Kwame Nkrumah University Multidisciplinary Journal
Vol. 3 No.2 (September 2026)
© Kwame Nkrumah University 2026
https://accessjournal.nkrumah.edu.zm/index.php/knuj

THE ROLE OF A NATIONAL RESEARCH CORPUS AS CRITICAL INFRASTRUCTURE FOR ACADEMIA-INDUSTRY COLLABORATION

Pethias Siame

Kwame Nkrumah University, Zambia, Email: psiame@yahoo.com, https://orcid.org/0000-0003-3203-5307

Abstract

This conceptual paper argues for the establishment of a large-scale, open-access National Research Corpus (NRC) as essential infrastructure for strengthening Zambia's research and innovation ecosystem. While physical laboratories and scientific equipment are widely recognised as research infrastructure, the linguistic and knowledge-based assets that facilitate collaboration between academic and industry sectors are frequently overlooked. A National Research Corpus, comprising academic publications, technical reports, policy documents, and industry patents, would function as a shared knowledge platform. Using corpus linguistics tools and natural language processing techniques, researchers and industry partners could identify terminological gaps, track innovation trends, and analyse knowledge flows between sectors. Such infrastructure would directly address partnership challenges by providing a common linguistic and conceptual foundation, thereby accelerating problem-solving, reducing research redundancy, and enhancing the value creation networks vital for national economic transformation. The paper contributes a normative conceptual framework, grounded in the Triple Helix model and Knowledge Management theory, for the design and phased implementation of this infrastructure.

Keywords: *Research Infrastructure, Language Corpora, Knowledge Transfer, Academia-Industry Partnerships, Terminology Management.*

Published online: 25 September 2026

1. INTRODUCTION

The global economy is increasingly driven by knowledge and innovation, where the effective translation of research into practical applications has become a critical determinant of national competitiveness (Etzkowitz & Leydesdorff, 2000). In this context, robust collaboration between academic institutions and industry sectors is not merely beneficial but essential for sustained development. Academic institutions generate fundamental knowledge and novel discoveries, while industry possesses the capacity for scaling, market application, and addressing real-world societal challenges. The synergy between these two sectors can fuel economic diversification,

technological advancement, and sustainable development across developing economies.

However, in many developing nations, including Zambia, this potential synergy remains largely untapped. Academia-industry partnerships are frequently hampered by a myriad of challenges, including misaligned incentives, a lack of mutual trust, and, most fundamentally, a communication and knowledge gap (Perkmann *et al.*, 2013; Anibueze & Okoye, 2019). Researchers and industry professionals often operate in separate professional silos, each with its own distinct jargon, priorities, and knowledge repositories. For instance, an agricultural researcher might publish findings on phytonutrient enhancement in a specific legume variety, while an agribusiness executive searches for ways to improve the nutritional value and shelf life of processed food products. The two may be working on complementary problems but lack the shared linguistic and conceptual infrastructure to recognise their mutual relevance.

The collaboration between universities and industries represents a well-established field of study in innovation research. This relationship is frequently framed by models such as the Triple Helix, which emphasises the overlapping and mutually reinforcing roles of university, industry, and government in fostering innovation (Etzkowitz & Leydesdorff, 2000). Globally, successful collaborations between these sectors have been linked to increased research productivity, effective commercialisation of inventions, and sustained regional economic development (Perkmann *et al.*, 2013). The strength of these linkages often distinguishes knowledge-intensive economies from those that remain trapped in resource-dependent growth patterns.

In the African context, scholarly discourse on university-industry collaboration has intensified considerably, with a growing recognition of the need for homegrown solutions to developmental challenges (Mégnigbéto, 2018). However, studies conducted across the continent consistently highlight persistent barriers that hinder effective partnership formation. Anibueze and Okoye (2019) identified factors such as weak policy frameworks, limited funding for collaborative research, and divergent institutional objectives as major impediments in the Nigerian context. In Zambia, a comprehensive study by the Zambian Academy of Sciences (2021) found that while there is genuine willingness to collaborate among both academics and industry practitioners, practical mechanisms for facilitating such collaboration are largely lacking. The same study noted that researchers often perceive industry as solely profit-driven, while industry practitioners view academic research as too theoretical, slow-paced, and disconnected from market realities.

