

นวัตกรรมวิธีในมนุษยศาสตร์ดิจิทัล: จากบริบทโลกสู่สังคมไทย¹

Innovative Methods in Digital Humanities: From Global Contexts to Thai Society

วิโรจน์ อรุณมานะกุล^{1,*} Wirote Aroonmanakun^{1,*}

¹ สถาบันภาษาไทยสิรินธร จุฬาลงกรณ์มหาวิทยาลัย ประเทศไทย; Sirindhorn Thai Language Institute, Chulalongkorn University, Thailand

* Corresponding author email: wirote.a@chula.ac.th

บทคัดย่อ

มนุษยศาสตร์ดิจิทัลเป็นศาสตร์ที่นำความก้าวหน้าทางเทคโนโลยีคอมพิวเตอร์มาประยุกต์ใช้ในทางมนุษยศาสตร์ ซึ่งไม่ได้เป็นเพียงการใช้ประโยชน์จากคอมพิวเตอร์มาประมวลผลข้อมูลเท่านั้น แต่ยังส่งผลให้เกิดกระบวนการศึกษาวิจัยทางมนุษยศาสตร์ในรูปแบบใหม่ ๆ ขึ้น และเกิดนวัตกรรมวิธีสำหรับการศึกษาทางมนุษยศาสตร์ที่หลากหลาย เพื่อให้เข้าใจภาพรวมถึงผลกระทบและการเปลี่ยนแปลงของการศึกษาทางมนุษยศาสตร์ดิจิทัล ในบทความนี้ส่วนแรกจึงนำเสนอวิธีต่าง ๆ ที่ได้มีการนำมาใช้ศึกษาวิจัยทางมนุษยศาสตร์ นวัตกรรมเหล่านี้ได้แก่ คลังข้อมูลภาษา หอเอกสารดิจิทัล เมทาเดตา วิธีการทางสถิติ การแสดงภาพนิทัศน์ แบบจำลองตัวแทน การประมวลผลภาษา และสื่อสังคมออนไลน์ หลังจากนำเสนอภาพรวมของนวัตกรรมวิธีในการศึกษาทางมนุษยศาสตร์ดิจิทัลแล้ว ส่วนที่สองจึงนำเสนอผลการวิจัยของโครงการ “นวัตกรรมวิธีในมนุษยศาสตร์ดิจิทัล: การพัฒนาและการประยุกต์ใช้” ซึ่งประกอบด้วยโครงการย่อยห้าโครงการเพื่อเป็นตัวอย่างแสดงให้เห็นถึงการพัฒนาและประยุกต์ใช้นวัตกรรมวิธีในการศึกษาทางมนุษยศาสตร์ด้านต่าง ๆ ในบริบทสังคมไทย โครงการเหล่านี้ช่วยให้เห็นภาพและเข้าใจว่าการนำนวัตกรรมวิธีต่าง ๆ ที่กล่าวถึงในตอนแรกนั้นสามารถประยุกต์ใช้กับการศึกษาทางมนุษยศาสตร์ในบริบทสังคมไทยได้ และเป็นการเปลี่ยนแปลงที่สำคัญของการศึกษามนุษยศาสตร์ดิจิทัล

คำสำคัญ: มนุษยศาสตร์ดิจิทัล นวัตกรรมวิธี

Abstract

Digital Humanities is a discipline that integrates advancements in computer technology with the humanities. It goes beyond merely leveraging computers for data processing, fostering novel approaches to humanities research and introducing diverse methodologies for

