

## INSIGHTS

### **Towards Developing Digital Humanities as a Discipline in Cambodia**

*Dechkunn Chhay<sup>1</sup>, Master in Digital Humanities*

---

<sup>1</sup> *Dechkunn Chhay is a Lecturer of AI and Data Science, CamTech University.*

## 1. Introduction

Digital Humanities (DH) refers to a field of scholarship that involves the systematic use of digital tools and resources to investigate research questions within the humanities. It is widely understood as a collaborative and interdisciplinary domain, integrating diverse practices, methods, and perspectives from both the humanities and computational sciences (Brier, 2012). The emergence of DH has been closely tied to the expansion of digital archives, the development of quantitative and computational methods, and the increasing accessibility of digital tools for research. More recently, advances in artificial intelligence (AI), particularly large language models (LLMs) such as GPT models, have opened new methodological possibilities for DH research.

The origins of Digital Humanities (DH) can be traced back to the mid-twentieth century, marked by early initiatives in humanities computing and the establishment of foundational platforms such as *Computers and the Humanities* (1966), one of the first specialized journals in the field (Hockey, 2004). Since then, scholarly communities and professional organizations have played a significant role in shaping DH as a discipline. However, DH has historically been concentrated in Western contexts. Anglophone countries such as the United States, the United Kingdom, and Canada dominate the field in terms of publication output and institutional development, followed by several European countries, including the Netherlands and Germany (Kirtania, 2021).

This concentration raises critical concerns regarding the epistemological orientation of the field. On the one hand, DH risks reinforcing Eurocentric narratives through the uneven distribution of resources, infrastructures, and scholarly visibility. On the other hand, it also presents an opportunity to function as a mechanism of decolonial praxis, which risks foregrounding underrepresented voices, challenging dominant knowledge systems, and incorporating non-Western epistemologies (Risam, 2019).

This article examines the institutional, data-related, and technological constraints shaping DH scholarship, with particular attention to the Cambodian context. It further proposes strategic recommendations for educational institutions and researchers in Cambodia to support the sustainable development of DH as an inclusive and locally grounded discipline.

## 2. Institutional Constraints

At the institutional level, DH has become increasingly established within higher education systems in North America and Europe, where it is offered as specialized courses, summer schools, and full degree programs. The flexibility of the U.S. higher education system, in particular, has facilitated interdisciplinary collaboration by enabling humanities students to engage with computational methods since as early as the 1970s (Brier, 2012).

In Europe, DH has experienced significant growth due to the presence of established academic networks and funding frameworks. For example, Germany has witnessed a rapid expansion of DH programs across public universities, ranging from individual courses to comprehensive degree programs (Fiormonte, 2014). In contrast, the development of DH in Asia remains uneven. While countries such as Japan, China, and Singapore have made notable progress, many other nations in the region are still in the early stages of adopting DH practices.

In Japan, dedicated research centers such as the Digital Humanities Center for Japanese Arts and Cultures at Ritsumeikan University and the Center for Digital Humanities and Social Sciences (CDHSS) at Nagoya University have contributed significantly to the field. Similarly, Singapore has established the Centre for Computational Social Science and Humanities (CSSH), one of the first institutions in Southeast Asia to focus on DH research, with particular emphasis on text, heritage, and culture. In other parts of the Asia-Pacific region, DH is often introduced through specialized courses rather than full academic programs. For instance, the Cambodia University of Technology and Science (CamTech) has incorporated DH into its curriculum at the foundational level. Introducing DH through coursework serves as an initial step that allows students and practitioners to engage with its interdisciplinary nature and explore its methodological diversity. Such exposure also lays the groundwork for future interdisciplinary collaboration and joint research initiatives across the humanities and computational sciences.

## 3. Data Constraints

The expansion of Digital Humanities (DH) in North America and Europe has been largely driven by increased access to digitized collections from libraries and archives. Advances in technology have enabled large-scale digitization projects, making vast amounts of cultural and historical materials available for computational analysis (Fiormonte, 2014). However, these developments

have not equally benefited all linguistic and cultural communities. A significant limitation of DH lies in its dependence on the availability and accessibility of digitized resources.