A critical yet underexplored barrier in the existing literature is the socio-linguistic divide between academic and industry communities. As Goring (2020) observed, effective collaboration requires not just a meeting of minds, but a meeting of languages. The specialised lexicon of academic disciplines can constitute a significant barrier to knowledge transfer, a problem that is exacerbated in contexts where English, the primary language of scientific publication, is a second language for many practitioners (Ngulube, 2018). Research on knowledge transfer in multilingual African contexts has demonstrated that linguistic factors play a substantial role in shaping access to and participation in knowledge networks (Ngulube, 2018; Baker & Egbert, 2022). These findings underscore the need for infrastructure that explicitly addresses terminological and communicative barriers. Despite the recognised importance of research infrastructure for innovation, the concept has traditionally been narrowly construed. This paper argues for an expanded understanding of research infrastructure that encompasses knowledge-based digital assets, specifically a National Research Corpus. The paper addresses the following research questions:

1. What is the rationale for reconceptualising a National Research Corpus as critical research infrastructure within a developing-economy innovation system?
2. How can corpus linguistics methods and tools be deployed at a national scale to bridge terminological and communication gaps between academia and industry?
3. What design framework, governance structures, and implementation pathway would be required to establish a sustainable National Research Corpus in Zambia?
4. What are the anticipated applications, limitations, and challenges associated with such an infrastructure?

These questions are addressed through a normative conceptual approach, drawing on literature synthesis and framework development. The paper is structured as follows: Section 2 reviews relevant literature on research infrastructure, corpus linguistics, and African knowledge systems; Section 3 presents the theoretical framework; Section 4 outlines the proposed design framework; Section 5 discusses potential applications; Section 6 addresses implementation challenges; Section 7 acknowledges limitations; Section 8 concludes; and Section 9 offers recommendations.

2. LITERATURE REVIEW

2.1 The Evolving Concept of Research Infrastructure

Traditionally, national research infrastructure has been synonymous with big science facilities such as telescopes, particle accelerators, and genomic sequencing centres (Hodson, 2018). This narrow conceptualisation has historically marginalised investments in knowledge-based infrastructure in developing countries. However, the concept of research infrastructure has evolved dramatically in the digital age. The Organisation for Economic Cooperation and Development (2021) now defines national research infrastructure to include major scientific equipment or sets of instruments, knowledge-based resources such as collections, archives, or scientific data, electronic infrastructures such as data, computing, and communication networks, and any other entity of a unique nature essential to achieve excellence in research and innovation.

This expanded definition clearly encompasses digital repositories, data libraries, and computational platforms as legitimate and essential components of national research infrastructure. Major international initiatives such as the European Open Science Cloud and South Africa's National Integrated Cyberinfrastructure System (NICIS) exemplify this conceptual shift, treating curated data and software as critical national assets (Scholtens & Avenant, 2022). Similar initiatives have emerged across East and West Africa, though their focus has largely remained on scientific data rather than the broader corpus of research texts.

Unlike conventional institutional repositories that primarily function as archival storage systems for individual universities, the proposed National Research Corpus is distinguished by three key features: (1) its national scope, aggregating outputs from all Zambian research-producing institutions; (2) its analytical capabilities, offering corpus linguistics tools such as concordancing, collocation analysis, and keyword extraction rather than simple search and retrieval; and (3) its explicit design as an interface between academia and industry, including industry-generated documents such as patents and technical reports that are typically excluded from academic repositories.

The argument for a Zambian National Research Corpus aligns with this modern, inclusive understanding of infrastructure as a knowledge-based resource and as an electronic infrastructure essential for research excellence and innovation capacity.

2.2 Precedent National and Regional Corpora

The proposal for a Zambian National Research Corpus must be situated within the established tradition of national corpus development. The British National Corpus

(BNC), developed in the 1990s, remains a canonical reference for national-level corpus design, providing a principled model for sampling, representativeness, and balance across genres (McEnery & Hardie, 2012). The Brown Corpus, created in the 1960s, established the foundational methodology for creating balanced, synchronic corpora that could serve as benchmarks for linguistic research. More recently, the Corpus of Contemporary American English (COCA) has demonstrated the value of continuously updated, large-scale corpora for tracking linguistic change and supporting a wide range of research applications (Baker & Egbert, 2022).

These precedents offer several design principles relevant to the Zambian context. First, they demonstrate the importance of principled sampling frameworks that ensure representativeness across genres, registers, and time periods. Second, they highlight the need for sustainable governance and funding models to ensure long-term maintenance and updating. Third, they illustrate the value of user-friendly interfaces and analytical tools that lower barriers to entry for non-specialist users. The proposed National Research Corpus would adapt these principles to the specific needs of a developing-economy innovation system.

