RESEARCH ARTICLE

From "Textual Interpretation" to "Cognitive Infrastructure": The Data-Driven Transformation and Paradigm Reconstruction in Country and Regional Studies for the Digital Age 一 With a Strategic Value Analysis of Multilingual Open Source INTelligence (OSINT) Data Annotation

Hang Dong[1],*
[1]*Shanghai Normal University, P.R. China*

Abstract: At a historical juncture when Artificial Intelligence Generated Content (AIGC) is profoundly reshaping global knowledge production and the landscape of geopolitical analysis, traditional area studies confront a dual crisis: the declining explanatory power of qualitative paradigms and a "knowledge supply crisis" mediated by algorithmic systems. This paper argues that area studies must undergo a paradigmatic transformation—from "linguistic mediation" to "semantic governance," and from tacit experiential knowledge to the construction of "semantic infrastructure." By conceptualizing high-quality data annotation as both a confirmation of semantic rights and the building of cognitive infrastructure, the study proposes an integrated framework encompassing ontology construction, semantic alignment, and domain-specific evaluation. Through this engineering-oriented architecture, unstructured regional knowledge can be transformed into structured data assets. Driven by a dual-engine model integrating "United Nations normative corpora" and "Open Source INTelligence (OSINT)," area studies can evolve from post hoc interpretation toward a decision-support paradigm characterized by real-time perception, structured computation, and scenario simulation. Such a transformation will consolidate the epistemic foundations of China's autonomous knowledge system and safeguard national cognitive sovereignty.

Keywords: country and regional studies, data annotation, semantic governance, international communication, new liberal arts, ESG, industry–education integration

1. Introduction: The "Legitimacy Crisis" in the Algorithmic Era and the Reconstruction of Disciplinary Niches

The intervention of Artificial Intelligence Generated Content (AIGC) signifies far more than a technological upgrade; it is fundamentally restructuring the established order of global knowledge production and decision-support systems. Against this backdrop, traditional area studies are experiencing an unprecedented paradigm crisis. Historically, the discipline's core competitiveness rested upon scholars' long-term embodied immersion in specific regions and their capacity for qualitative interpretation, thereby fulfilling its mission of facilitating cross-cultural academic dialogue.

However, as large language models (LLMs) push the efficiency of linguistic transformation and information retrieval to unprecedented levels, the "instrumental barriers" once sustained by foreign language expertise are rapidly eroding. This so-called legitimacy crisis is, in essence, the reconfiguration of epistemic authority under algorithmic mediation, whereby conventional interpretive authority appears increasingly inadequate in the face of intelligent decision-making systems.

More strikingly, this crisis has begun to surface within the macro-level logic of national disciplinary planning. For instance, the 2025 policy document Opinions on Strengthening the Development of Data-Element Disciplines and Digital Talent Cultivation assigns strategic urgency to fields such as digital finance and digital governance, while the relative absence of foreign language and area studies

Corresponding Author: Hang Dong
Shanghai Normal University, P.R. China

implicitly reflects their marginal position within the ongoing digital transformation.

This phenomenon signals a critical shift: within the framework of the national data strategy, linguistic proficiency or qualitative description detached from specific vertical industry scenarios is losing its status as an independent strategic data element (National Development and Reform Commission et al., 2025). To mitigate the risk of marginalization, area studies must undertake a paradigmatic shift—from a focus on linguistic form to semantic governance, and from linguistic mediation to cognitive infrastructure construction (State Council of the People's Republic of China, 2025).

The endogenous motivation for this reconstruction lies in the discipline's pressing need to overcome a longstanding "knowledge supply impasse." For decades, vast bodies of regional knowledge have remained embedded in unstructured, highly tacit textual forms. This condition has constrained knowledge reusability and erected barriers to inter-institutional collaboration.

From the perspective of computational social science, these high-value scholarly insights—lacking explicit structural features—have effectively become "dormant assets," excluded from the parameter space of intelligent decision systems and unable to participate in algorithmic computation or iterative optimization. This disciplinary dilemma—possessing insight without assetization—reveals a structural misalignment between traditional research paradigms and the data-driven era.

Accordingly, the urgent task for area studies is to translate its core interpretive competencies into computable, reusable, and interoperable digital assets through engineering methodologies. Through such paradigm reconstruction, the discipline will no longer remain a passive interpreter of global developments but will instead emerge as an active architect of cognitive frameworks for global regional analysis in the intelligent age. In doing so, it can become a strategic growth pole within China's autonomous knowledge system, contributing substantively to the digital economy and global governance.