¹ เนื้อหาส่วนแรกในบทความนี้แก้ไขดัดแปลงจากบทความที่ผู้เขียนเผยแพร่ในเว็บเมื่อ 1 ม.ค. 2567, “นวัตกรรมวิธีในมนุษยศาสตร์ดิจิทัล” (<https://awrote.medium.com/นวัตกรรมวิธีในมนุษยศาสตร์ดิจิทัล-990f1d27bf04>)

humanities studies. To provide an overview of the impacts and transformations brought about by Digital Humanities, this article first presents various innovative methods employed in humanities research, including language corpora, digital archives, metadata, statistical methods, data visualization, agent-based modeling, natural language processing, and social media. After discussing these methodologies, the second part of the article highlights findings from the research project titled “Innovative Methods in Digital Humanities: Development and Applications.” This project comprises five sub-projects that serve as examples of the development and application of these methodologies in different areas of humanities research within the context of Thai society. These projects illustrate how the innovative methods can be effectively applied to humanities studies in the Thai context, marking a significant transformation in the field of Digital Humanities.

Keyword: Digital humanities, Innovative methods

1. มนุษยศาสตร์ดิจิทัลคืออะไร

มนุษยศาสตร์ดิจิทัล (Digital Humanities) เป็นสาขาที่ได้รับความนิยมอย่างมากในปัจจุบัน หากพิจารณาในภาพรวมจะพบว่าสาขาวิชาต่าง ๆ ล้วนมีการนำเทคโนโลยีคอมพิวเตอร์มาใช้ประโยชน์ในลักษณะใดลักษณะหนึ่งอยู่แล้ว การเปลี่ยนผ่านเข้าสู่ "มนุษยศาสตร์ดิจิทัล" จึงไม่ใช่เรื่องที่น่าแปลกแต่อย่างใด อย่างไรก็ตาม หากพิจารณาสาขาวิชาอื่น ๆ ที่นำคำว่า "ดิจิทัล" มาขยาย เช่น Digital Social Science, Digital Physics, Digital Mathematics, Digital Biology หรือ Digital Education จะพบว่าคำเหล่านี้ไม่ได้ถูกใช้แพร่หลายเท่ากับคำว่า Digital Humanities และส่วนใหญ่ยังไม่มีนิยามใน วิกิพีเดีย (Wikipedia) โดยมีเพียง Digital Physics เท่านั้นที่มีการให้นิยามไว้ แต่ก็ไม่ได้หมายถึงการใช้เทคโนโลยีคอมพิวเตอร์ในการศึกษาฟิสิกส์โดยตรง หากแต่เป็นแนวคิดทางอภิปรายที่มองว่าเอกภพถูกสร้างขึ้นจากระบบการเชิงคำนวณคล้ายโปรแกรมคอมพิวเตอร์ การใช้คำว่า Digital Humanities จึงมีลักษณะเฉพาะที่แตกต่างจากศาสตร์อื่น ๆ กล่าวคือ มิได้หมายถึงเพียงการนำเทคโนโลยีคอมพิวเตอร์มาใช้เป็นเครื่องมือเท่านั้น หากแต่สะท้อนถึงการเปลี่ยนแปลงเชิงแนวคิดและระเบียบวิธีวิจัย ที่นำไปสู่กระบวนการศึกษาแบบใหม่ซึ่งขยายขอบเขตออกไปจากมนุษยศาสตร์ในรูปแบบดั้งเดิม

ก่อนหน้านี้คำว่า Digital Humanities จะใช้กันแพร่หลาย คำว่า Humanities Computing เป็นคำที่ใช้กับงานที่พูดถึงการนำคอมพิวเตอร์มาใช้ทางด้านมนุษยศาสตร์ ในปีค.ศ. 1946 บาทหลวง Roberto Busa เริ่มสร้างข้อมูลรายการ concordance จากหนังสือของ Thomas Aquinas จนสำเร็จและเผยแพร่ในปีค.ศ. 1951 (Winter, 1999) ความสนใจของเขาช่วงนั้นอยู่ที่การนำคอมพิวเตอร์มาช่วยงานทางมนุษยศาสตร์ ข้อมูล concordance เป็นตัวอย่างของการใช้คอมพิวเตอร์ช่วยแจกแจงข้อมูลคำพร้อมบริบท ทำให้การศึกษาคำและความหมายต่าง ๆ ในบริบททำได้ง่ายขึ้น นอกจากการใช้คอมพิวเตอร์เพื่อช่วยวิเคราะห์ตัวบทสะกดขึ้น ยังมี