2.3 African Regional Initiatives and Precedents

In the African context, several regional initiatives provide directly relevant precedents for the proposed National Research Corpus. The South African Bibliographic and Information Network (SABINET) represents one of the oldest and most established bibliographic networks on the continent, offering a model for aggregating and providing access to research outputs across multiple institutions (Ngulube, 2018). The Database of African Theses and Dissertations (DATAD), coordinated by the Association of African Universities, demonstrates the feasibility of cross-national collaboration in building shared research repositories. The African Journals OnLine (AJOL) platform has successfully enhanced the visibility and accessibility of African-published research, providing a model for the aggregation and dissemination of scholarly content.

These initiatives offer valuable lessons for the proposed National Research Corpus. They demonstrate both the feasibility and the challenges of building shared research infrastructure in resource-constrained environments. They also highlight the importance of institutional partnerships, capacity building, and sustainable funding models. However, these initiatives have largely focused on bibliographic metadata and document access rather than on the linguistic analysis and terminological bridging functions that are central to the proposed National Research Corpus. The proposed NRC would complement and extend these initiatives by adding analytical capabilities specifically designed to facilitate academia-industry collaboration.

Tanzania's COSTECH repository and Ethiopia's national research data platform provide additional relevant comparators, demonstrating the growing recognition across Africa of the importance of national research infrastructure (Scholtens & Avenant, 2022). The proposed NRC would build upon these precedents while introducing novel features designed to address the specific challenge of terminological and communicative barriers between academic and industry communities.

2.4 Corpus Linguistics as a Tool for Knowledge Discovery

Corpus linguistics provides a robust methodological toolkit for analysing large collections of natural language text. This approach moves beyond anecdotal evidence or small-scale case studies to reveal systematic patterns of language use, including word frequency, collocations (words that frequently appear together in meaningful sequences), and discourse structures (McEnery & Hardie, 2012). While traditionally used in linguistic analysis and language education research, the applications of corpus methods have expanded substantially into other fields,

including knowledge management, information science, and innovation studies.

In knowledge management research, corpora have been successfully used to map emerging research trends and intellectual structures within scientific literature (Lei & Liu, 2019). In terminology science, corpora are fundamental for building standardised glossaries and ontologies, ensuring consistent communication within specialised domains (Kockaert & Steurs, 2015). For example, a study by Zhang and Li (2021) used a corpus of engineering and business documents to identify and align terminology across departments, facilitating better communication between research and development teams and marketing departments in a multinational corporation. More recently, researchers have begun applying corpus methods to map knowledge flows between academic and policy communities, with promising results for evidence-informed policymaking (Baker & Egbert, 2022).

The application of corpus linguistics to bridge the academia-industry divide, however, remains a nascent area of inquiry, particularly in the Global South. While individual studies have examined terminology alignment in specific sectors, no research to date has proposed a systematic, national-level application of these principles for infrastructure development in the Zambian context. This paper seeks to contribute to filling this gap by proposing a structured, nationally coordinated approach to corpus development and deployment.

2.5 Digital Humanities and Metadata Standards

The proposed National Research Corpus would also draw on scholarship in the Digital Humanities and Text Encoding Initiative (TEI) communities, which have developed robust standards for corpus governance, metadata creation, and open access design (Wilkinson *et al.*, 2016). The FAIR principles for scientific data management, Findable, Accessible, Interoperable, and Reusable, provide a guiding framework for ensuring that the corpus maximises its value as a public good. The TEI guidelines offer detailed specifications for encoding textual features, enabling sophisticated search and analysis capabilities. These standards and principles would inform the technical architecture and governance framework of the proposed NRC.

3. THEORETICAL FRAMEWORK

This study is underpinned by two interrelated theoretical perspectives: the Triple Helix Model of innovation and Knowledge Management Theory. These frameworks provide complementary lenses for understanding both the challenges and opportunities associated with academia-industry collaboration.

3.1 The Triple Helix Model

The Triple Helix model, developed by Etzkowitz and Leydesdorff (2000), posits that the innovation system in a knowledge-based economy is comprised of three overlapping and interacting helices representing University, Industry, and Government. The synergies and interactions occurring at the overlaps between these helices constitute the primary drivers of innovation within a national system. The model anticipates the emergence of hybrid organisations such as technology transfer offices, science parks, and industry liaison units at these interfaces, serving as institutional mechanisms for cross-sectoral collaboration.