2. Ontological Reconstruction: Data Annotation as Strategic "Cognitive Infrastructure"

To elucidate the symbiotic mechanism between data annotation and area studies, it is first necessary to dispel the long-standing misconception that reduces annotation to "digital piecework" or a mere "corpus assembly line." In the deep waters of digital-intelligent transformation, the ontological character of data annotation has fundamentally shifted. It is no longer peripheral labor at the lower end of the AI value chain; rather, it constitutes the core mechanism of semantic rights confirmation through which interpretive authority is exercised. More profoundly, it functions as the cognitive infrastructure underpinning the digital transformation of area studies. This paradigmatic shift—from instrumental utility to ontological centrality—provides the logical anchor for reconstructing the discipline's academic territory.

2.1 From labor processing to "semantic governance"

Within the iterative logic of large language models (LLMs), particularly during the stage of Reinforcement Learning from Human Feedback (RLHF), the nature of data annotation has undergone a structural transformation from quantitative accumulation to qualitative reconfiguration. Algorithmic learning has moved beyond basic keyword mapping toward deep alignment with complex reasoning patterns, affective tonality, and cultural positioning (School of Foreign Languages, East China Jiaotong University, 2025).

When addressing highly sensitive cross-linguistic texts involving geopolitical rivalry, multilateral institutions, or global governance, annotation ceases to be a neutral technical procedure. Each classificatory decision and interpretive calibration effectively shapes the normative orientation of AI systems. This practice of semantic governance directly determines how intelligent systems delineate factual boundaries and calibrate value hierarchies amid heterogeneous cultural conflicts.

2.2 Constructing the cognitive infrastructure of a digitally unified semantic order

Competition over governance in the digital realm ultimately reduces to competition over semantic definition. Those who define labeling taxonomies define the interpretive logic of AI worldviews. One strategic mission of area studies in the digital-intelligent era is thus to construct cognitive infrastructure that supports national decision-making and international communication. The core of such infrastructure lies in translating ambiguous and polysemous natural language into machine-executable and rule-governed logical architectures (Frontier Domestic Universities Teaching Reform Research Group, 2025).

Taking transnational legal governance and global compliance as empirical domains, traditional area studies often remain at the level of surface

translation and doctrinal commentary. By contrast, a cognitive-infrastructure paradigm emphasizes deep semantic deconstruction of legal texts—disaggregating abstract provisions into computable units such as responsible entities, triggering thresholds, jurisdictional boundaries, and remedial logics—while constructing cross-cultural and cross-linguistic mapping matrices.

Such millimeter-level semantic calibration and logical anchoring are becoming invisible yet decisive stakes in global governance. By forging high-granularity semantic infrastructure, area studies do more than populate databases; they construct a logically coherent and autonomously controllable digital discourse foundation. The ultimate objective is to counterbalance the dilution of interpretive authority under algorithmic mediation by mastering underlying semantic standards, thereby reclaiming and consolidating strategic interpretive sovereignty.

2.3 Knowledge assetization: resolving the "insight without asset" dilemma

Historically, the outputs of area studies have been embedded in unstructured research reports and print-based scholarship. While rich in geopolitical intuition and analytical acuity, such knowledge forms exhibit high dispersion and lack standardized logical interfaces. Consequently, they remain difficult to integrate into intelligent decision systems, becoming isolated "intellectual islands" resistant to reuse.

An ontological perspective provides the methodological blueprint for engineering and assetizing knowledge. In this framework, data annotation is reconstituted as the core craft of epistemic transformation—structurally extracting and logically reorganizing dispersed insights into algorithmically operable assets capable of entering iterative computational cycles.

When scholars' tacit knowledge is codified through scientific annotation schemas into structured datasets, ephemeral intellectual insights are transformed into infinitely reusable digital assets. This assetization enhances the discipline's capacity to serve national strategy while establishing a closed-loop mechanism of corpus–annotation–evaluation that enables self-evolution and dynamic value appreciation. Through this transition, area studies move from individualized interpretive practice toward clustered, engineering-based research grounded in semantic infrastructure—realizing a fundamental shift from textual exegesis to structured computation (Ng, 2021).

3. Pathway Reconstruction: A Trinity Framework Driven by Deep Knowledge Engineering

The ultimate mission of area studies is to serve national strategy. In an era when algorithmic systems dominate information distribution, the primary battleground of international communication and intelligence analysis has descended to the level of algorithmic black boxes and dataset architectures.