In the Cambodian context, these limitations are further compounded by historical factors. The country experienced prolonged periods of civil war and the Cambodian genocide in the late twentieth century, which resulted in the destruction and loss of a substantial portion of its cultural and archival heritage. Consequently, many primary sources are either incomplete, inaccessible, or no longer exist. This scarcity puts scholars at a disadvantage as they have to rely on colonial archives or foreign-produced records such as those generated by French researchers. These materials often reflect non-Khmer perspectives and epistemologies (Miley & Read, 2021). Such reliance raises important concerns regarding representation, bias, and the framing of Cambodian history within DH scholarship.

This imbalance is also evident on digital platforms. For example, while the English version of Wikipedia contains over 7 million articles and the French version over 2.7 million, the Khmer version includes only around ten thousand articles. This disparity highlights the structural challenges faced by researchers working with underrepresented languages and underscores the urgent need for targeted efforts to expand digital resources.

Despite these challenges, Cambodia is ripe for the development of DH with its rich historical and cultural heritage. Numerous initiatives have been undertaken to preserve and restore cultural artifacts, many of which remain vulnerable due to age, environmental conditions, and limited preservation infrastructure. DH methodologies offer innovative approaches to support these efforts through digitization, computational analysis, and enhanced accessibility, thereby creating new opportunities for both scholarship and cultural preservation.

#### **4. Technological Constraints**

In addition to limitations in resources, DH scholarship in the Khmer language faces significant challenges related to the availability of computational tools and research methodologies. As many DH studies rely heavily on textual datasets, the application of methods from Natural Language Processing (NLP) is essential for enabling systematic analysis. However, in the case of the Khmer language, NLP tools remain underdeveloped due to both the limited number of specialized researchers in this domain and the scarcity of large, high-quality linguistic datasets.

A key example of this limitation is tokenization. Tokenization is the process of segmenting text into smaller units such as words, subwords, or morphemes. The process serves as a foundational step in most NLP pipelines. Unlike many high-resource languages, particularly those using Latin scripts where words are typically separated by spaces, Khmer script does not employ explicit word delimiters. This structural characteristic makes tokenization considerably more complex and necessitates the development of language-specific approaches. As a result, tokenization has emerged as a distinct research area within Khmer NLP, with several scholars proposing algorithms and techniques aimed at improving segmentation accuracy such as conditional random fields (CRF) and recurrent neural networks (RNN) (Kaing et al., 2021). Beyond tokenization, several other fundamental tasks in Natural Language Processing including part-of-speech (POS) tagging, named entity recognition (NER), stemming, and lemmatization, remain underexplored in the context of the Khmer language. This limited research attention further constrains the development of robust linguistic tools and hampers the advancement of data-driven studies in both NLP and Digital Humanities.

Beyond the constraints posed by limited NLP technologies, researchers have increasingly leveraged the availability of data from high-resource languages such as English to perform tasks in low-resource languages through cross-lingual transfer. This approach enables models trained on resource-rich languages to be adapted for use in languages with limited annotated data. Cross-lingual transfer has shown promising results in tasks such as machine translation and text recognition (Nom et al., 2025, Shurtz & Richardson., 2025).

However, this method is not without limitations. Its effectiveness is often dependent on the degree of linguistic similarity between the source and target languages (Lauscher et al., 2020). In the case of Khmer, these challenges are particularly pronounced. Khmer belongs to the abugida writing system, which differs significantly in both syntax and orthographic structure from high-resource languages such as English and Chinese. As a result, the performance of cross-lingual approaches may be limited, necessitating the development of language-specific models and resources tailored to the unique characteristics of Khmer.

## **5. Recommendations for Future Development**

To foster the growth of Digital Humanities in Cambodia, several strategic actions should be considered:

1. Enhancing Digital Infrastructure and Access

National and local libraries should adopt standardized practices for digitizing historical documents and archives. Ensuring that these materials are machine-readable will facilitate computational research and broaden access for scholars.

2. Promoting Interdisciplinary Education

Universities should encourage collaboration between departments of computer science and the humanities. Introducing DH as specialized courses for students in both fields can serve as an effective starting point for interdisciplinary engagement.

3. Strengthening Research Initiatives

Research institutions should prioritize DH-themed projects that integrate computational methods with cultural and historical studies, fostering innovation within both domains.