Subsequent scholarship has extended and critiqued the Triple Helix model in several important respects. Mégnigbéto (2018) has developed a game-theoretic extension of the model, demonstrating how collaborative incentives can be modelled and optimised. This work is directly relevant to the proposed National Research Corpus, as it provides analytical tools for understanding the conditions under which academics and industry practitioners will invest in collaborative activities. Critiques of the Triple Helix model from developing-country contexts have highlighted that the assumed symmetry of university-industry-government relations rarely holds in resource-constrained environments, where government capacity may be limited,

industry may be dominated by extractive sectors with weak research linkages, and universities may lack the resources to engage effectively (Anibueze & Okoye, 2019). These critiques underscore the need for infrastructure that can compensate for structural asymmetries by providing shared platforms for interaction and knowledge exchange.

This paper positions the National Research Corpus as a digital hybrid organisation, that is, a platform that exists at the nexus of all three helices. The corpus would be initiated and governed by the state, drawing on government leadership and policy support. It would be populated by the knowledge outputs of scholars from universities and research institutions. Finally, it would be utilised for problem-solving and product development by private sector entities. The National Research Corpus thus acts as a mediator, reducing the friction between the helices and strengthening the ties that bind them together, thereby enhancing the overall innovation capacity of the nation.

The concept of the National Research Corpus as a digital hybrid organisation can be further developed with reference to organisational theory and platform theory. Hybrid organisations are those that combine elements from different institutional logics, in this case, the logics of academic scholarship, commercial enterprise, and public governance (Etzkowitz & Leydesdorff, 2000). Platform theory, as developed in the study of digital economies, emphasises the role of shared infrastructure in reducing transaction costs, enabling complementarities, and creating value through network effects. The National Research Corpus would function as a multi-sided platform, connecting academic producers of knowledge with industry users, while also serving government interests in evidence-based policy and national innovation monitoring. This theoretical framing positions the NRC not merely as a repository but as an active mediator of knowledge flows and collaborative relationships.

3.2 Knowledge Management Theory

Knowledge Management theory distinguishes between tacit knowledge and explicit knowledge (Nonaka & Takeuchi, 1995). Tacit knowledge refers to personal, context-specific knowledge that is difficult to formalise, articulate, or transfer to others. It includes insights, intuitions, and skills that individuals acquire through experience. Explicit knowledge, by contrast, refers to codified knowledge that has been documented in written or symbolic form and is therefore easily transferable across individuals and contexts. The process of converting tacit knowledge to explicit knowledge, and subsequently combining and internalising explicit knowledge, is critical for organisational learning and innovation.

The National Research Corpus is fundamentally a tool for managing a nation's explicit knowledge assets. It systemises the knowledge assets scattered across various institutions, making them findable, accessible, interoperable, and reusable in accordance with the FAIR principles for scientific data management (Wilkinson *et al.*, 2016). By providing powerful search and analysis tools, the National Research Corpus supports the key knowledge management processes of knowledge capture, knowledge sharing, and knowledge application. It helps industry partners capture relevant academic research that might otherwise remain invisible. It enables the sharing of best practices documented in policy papers and technical reports across sectoral boundaries. Finally, it facilitates the application of patented technologies and research findings in new commercial and policy contexts. In doing so, the National Research Corpus creates a public good that substantially lowers the transaction costs of finding, evaluating, and using existing knowledge.

The knowledge management perspective also highlights the importance of addressing tacit knowledge dimensions of academia-industry collaboration. While the National Research Corpus focuses on explicit knowledge, it can also support the development of shared tacit knowledge by facilitating interactions and building trust between communities. For example, the terminological mapping functions of the

corpus can help researchers and industry practitioners understand each other's conceptual frameworks, laying the groundwork for more substantive collaboration.

4.PROPOSED FRAMEWORK FOR NATIONAL RESEARCH CORPUS DEVELOPMENT

This section presents a conceptual design framework and implementation proposal. No empirical data were collected or analysed. The framework described below outlines a proposed approach for future corpus development rather than reporting on an existing implementation.

4.1 Corpus Design and Compilation

The National Research Corpus would be designed as a dynamic, open-access, digital repository. Its design would follow established corpus linguistics principles to ensure representativeness, balance, and long-term usability (McEnery & Hardie, 2012). The corpus would be stratified to include multiple genres that are crucial for the innovation ecosystem.

First, the academic sub-corpus would include peer-reviewed journal articles by Zambian authors, doctoral and master's theses from Zambian universities, and conference proceedings from national and regional academic gatherings. Second, the government and policy sub-corpus would include national development plans such as the Eighth National Development Plan, sector-specific policies in agriculture, health, energy, and other domains, and technical reports from ministries and government agencies. Third, the industry sub-corpus would include patent filings from the Patents and Companies Registration Agency, annual reports of major Zambian companies, technical manuals, and market analysis reports where these are publicly available.