3.1 Correcting algorithmic bias and the "discourse deficit"

Mainstream large models are predominantly trained on English-language corpora, embedding implicit Western-centric value orientations and generating latent cultural hegemony and discourse deficits. Absent high-quality alignment data reflecting China's discursive framework, AI systems will reproduce Western normative logics by default.

Organizing area specialists to conduct large-scale, high-quality multilingual annotation thus constitutes a structural intervention to inject a Chinese epistemic perspective into AI systems. For instance, precise semantic annotation of core political concepts ensures accurate policy transmission. Such upstream semantic engineering represents the foundational project in contesting discursive authority in the intelligent age.

3.2 Cross-Linguistic semantic alignment: cognitive synchronization and logical anchoring

In multilingual environments, semantic alignment transcends translation; it requires precise cognitive synchronization. In global governance contexts, identical terms may carry divergent semantic boundaries across cultures—for example, differing conceptualizations of "the right to development" or "security." Semantic alignment aims to render such implicit differences explicit and rule-governed.

High-dimensional semantic mapping standards reduce algorithmic hallucination and misinterpretation when AI systems process sensitive transnational issues, thereby safeguarding analytical accuracy in national strategic assessment. The capacity to define semantics thus emerges not merely as a technical translation skill but as a foundational competitive advantage in global governance.

3.3 Professionalized benchmarks: a closed loop for dynamic knowledge appreciation

The systematic deep representation of area studies requires that research efficacy be transformed into quantifiable and falsifiable scientific metrics. This necessitates converting tacit expert intuition into algorithmically decodable benchmarks—such as a geopolitical stance-classification matrix for the

Middle East or a regulatory-text parsing benchmark for Southeast Asia.

These evaluation sets should not be treated as static examination tools but as engineering engines driving iterative cycles of evaluation, diagnosis, replenishment, and evolution. Through high-frequency stress testing, semantic drift and logical blind spots can be precisely identified and corrected via targeted data augmentation. In doing so, disciplinary knowledge assets enter a spiral trajectory of value appreciation (Bai, 2026).

4. Driving Engines: Dual Empowerment Through United Nations Normative Corpora and OSINT

Supporting this paradigmatic reconstruction requires forging a dual digital engine that integrates high-order normative standards with real-time situational awareness—bridging the divide between static rule systems and dynamic strategic contestation.

4.1 United Nations normative corpora: a high-energy semantic "rule mine"

Multilingual parallel texts produced within the United Nations system represent not merely translation templates but semantic archetypes crystallized through global power negotiations (Sun & Cui, 2025). They encapsulate the core normative contracts and agenda-setting logics of the international order.

Deep semantic annotation of such corpora implants authoritative cognitive kernels into large models, enhancing their capacity for policy inference and discursive negotiation in multilateral diplomacy. This top-down semantic calibration establishes a discursive firewall capable of resisting cognitive erosion in intelligent decision systems (He & Xue, 2025).

4.2 Multilingual OSINT: real-time perception in a high-noise narrative battlefield

In contrast, Open Source INTelligence (OSINT) constitutes a high-frequency, high-noise arena of cognitive contestation. From social media dynamics to real-time statements by foreign officials, OSINT reflects fragmented yet revealing societal psychological currents.

Here, scholars transition from textual commentators to situational intelligence analysts. Their task extends beyond translation to affective polarization detection, narrative deconstruction, and source credibility tracing. Through real-time slicing and annotation of OSINT data, area studies can enable a closed-loop transformation from perception to quantification to scenario simulation.

4.3 Dual-Engine synergy: a closed loop of annotation–evaluation–replenishment

The power of the dual-engine model lies in synergy: UN corpora provide stable, long-cycle semantic anchors, while OSINT offers short-cycle empirical stress tests. Under a human-in-the-loop annotation paradigm, domain experts integrate tacit regional knowledge into dataset processing workflows.

Continuous stress testing identifies contextual biases, which are then corrected through targeted data replenishment and model fine-tuning. This closed-loop system enhances AI comprehension of global affairs while strengthening discursive precision in international arenas—thereby safeguarding national cognitive sovereignty and information security (Li & Xu, 2025).

5. Application Scenarios: From Think Tank Reports to Decision Intelligence

The engineering transformation of area studies aims to convert individualized think tank reporting into interactive and computational decision intelligence. Supported by robust semantic infrastructure, disciplinary empowerment shifts from retrospective summarization to prospective prediction.