ความสนใจเรื่องการทำข้อมูลให้เป็นดิจิทัลและสร้างฐานข้อมูลงานทางมนุษยศาสตร์เพื่อเป็นแหล่งข้อมูลสำหรับการศึกษาโดยใช้คอมพิวเตอร์ได้สะดวกขึ้น

หลังทศวรรษ 1980 ถึง 1990 เมื่อการใช้คอมพิวเตอร์ส่วนบุคคลเริ่มแพร่หลายมากขึ้น และการนำข้อมูลต่าง ๆ เผยแพร่บนอินเทอร์เน็ตในรูปไฮเปอร์เท็กซ์ (Hypertext) เริ่มมีมากขึ้น และช่วงนี้ก็มีการทำฐานข้อมูลหอเอกสารดิจิทัล (Digital archive) ต่าง ๆ มากขึ้น มีการกำหนดมาตรฐานกำกับข้อมูลโดย TEI (Text Encoding Initiative) เริ่มมีการใช้เทคโนโลยีใหม่จากงานด้านภาษาศาสตร์คอมพิวเตอร์เป็นเครื่องมือในการศึกษามากขึ้น พอเริ่มเข้าสู่ศตวรรษที่ 21 มนุษยศาสตร์ดิจิทัลได้รับความสนใจในวงกว้างมากขึ้น หนังสือ A Companion to Digital Humanities (2004) ทำให้คำนี้เป็นที่รู้จักกว้างขวาง และบทความในหนังสือนอกจากจะกล่าวถึงความเป็นมาของ Humanities Computing จากอดีตมาถึงปัจจุบัน ยังทำให้เห็นว่า Digital Humanities ไม่ได้เป็นเพียงการจัดเก็บและใช้ข้อมูลทางมนุษยศาสตร์เท่านั้น แต่ได้เสนอวิธีใหม่ ๆ รวมถึงการตั้งคำถามใหม่ ๆ ที่ไม่สามารถทำได้ในอดีต เกิดการประยุกต์ใช้ในการศึกษาเรื่องใหม่ เช่น การศึกษาสไตล์การเขียน การระบุตัวผู้เขียน การเปลี่ยนแปลงทางวัฒนธรรม และทำให้เห็นความเป็นสหวิทยาการ (Interdisciplinary) ของสาขามากขึ้น