4.2 Multilingual Corpus Design Considerations

Zambia is a linguistically diverse nation with 73 indigenous languages, of which seven are recognised as official languages for educational and administrative purposes. While English functions as the primary language of formal education and scholarly publication, it is a first language for only a very small minority of the population. This linguistic reality presents both challenges and opportunities for the proposed National Research Corpus.

The initial phase of corpus development would focus on English-language texts, for several practical reasons. First, English is the dominant language of academic publication in Zambia and internationally, making English-language texts the primary vehicle for scholarly knowledge transfer. Second, the technical infrastructure for corpus analysis is most developed for English, reducing initial implementation costs. Third, the majority of policy documents and formal industry reports are produced in English.

However, the limitation to English-language texts would fundamentally constrain the corpus's claim to comprehensiveness. Many industry documents, community-level reports, and policy consultations are conducted in Bemba, Nyanja, Tonga, and other local languages. These texts may contain crucial knowledge about local contexts, indigenous innovations, and community needs that would remain invisible to an English-only corpus. The paper's argument about addressing language barriers would be undermined if the corpus itself perpetuated linguistic exclusion.
To address this limitation, the proposed framework includes a phased approach to multilingual corpus development. Phase one would establish the English-language corpus while simultaneously developing the technical and institutional capacity for multilingual corpus development. Phase two would add major Zambian language corpora, beginning with the seven official languages and potentially expanding to other languages as resources permit. This phased approach would allow the project to demonstrate value and build momentum while working toward the long-term goal of a truly multilingual national corpus. The experience of other multilingual societies

in developing corpus resources would inform this work (Ngulube, 2018).

4.3 Corpus Analysis Tools and Techniques

The raw text of the corpus would be processed and made accessible through a user-friendly web interface equipped with specialised corpus analysis tools. The rationale for selecting these specific tools is grounded in established corpus linguistics principles: concordancing reveals contextual usage patterns, frequency analysis identifies dominant themes, collocation analysis exposes conceptual networks, and keyword analysis defines domain-specific preoccupations. Key functionalities would include the following components:

The first component is a concordancer. A concordancer would display every occurrence of a search word or phrase in its immediate context (keyword-in-context display). This tool is fundamental for understanding how a term is used differently in academic texts compared to industry documents.

The second component is frequency analysis. Frequency analysis would generate lists of the most frequent words, phrases, or technical terms in the entire corpus or within specific sub-corpora, such as mining engineering theses compared to mining company reports.

The third component is collocation analysis. Collocation analysis would identify words that statistically tend to co-occur with a target term, revealing conceptual networks and domain-specific associations. For example, in the agriculture sub-corpus, the term soil might collocate with erosion in academic texts, while collocating with fertility and yield in industry documents.

The fourth component is keyword analysis. Keyword analysis would compare a specialised sub-corpus against a larger reference corpus to identify words that are unusually frequent in the specialised corpus. These keywords effectively define a domain's core concepts and preoccupations.

The fifth component is visualisation tools. Visualisation tools would generate word clouds, trend graphs over time, and network diagrams showing the connections between key concepts, institutions, and research communities.

5. Potential Applications and Expected Outcomes

This section presents hypothetical but realistic examples based on the proposed methodology and existing corpus linguistics research. No empirical data were collected or analysed. The following scenarios illustrate anticipated outcomes rather than reporting actual findings.

5.1 Bridging the Terminological Divide

Consider a Zambian agribusiness seeking to develop drought-resistant maize varieties. Their research and development team might search for terms such as drought tolerance, water stress, and maize yield. A conventional search of academic databases would return relevant documents, but the distinctive power of the National Research Corpus lies in its analytical capabilities rather than simple retrieval.

Using the concordancer, the agribusiness team could search for the term drought and immediately see how this concept is used in academic theses compared to industry reports. They might discover that academic research frequently mentions specific genes or physiological processes, such as proline accumulation, while industry reports focus on field performance and cost-benefit analysis. This insight alone is valuable because it tells the industry team which language to use when searching for academic partners and which type of evidence (whether field trial data or genomic data) will be most persuasive to their management. As noted by Kockaert and Steurs (2015), such terminological alignment is essential for successful knowledge transfer across specialised communities.

Furthermore, a collocation analysis for the term maize in the academic sub-corpus might reveal strong statistical links to biofortification and Vitamin A, suggesting a parallel body of research that the agribusiness could integrate into their product development for enhanced nutritional marketing. This finding aligns with the observations of Zhang and Li (2021), who demonstrated that corpus-based terminology mapping can reveal unexpected connections between previously siloed research domains.