4. Encouraging Regional Collaboration

Cambodian institutions should establish partnerships with leading DH centers in Asia, such as the Centre for Computational Social Science and Humanities and the Center for Digital Humanities and Social Sciences. Such collaborations can facilitate knowledge exchange, capacity building, and the promotion of DH scholarship within the region.

## 6. Conclusion

Digital Humanities (DH) has matured as a discipline in North America and Europe, where scholars benefit from abundant research resources, well-established institutional support, and active academic networks for the exchange of ideas and practices. However, much of this scholarship remains grounded in Western data, infrastructures, and methodological frameworks. As a result, researchers working on non-Western contexts often encounter structural and epistemological challenges that limit the inclusivity and global applicability of the field.

In the case of Cambodia, these challenges are particularly pronounced. Institutional constraints continue to hinder the development of DH, especially due to limited academic support and the lack of sustained interdisciplinary collaboration between computer science and humanities scholars. Furthermore, DH scholarship in Cambodia faces significant obstacles related to the scarcity of digitized materials, limited availability of high-quality datasets, and the underdevelopment of computational tools tailored to the Khmer language and script.

Addressing these challenges will require coordinated efforts to strengthen institutional frameworks, expand digital resources, and foster interdisciplinary and regional collaborations. Such initiatives

are essential not only for advancing DH scholarship in Cambodia but also for contributing to a more inclusive and globally representative DH landscape.

## References

- Brier, S. (2012). Where's the pedagogy? the role of teaching and learning in the Digital Humanities. *Debates in the Digital Humanities*, 390–401. <https://doi.org/10.5749/minnesota/9780816677948.003.0038>
- Fiormonte, D. (2014). Digital Humanities from a global perspective. *Laboratorio dell'ISPF*, 11(10.12862), 121-136. [http://www.ispflab.cnr.it/2014\\_203.pdf](http://www.ispflab.cnr.it/2014_203.pdf).
- Hockey, S. (2004). The history of humanities computing. *A companion to digital humanities*, 1-19. <https://doi.org/10.1002/9780470999875.ch1>
- Kaing, H., Ding, C., Utiyama, M., Sumita, E., Sam, S., Seng, S., Sudoh, K., & Nakamura, S. (2021). Towards tokenization and part-of-speech tagging for Khmer: Data and discussion. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 20(6), 1–16. <https://doi.org/10.1145/3464378>
- Kirtania, D. K. (2021). Bibliometric Analysis of Open Access Digital Humanities Publications. *Library Philosophy and Practice (e-journal)*. 6667. <https://digitalcommons.unl.edu/libphilprac/6667>
- Lauscher, A., Ravishankar, V., Vulić, I., & Glavaš, G. (2020). From Zero to hero: On the limitations of zero-shot language transfer with multilingual transformers. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 4483–4499. <https://doi.org/10.18653/v1/2020.emnlp-main.363>
- Miley, F., & Read, A. (2021). Entertainment as an archival source for historical accounting research. *Accounting History*, 26(1), 146-167. <https://doi.org/10.1177/103237322096921>
- Nom, V., Keo, S., Bakkali, S., Luqman, M. M., Coustaty, M., & Ogier, J.-M. (2025, September). Cross-lingual learning for low-resource khmer scene text detection and recognition. *ICDAR 2025 Workshop on Documents Analysis of Low-Resource Languages*, <https://hal.science/hal-05191219v1>
- Risam, R. (2019). *New Digital Worlds: Postcolonial Digital Humanities in Theory, Praxis, and Pedagogy*. Northwestern University Press. <https://doi.org/10.2307/j.ctv7tq4hg>

Shurtz, A., Richardson, C., & Richardson, S. D. (2025). When scripts diverge: Strengthening low-resource neural machine translation through phonetic cross-lingual transfer. *Proceedings of the 5th Workshop on Multilingual Representation Learning (MRL 2025)*, 336–346.  
<https://doi.org/10.18653/v1/2025.mrl-main.22>

\* Edited with ChatGPT for clarity; all views expressed are solely those of the author.

\* The opinions expressed are the author's own and do not reflect the views of AI Forum.