2. นวัตกรรมสำหรับงานมนุษยศาสตร์

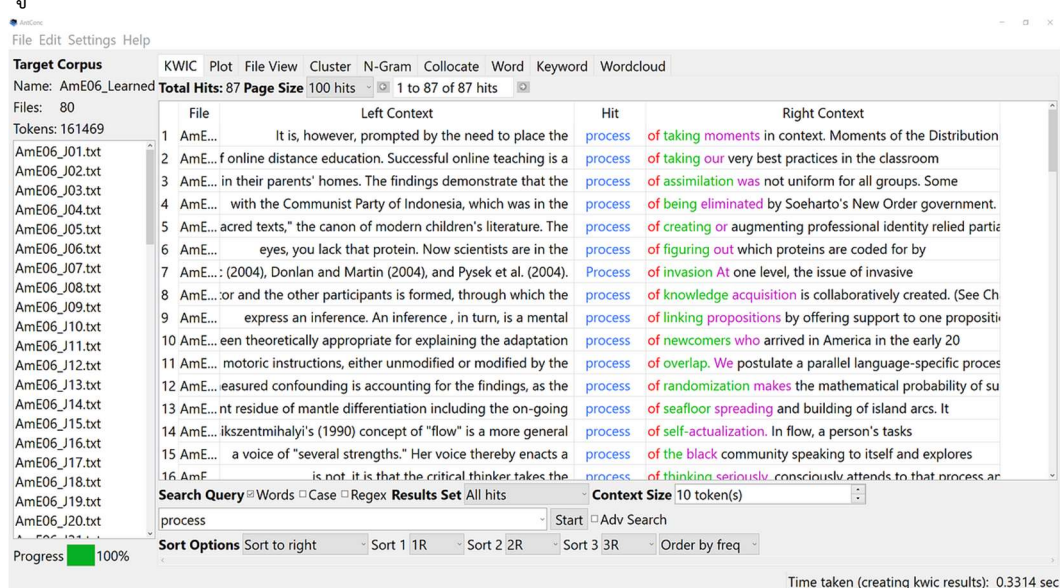
มนุษยศาสตร์ดิจิทัลนำมาซึ่งนวัตกรรมหรือวิธีการศึกษาใหม่ที่ขยายขอบเขตการศึกษางานทางมนุษยศาสตร์ ช่วยให้สามารถศึกษาเรื่องใหม่ ๆ ที่ทำได้ยากในอดีต เช่น การค้นหาตัวอย่างการใช้คำที่ต้องการทั้งหมดในหนังสือแต่ปัจจุบันสามารถดึงตัวอย่างเหล่านี้ได้อย่างง่ายดาย หรือการที่ตัวละครหนึ่งมีปฏิสัมพันธ์กับใครบ้างมากนักน้อยเพียงใดก็สามารถแสดงเป็นภาพนิทัศน์ให้เห็นความต่างได้ชัดเจนขึ้น จากจุดเริ่มต้นที่มีผู้เห็นประโยชน์จากการใช้คอมพิวเตอร์มาช่วยจัดการข้อมูลและประมวลผลข้อมูล จำเป็นต้องทำข้อมูลต่าง ๆ ให้เป็นดิจิทัลเพื่อให้ใช้งานกับคอมพิวเตอร์ได้ จนมาสู่ยุคปัจจุบันที่ข้อมูลส่วนใหญ่เป็นดิจิทัลแต่กำเนิดและข้อมูลมีจำนวนมากมายมหาศาล ความก้าวหน้าทางเทคโนโลยีที่พัฒนาขึ้นอย่างต่อเนื่องเป็นตัวเร่งให้การเติบโตของมนุษยศาสตร์ดิจิทัลได้รับความสนใจจากนักวิชาการสายมนุษยศาสตร์มากขึ้น ตัวอย่างนวัตกรรมที่ใช้ในงานทางมนุษยศาสตร์มีดังนี้

2.1 คลังข้อมูลภาษา

คลังข้อมูลภาษาหรือ Corpora เป็นนวัตกรรมตั้งแต่ในยุคแรก ๆ ของการศึกษาทางมนุษยศาสตร์ โดยเฉพาะการศึกษาผ่านข้อมูลภาษา การรวบรวมข้อมูลมาแสดงในรูปของ concordance เป็นการแสดงผลที่ช่วยให้เราเห็นคำที่ต้องการศึกษาและบริบทการปรากฏของคำโดยที่เราสามารถจัดเรียงข้อมูลตามเงื่อนไขที่กำหนดได้ ช่วยให้เห็นภาพรวมของการใช้คำและวิเคราะห์ความหมายและการใช้คำนั้นได้ง่ายขึ้น จึงไม่แปลกที่ความคิดที่จะทำ concordance ของหนังสือสำคัญมีมาตั้งแต่ก่อนจะมีโปรแกรม Concordance ใช้กันแพร่หลาย บาทหลวง Roberto Busa สร้าง concordance ด้วยการเขียนประโยคที่มีคำที่ต้องการศึกษาลงในบัตรรายการ 10,000 ใบสำหรับใช้ในงานวิทยานิพนธ์ของตน (Winter, 1999) หลังจากนั้น ความต้องการหาวิธีการที่จะช่วยสร้างรายการ concordance ให้โดยอัตโนมัติ ทำให้เขาเดินทางมาพบ Thomas J. Watson