5.2 Tracking Knowledge Flows and Innovation Trends

A National Research Corpus can act as a barometer for the national innovation system. By analysing the metadata and textual content over time, policymakers and researchers can answer critical questions about the health and direction of research and innovation activities.

One important question concerns whether academic research is addressing national priorities. A frequency analysis of keywords drawn from national development plans, such as diversification, value addition, and climate-smart agriculture, within the academic sub-corpus over a decade can reveal whether research focus is aligning with policy goals. In a hypothetical scenario, the results could be visualised as trend lines showing increasing or decreasing attention to these priority areas. If alignment is weak, this finding would prompt targeted interventions such as funding incentives or research priority setting exercises.

Another important question concerns the pattern of collaboration among researchers and between academia and industry. Network analysis based on co-authorship patterns among academics and citations of patents in academic papers can map the collaboration network within Zambia. A weak or non-existent link between a university's school of mines and major mining corporations, for instance, would be visually apparent from the network diagram and could prompt targeted interventions to strengthen those connections. As Lei and Liu (2019) demonstrated, bibliometric and corpus-based methods can reveal the intellectual structure of research fields and identify influential institutions and research groups.

5.3 Reducing Redundancy and Identifying Gaps

A common complaint from industry practitioners is that academic researchers frequently reinvent the wheel by undertaking studies that have already been completed by others. A comprehensive National Research Corpus can mitigate this problem substantially. Before embarking on a new research project, a graduate student or experienced researcher would be required to query the National Research Corpus. A simple search could reveal a technical report from a national research institute or a patent from a local company that has already addressed a similar problem, allowing the researcher to build upon existing work rather than duplicating effort. This function directly supports the knowledge management principle of knowledge reuse and reduces wasteful expenditure of scarce research resources (Nonaka & Takeuchi, 1995).

Conversely, the National Research Corpus can reveal terminological and research gaps that represent opportunities for new investment. If policy documents heavily emphasise renewable energy and solar power, but the academic and industry sub-corpora show very few mentions of battery storage technology or grid integration, this finding clearly signals a critical area for future research investment and industrial development. Such gap identification is a fundamental function of knowledge management systems, enabling strategic resource allocation rather than reactive responses to funding opportunities.

The projected outcomes discussed above are consistent with the broader literature on corpus applications in knowledge management. For example, Baker and Egbert (2022) demonstrated that corpus analysis of policy and research documents can identify areas where evidence is lacking or where terminological mismatches impede

effective knowledge utilisation. Similarly, Scholtens and Avenant (2022) showed that national-level cyberinfrastructure can transform research practices by making previously siloed knowledge assets discoverable and reusable.

6. ADDRESSING IMPLEMENTATION CHALLENGES

The successful implementation of a National Research Corpus faces several significant challenges that must be acknowledged and addressed proactively. This section discusses these challenges and proposes mitigation strategies.

Funding sustainability: Initial development costs and long-term maintenance require predictable funding. A mixed model of government baseline funding, donor support for capital expenditure, and cost-recovery fees for commercial users is recommended. The experience of comparable initiatives, such as South Africa's NICIS and the European Open Science Cloud, suggests that sustainable funding requires diversified sources and strong institutional commitment (Scholtens & Avenant, 2022).

Copyright and intellectual property concerns: Publishers' copyright restrictions may limit the deposit of published articles. Negotiations with publishers, reliance on green open access mandates, and clear embargo policies will be necessary. The FAIR principles provide a framework for balancing openness with legitimate intellectual property concerns (Wilkinson *et al.*, 2016).

Data governance and privacy: Industry documents may contain commercially sensitive information. A tiered access model (open, registered user, restricted) with clear data use agreements is required. This model would allow industry partners to contribute sensitive documents while protecting their commercial interests.

Technical infrastructure requirements: Adequate server capacity, data backup systems, and ongoing software development expertise must be budgeted and secured. The technical architecture should be designed for scalability and interoperability with regional and international systems.

Institutional resistance and adoption barriers: Universities and companies may be reluctant to share knowledge assets. Incentives such as research funding preferences for depositors and recognition mechanisms for contributing institutions should be established. Capacity building programmes for researchers, graduate students, and industry professionals would help to build a culture of corpus use.
Multilingual capacity: The development of multilingual corpus resources requires specialised expertise in computational linguistics and language technology for Zambian languages. Partnerships with regional and international research centres could help to build this capacity over time.

