approaches by surfacing candidates beyond visible reference trails and shows how document embeddings and similarity networks can support a more inclusive knowledge infrastructure for policy research. The workflow is intended as a retrieval and prioritization layer before expert validation, not as an automated claim of policy influence.

Next steps will strengthen this proof of concept in three ways. First, we will expand the cross-boundary corpus to cover more Global South jurisdictions and policy domains. Second, we will validate visibility categories through systematic citation, indexing, and expert-review evidence. Third, we will test alternative embedding models and use policy-domain experts to assess high-similarity pairs, since embedding-based similarity identifies substantive proximity rather than causal influence, adoption, or diffusion.

References

1. Hjorland, B.: Fundamentals of knowledge organization. Knowledge Organization 30(2), 87-111 (2003).
2. Britz, J.J.: A moral reflection on the information flow from South to North: an African perspective. Libri 53(3), 160-174 (2003).
3. Wilkerson, J., Smith, D., Stramp, N.: Tracing the flow of policy ideas in legislatures: A text reuse approach. American Journal of Political Science 59(4), 943-956 (2015).
4. Li, Z., Zhang, X., Zhang, Y., Long, D., Xie, P., Zhang, M.: Towards general text embeddings with multi-stage contrastive learning. arXiv:2308.03281 (2023).