ที่บริษัท IBM และเริ่มใช้บัตรเจาะรูเพื่อเตรียมข้อมูลการทำ concordance ด้วยเครื่อง บาทหลวง Busa ยิ่งวาดฝันถึงงานที่คนรุ่นหลังจะสามารถทำได้ เช่น การระบุคำสรรพนามอ้างอิงใครหรืออะไรก่อนหน้า การหาหน่วยภาษาที่เป็นชื่อเฉพาะ คำประสม การระบุหมวดคำ ความหมายคำ และอื่น ๆ เพื่อให้การวิเคราะห์ตัวบทเป็นไปได้ง่ายขึ้น ซึ่งความฝันในตอนนั้นก็ได้กลายเป็นจริงในปัจจุบัน

โปรแกรม Concordance ปัจจุบันอย่าง AntConc², LancsBox³ ช่วยให้เราทำงานพื้นฐานอย่าง การหาความถี่คำ เรียงข้อมูลคำพร้อมบริบท (ภาพที่ 1) แจกแจงคำที่มักปรากฏร่วมกัน หารายการคำสำคัญในข้อมูล และด้วยเครื่องมือการประมวลผลภาษาที่ซับซ้อนขึ้น ทำให้โปรแกรมอย่าง Word Sketch⁴ ขยายความสามารถในการหาความสัมพันธ์ทางวากยสัมพันธ์ระหว่างคำและวิเคราะห์จัดกลุ่มความหมายของคำได้ โปรแกรมเหล่านี้จึงเป็นเครื่องมือที่ช่วยให้แก่นมนุษยศาสตร์ศึกษาจากข้อมูลภาษาขนาดใหญ่ได้ การศึกษาทางวรรณกรรมเองก็ได้ใช้ประโยชน์จากคลังข้อมูลภาษา ด้วยการรวบรวมข้อมูลงานเขียนทั้งหมดของนักเขียนทำให้สามารถศึกษาเนื้อหาและรูปแบบการใช้ภาษาของนักเขียนนั้นได้อย่างครอบคลุม (Biber, 2011) คลังข้อมูลภาษาเมื่อใช้ร่วมกับโปรแกรม Concordance จึงเป็นนวัตกรรมวิธีพื้นฐานสำหรับการใช้การศึกษาที่ต้องอาศัยตัวบทเป็นข้อมูลอย่างเช่น งานทางภาษาศาสตร์ งานวรรณกรรมวิเคราะห์ เป็นต้น



ภาพที่ 1 ตัวอย่าง Concordance จากการใช้โปรแกรม AntConc

2.2 หอเอกสารดิจิทัล

นอกจากคลังข้อมูลภาษาที่เป็นการรวบรวมข้อมูลมาเพื่อเป็นตัวแทนของภาษาที่ต้องการศึกษา เช่น ภาษาของนักเขียนคนใดคนหนึ่ง ภาษากฎหมาย ภาษาวิชาการ ภาษาพูด ฯลฯ การเก็บรวบรวมข้อมูลอีกลักษณะคือการเก็บข้อมูลเป็นหอเอกสารดิจิทัล (Digital archives) ในหัวเรื่องใดหัวเรื่องหนึ่งเพื่อให้ผู้สนใจเรื่องนั้นเข้าถึงข้อมูลศึกษาเรื่องนั้นได้ง่าย ในงานด้านมนุษยศาสตร์ดิจิทัล ความสนใจสร้างหอเอกสารดิจิทัล