Addressing these challenges from the outset through a multi-stakeholder governance structure will be critical to the long-term viability and impact of the National Research Corpus.

7. LIMITATIONS AND FUTURE RESEARCH

This paper presents a conceptual framework for a National Research Corpus in Zambia. Several limitations must be acknowledged, and these limitations also suggest directions for future research.

First, the paper's entire argument rests on hypothetical scenarios (Sections 5.1, 5.2, 5.3) that are presented with illustrative plausibility but without empirical validation. The limits of prospective, non-empirical reasoning should be explicitly acknowledged. Future research should include pilot studies that test the proposed corpus design and analytical tools with real-world users in academic and industry settings. Such studies would provide empirical evidence on the utility and usability of the proposed infrastructure.

Second, the transferability of corpus-based methods from OECD or large-economy

contexts (the bulk of the literature cited) to a low-resource, multilingual environment like Zambia is assumed but not critically examined. Future research should investigate the specific challenges and opportunities of applying corpus methods in African contexts, including issues of data scarcity, language technology capacity, and institutional readiness.

Third, the paper's scope is limited to English-language texts in its initial phase, yet many Zambian industry and community-level documents exist in Bemba, Nyanja, Tonga, and other local languages. This limitation fundamentally affects the claimed comprehensiveness of the proposed corpus. Future research should explore the technical and institutional requirements for multilingual corpus development, including the development of language technology resources for Zambian languages and the establishment of partnerships with language communities.

Fourth, the paper does not empirically validate the claim that terminological and communication gaps are a primary barrier to academia-industry collaboration in Zambia. While the literature supports this claim (Goring, 2020; Ngulube, 2018), further empirical research in the Zambian context would strengthen the evidence base for the proposed intervention.

Fifth, the governance and sustainability model proposed in this paper is preliminary and would require detailed elaboration and stakeholder consultation before implementation. Future research should investigate appropriate governance structures, funding models, and incentive mechanisms for national research infrastructure in developing-economy contexts.

8. CONCLUSION

This paper has argued that the establishment of a National Research Corpus is a strategic imperative for Zambia's socioeconomic development. Moving beyond the traditional view of research infrastructure as limited to physical facilities and equipment, the National Research Corpus conceptualises a nation's collective, codified knowledge as a critical and analysable national asset. By serving as a shared linguistic and conceptual platform, this infrastructure directly addresses the often invisible barrier of terminology and communication that plagues academia-industry collaboration in developing-economy contexts.

Grounded in the Triple Helix model of innovation and Knowledge Management theory, the National Research Corpus is envisioned as a digital public good that enhances the findability, accessibility, and utility of research outputs across sectors. Through practical applications including terminological mapping, trend analysis, and gap identification, the corpus can accelerate problem-solving, foster strategic partnerships, and ensure that national research efforts are more closely aligned with both industrial needs and policy priorities. The framework proposed in this paper builds upon and extends the existing literature on corpus applications in knowledge management, while adapting these methods to the specific challenges facing developing-economy innovation systems.

The development of a National Research Corpus is not a trivial undertaking. It requires sustained political will, adequate and predictable funding, and genuine multi-stakeholder collaboration across government, academic, and industry sectors. However, the potential returns on this investment, namely, a more integrated, efficient, and impactful national innovation system, are profound. In an era of intensifying global competition for knowledge-intensive investment, the establishment of such infrastructure represents a foundational step towards building a truly knowledge-based Zambian economy.

9. RECOMMENDATIONS

To translate this conceptual proposal into practical reality, the following specific actions are recommended for policymakers, institutional leaders, and development partners.

First, government leadership is essential. The Ministry of Higher Education, in collaboration with the Ministry of Technology and Science and the Zambian Academy of Sciences, should champion the development of the National Research Corpus. This process should begin with a comprehensive feasibility study followed by the formation of a multi-stakeholder task force representing universities, industry associations, and government agencies.

Second, the project should be implemented in carefully planned phases. Phase one would develop a prototype corpus focusing on one or two key sectors (e.g., agriculture and mining), aggregating existing open-access theses, policy documents, and patent data. Timeline: 12-18 months. Phase two would expand the corpus to include all major disciplines, strengthen data harvesting pipelines, and develop a user-friendly web interface with basic analytical tools. Timeline: 18-24 months. Phase three would integrate the National Research Corpus into the national research ecosystem, mandating deposition for publicly funded research and exploring sustainable funding models, including government allocation, donor support, and cost recovery mechanisms for commercial users. Timeline: 24-36 months.