5.1 Regional decision models and scenario simulation systems

Structured knowledge extracted from unstructured corpora enables the construction of region-specific foundational decision models. Compared to general-purpose LLMs, these vertical models integrate political power maps, legal architectures, and deep cultural structures of specific regions, thereby achieving stronger semantic penetration.

In foreign trade compliance, for example, fine-grained annotation of regulatory notices and judicial precedents enables automated compliance screening within seconds. Instead of vague risk alerts, the system conducts parameterized scenario simulations based on identified indicators—transforming personal expertise into scalable decision-support services.

5.2 Safeguarding cognitive security and narrative agency

Under algorithm-driven information distribution, engineering achievements in area studies constitute a frontline defense of national cognitive sovereignty. Through the corpus–annotation–evaluation loop, researchers can monitor evolving global narratives concerning China and identify algorithmically

mediated ideological bias.

By applying semantically aligned multilingual outputs consistent with international communication norms, misinterpretations can be countered at the source. This transition from textual mediation to semantic governance embeds autonomous knowledge systems into foundational data structures, correcting bias and mitigating discourse deficits.

5.3 Industry–Education integration and the new liberal arts talent paradigm

Paradigmatic transformation necessitates structural reform in talent cultivation, constructing a competency matrix integrating language proficiency, domain expertise, and data engineering. Universities should establish platforms such as regional data governance centers in collaboration with strategic industries.

Students thus move beyond surface-level translation tasks to participate in semantic rights confirmation processes within vertical sectors (e.g., infrastructure, green energy standard-setting). Through practical annotation and evaluation projects, they evolve into "interface engineers" who combine regional expertise with data logic—serving as strategic talent reserves for national digital transformation.

6. Conclusion: Achieving Strategic Leapfrogging in the Cognitive Revolution

The philosopher of technology Bernard Stiegler once asserted that technology constitutes the exteriorization of memory. In an era when artificial intelligence reshapes the architecture of civilization, data annotation and knowledge engineering transcend technical assistance; they represent the pivotal moment at which humanity transcribes its civilizational patterns, logical rules, and core values into digital life.

For area studies, this engineering reconstruction is not merely defensive adaptation to a legitimacy crisis but a strategic repositioning from peripheral interpretation to foundational digital architecture. In the digital-intelligent era, disciplinary competitiveness no longer depends on data accumulation alone but on the capacity to codify deep global insight—semantic boundary delineation, narrative trap detection, and causal deconstruction—into computable and dynamically evaluable semantic infrastructure.

This paradigm shift redefines the historical trajectory of the discipline. Scholars evolve from information carriers and observers into architects of global cognitive frameworks in the intelligent age.

By securing definitional authority over foundational semantic standards, area studies can construct for the nation an autonomous, resilient, and interpretively sovereign cognitive defense system amid the surging currents of data.

Conflict of Interest

The author declares that he has no conflicts of interest in this work.

Acknowledgement

References

Bai, Y. (2026). Theoretical construction and methodological innovation of China's area studies from the perspective of civilizational mutual learning. *Changbai Journal,* (1), 43–54. https://doi.org/10.19649/j.cnki.cn22-1009/d.2026.01.004

Frontier Domestic Universities Teaching Reform Research Group. (2025). *Special work plan for the reform of professional degree education in translation (2026–2030).*

He, S., & Xue, Y. (2025). A study on foreign language competence cultivation mechanisms for area studies talents in European universities. *Frontiers of Foreign Language Education Research, 8(*6), 87–98. https://doi.org/10.20083/j.cnki.fleic.2025.0062

Li, L., & Xu, R. (2025, December 12). After the suspension of pure foreign language majors, a new track of "Foreign Language+" emerges. *Jiefang Daily*, 005. https://doi.org/10.28410/n.cnki.njfrb.2025.010420

National Development and Reform Commission et al. (2025). *Opinions on strengthening the construction of data element disciplines and majors and the development of digital talent teams (No. 1425 [2025]).*

Ng, A. (2021). *Data-centric AI.* GitHub. https://github.com/

School of Foreign Languages, East China Jiaotong University. (2025). *Report on the establishment of the CEEMA Transportation ESG Data Governance Industry Center at East China Jiaotong University.*

State Council of the People's Republic of China. (2025). *Opinions of the State Council on deeply implementing the "Artificial Intelligence+" action plan.*

Sun, A., & Cui, W. (2025). Exploration and practice of cultivating top innovative foreign language talents under the background of the New Liberal Arts. *Caizhi,* (34), 133–136.