² <https://www.laurenceanthony.net/software/antconc/>

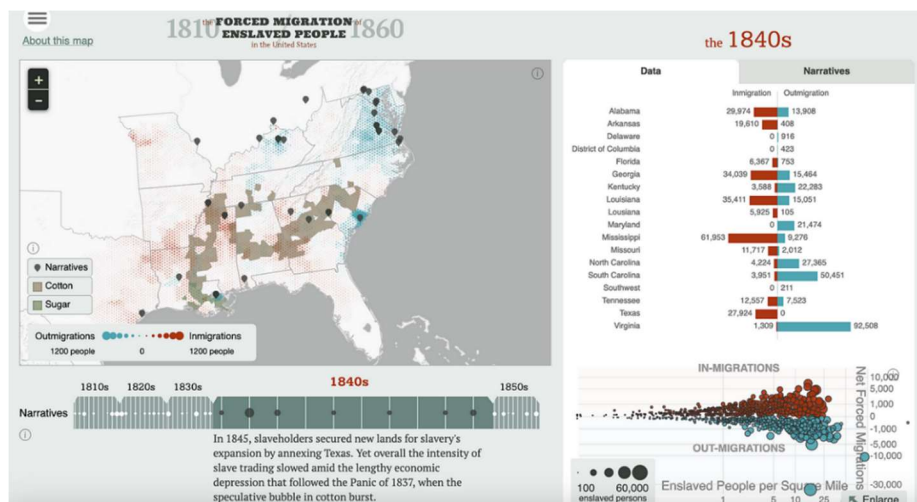
³ <https://lancsbox.lancs.ac.uk/>

⁴ <https://www.sketchengine.eu/>

เกิดขึ้นเพื่อรวบรวมข้อมูลเอกสารต่าง ๆ ในหัวเรื่องที่น่าสนใจเก็บเป็นข้อมูลดิจิทัลเพื่อให้เข้าถึงง่ายขึ้น ข้อมูลอาจเป็นเอกสารกระดาษที่สแกนหรือถ่ายภาพเก็บไว้ เป็นภาพวาด ภาพเขียน ประติมากรรม โน้ตดนตรี ฯลฯ ที่มีคุณค่าในการศึกษาทางมนุษยศาสตร์ ตัวอย่างหอเอกสารดิจิทัล เช่น African American Women Writers of the 19th Century (Divisions, n.d.) รวบรวมข้อมูลงานนักเขียนสตรีชาวอเมริกันผิวสีเพื่อให้เข้าถึงและค้นหางานได้ง่าย Jane Austen Fiction Manuscripts (n.d.) เป็นอีกหอเอกสารหนึ่งที่เก็บข้อมูลต้นฉบับงานเขียนของ Jane Austen เก็บเป็นภาพจากต้นฉบับทำให้เห็นถึงร่องรอยการแก้ไขระหว่างเขียนด้วย และสามารถสืบค้นคำต่าง ๆ ในเอกสารได้

หอเอกสารดิจิทัลในปัจจุบันไม่จำเป็นต้องมีลักษณะนิ่ง (Static) ที่จำกัดการใช้งานเพียงแค่การอ่านหรือการเข้าถึงไฟล์เอกสารเท่านั้น ด้วยพัฒนาการของเทคโนโลยีสารสนเทศ ทำให้สามารถนำเสนอเนื้อหาในรูปแบบสื่อประสม (Multimedia) ซึ่งประกอบด้วยเสียงและภาพเคลื่อนไหวได้ ส่งผลให้การนำเสนอข้อมูลมีความหลากหลายและน่าสนใจมากยิ่งขึ้น ตัวอย่างเช่น โครงการ The World of Dante (n.d.) ซึ่งเป็นหอเอกสารดิจิทัลบนเว็บ ที่เปิดให้ผู้ใช้เข้าถึงและสืบค้นตัวบทวรรณกรรม พร้อมทั้งแสดงแผนภาพและภาพประกอบที่มีคุณสมบัติแบบปฏิสัมพันธ์ (Interactive) ผู้ใช้สามารถคลิกเลือกดูเนื้อหาแต่ละส่วนหรือสำรวจข้อมูลตามลำดับเวลาได้ (Timeline navigation) ในทำนองเดียวกัน โครงการ Shakespeare's Circuits (n.d.) ก็เป็นอีกตัวอย่างหนึ่งในสาขามนุษยศาสตร์ดิจิทัลที่นำเสนอการเผยแพร่และการรับรู้ผลงานของเชกสเปียร์ในระดับโลก ผ่านการจัดทำแผนที่แบบมีปฏิสัมพันธ์ (Interactive mapping) และการนำเสนอเชิงเวลา (Temporal mapping) ซึ่งสะท้อนถึงแนวทางใหม่ในการจัดการเอกสารดิจิทัลและการเข้าถึงองค์ความรู้ในยุคดิจิทัล