Third, capacity building is essential for successful adoption. Universities must be supported to strengthen their institutional repositories, ensuring a high-quality and steady flow of content into the National Research Corpus. Regular workshops should be conducted for researchers, graduate students, and industry professionals on how to use the corpus effectively for partnership building and literature discovery.

Fourth, policy integration will drive sustained use. National research funding bodies, such as the National Science and Technology Council, should make consultation of the National Research Corpus a mandatory part of grant proposal development. This requirement would demonstrate the novelty and relevance of proposed research while also exposing applicants to potential industry partners. Measurable indicators of success would include: (a) number of cross-sectoral searches conducted annually; (b) number of documented academia-industry partnerships initiated through corpus-facilitated discovery; (c) reduction in redundant research proposals; and (d) citations of corpus data in policy documents and industry reports.
Fifth, international linkage will amplify impact. The Zambian National Research Corpus should be designed for interoperability with regional and international corpus initiatives, including the African Language Corpus and global open science platforms. This connectivity will ensure that Zambian research outputs are discoverable within the global knowledge network, attracting international collaboration and investment.

Sixth, multilingual development should be planned from the outset. While the initial phase would focus on English-language texts, the long-term vision should include the development of multilingual corpus resources. Partnerships with regional and international research centres should be established to build capacity in language technology for Zambian languages.

REFERENCES

Anibueze, C. O. & Okoye, I. E. (2019). Barriers to university-industry collaboration in a developing economy: A Nigerian perspective. *Journal of Research in National Development*, 17(1), 1-12.

Baker, P. & Egbert, J. (2022). *Triangulating Methodological Approaches in Corpus Linguistics Research*. Routledge, London.

Etzkowitz, H. & Leydesdorff, L. (2000). The dynamics of innovation: from national systems and "Mode 2" to a Triple Helix of university, industry, and government relations. *Research Policy*, 29(2), 109-123.

Goring, P. (2020). *The Rhetoric of Science*. Routledge.

Hodson, S. (2018). The evolving role of national research infrastructures in the European research area. *Insights*, 31(0), 1-10.

Kockaert, H. J. & Steurs, F. (eds) (2015). *Handbook of Terminology Management, Volume 2*. John Benjamins Publishing Company, Amsterdam.

Lei, L. & Liu, D. (2019). Research trends in applied linguistics from 2005 to 2016: A bibliometric analysis and its implications. *Applied Linguistics*, 40(3), 540-561.

McEnery, T. & Hardie, A. (2012). *Corpus Linguistics: Method, Theory and Practice*. Cambridge University Press.

Mégnigbéto, E. (2018). Modelling the Triple Helix of university-industry-government relationships with game theory: Core, Shapley value and nucleolus. *Journal of Informetrics*, 12(4), 1118-1132.

Ministry of Higher Education, Zambia (2017). *National Policy on Research and Development*. Government of the Republic of Zambia.

Ngulube, P. (2018). Overcoming the language barrier in African scholarship. *African Journal of Library, Archives and Information Science*, 28(2), 95-108.

Nonaka, I. & Takeuchi, H. (1995). *The Knowledge Creating Company: How Japanese Companies Create the Dynamics of Innovation*. Oxford University Press.

Organisation for Economic Cooperation and Development (2021). *Main Science and Technology Indicators*. OECD Publishing.

Perkmann, M., Tartari, V., McKelvey, M., Autio, E., Broström, A., D'Este, P., Fini, R., Geuna, A., Grimaldi, R., Hughes, A., Krabel, S., Kitson, M., Llerena, P., Lissoni, F., Salter, A. & Sobrero, M. (2013). Academic engagement and commercialisation: A review of the literature on university-industry relations. *Research Policy*, 42(2), 423-442.

Scholtens, M. & Avenant, M. (2022). The role of cyberinfrastructure in enabling a data-driven research ecosystem: The South African case. *Data Intelligence*, 4(2), 432-451.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J. W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., Gonzalez Beltran, A., Gray, A. J. G., Groth, P., Goble, C., Grethe, J. S., Heringa, J., 't Hoen, P. A. C., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S. J., Martone, M. E., Mons, A., Packer, A. L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S. A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M. A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J. & Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(1), 1-9.

Zambian Academy of Sciences (2021). *Landscape Study on Academia-Industry Linkages in Zambia*. ZAS Press.

Zhang, Y. & Li, J. (2021). Aligning research and development and marketing terminology through corpus analysis: A case study of a manufacturing firm. *Journal of Knowledge Management*, 25(5), 1125-1143.