โครงการ American Panorama (2020) นำเสนอแผนที่แบบปฏิสัมพันธ์ที่เปิดโอกาสให้ผู้ใช้สามารถสำรวจการเปลี่ยนแปลงที่เกิดขึ้นในประเทศสหรัฐอเมริกาในประเด็นต่าง ๆ ตามช่วงเวลาได้อย่างมีประสิทธิภาพ ตัวอย่างเช่น ภาพที่ 2 แสดงแผนที่แบบปฏิสัมพันธ์ที่นำเสนอจำนวนการย้ายถิ่นฐานซึ่งเป็นผลจากการค้าทาสในแต่ละทศวรรษ หอเอกสารดิจิทัลจึงเป็นนวัตกรรมที่ทำให้มีข้อมูลกลางสำหรับผู้สนใจศึกษาหัวเรื่องนั้นได้



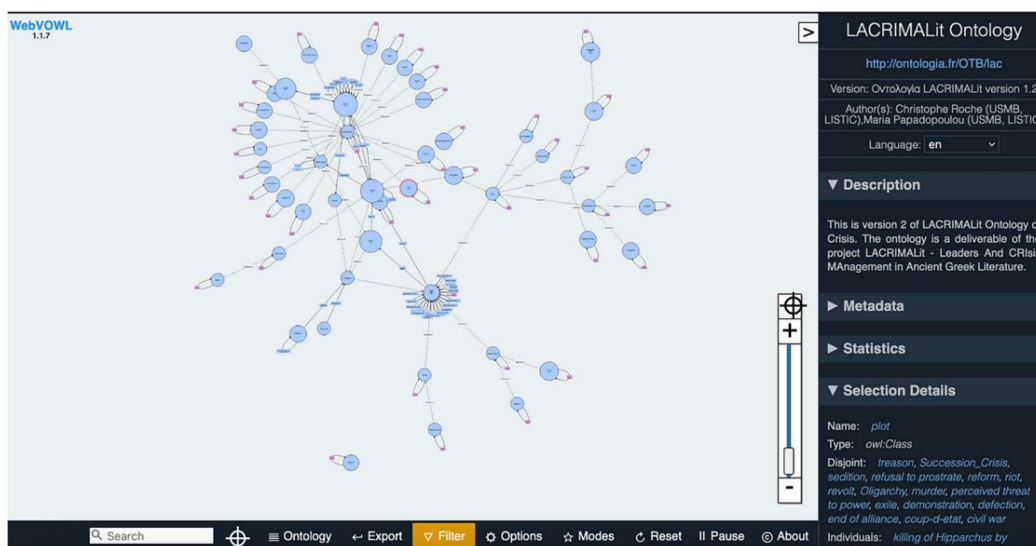
ภาพที่ 2 แผนที่แบบปฏิสัมพันธ์ในโครงการ American Panorama

นอกจากการสร้างหอเอกสารดิจิทัลที่เป็นแหล่งศึกษาค้นคว้าทางด้านมนุษยศาสตร์แล้ว การรวบรวมข้อมูลจำนวนมากที่เกิดขึ้นในช่วงระยะเวลาต่าง ๆ เช่น โครงการ Google Books Ngram Viewer (Michel et al., 2010) ที่รวบรวมข้อมูลหนังสือจากอดีตถึงปัจจุบันเป็นฐานข้อมูลให้สืบค้นหาคำและวลีต่าง ๆ ที่ปรากฏในฐานข้อมูลได้ ทำให้สามารถใช้วิเคราะห์การเกิดขึ้นและแนวโน้มการเปลี่ยนแปลงทางความคิดและวัฒนธรรมผ่านการเปลี่ยนแปลงของภาษาตามช่วงเวลาต่าง ๆ เกิดการศึกษาวัฒนธรรมในมิติใหม่ที่ว่า Culturomics (2024)

2.3 เมทาดาตา

ความสนใจในการสร้างและใช้ประโยชน์จากคลังข้อมูลภาษาและหอเอกสารดิจิทัล ทำให้ต้องสนใจนวัตวิธียุคใหม่ที่เกี่ยวข้องกับเรื่องการจัดเก็บและค้นคืนข้อมูลมาใช้ด้วย บางโครงการก็อาจต้องการใช้เครื่องมือเกี่ยวกับการแสดงผลภาพหรือแผนที่เพื่อให้ตอบโต้กับผู้ใช้ได้ และในการจัดเก็บข้อมูล นอกจากตัวข้อมูลงานโดยตรงแล้ว ข้อมูลบริบทอื่น ๆ ที่เกี่ยวข้องกับตัวงานเองก็มีความสำคัญสำหรับใช้จัดระบบและค้นคืนข้อมูลด้วย ข้อมูลเหล่านี้เรียกว่าเมทาดาตา (Metadata) ข้อมูลบริบท เช่น ชื่อผู้เขียน ปีที่เขียน ประเภทงาน ฯลฯ จะช่วยคัดกรองข้อมูลที่ตรงกับเงื่อนไขที่ต้องการได้

เมทาดาตาในที่นี้ ยังหมายรวมถึงข้อมูลออนโทโลยีที่กำหนดโครงสร้างความรู้ในหัวเรื่องต่าง ๆ ด้วย เพราะเป็นข้อมูลที่ทำให้สามารถเข้าใจความเชื่อมโยงข้อมูลและสืบค้นผ่านออนโทโลยีนั้นได้ โครงการ O4DH: Ontologies for Digital Humanities (AdminO4DH, 2019) เป็นโครงการเพื่อสร้างและใช้ออนโทโลยีที่เป็นเหมือนผังความรู้ (Knowledge graph) ที่ช่วยให้เข้าใจข้อมูลงานทางมนุษยศาสตร์ดิจิทัลมากขึ้น



ภาพที่ 3 ผังความรู้แบบปฏิสัมพันธ์ ในเรื่อง “วิกฤต” (Crisis)⁵

⁵ <https://service.tib.eu/webvowl/#iri=http://ontologia.fr/OTB/lac.owl>

เมทาเดตาอาจเป็นข้อมูลบริบทที่กำหนดร่วมกันเพื่อให้ใช้งานร่วมกันได้ เช่น โครงการ Pelagios (Pelagios Network, n.d.) ที่สร้างเครือข่ายทางมนุษยศาสตร์ดิจิทัลให้สามารถแลกเปลี่ยนข้อมูลกันได้โดยใช้ข้อมูลเชื่อมโยงแบบเปิด (Open linked data) แต่ละหน่วยงานหรือสถาบันที่มีงานทางมนุษยศาสตร์สามารถกำกับข้อมูลความหมาย (Semantic annotation) ของงานต่าง ๆ เช่น กำกับข้อมูล what-who-when-where เพื่อบอกว่าเป็นวัตถุอะไร ถูกพบโดยใคร ที่ไหนเมื่อไร และทำให้ข้อมูลนี้เป็นข้อมูลเปิดเข้าถึงได้ด้วย URI เฉพาะ (Ricogoto, n.d.) เมทาเดตาจึงเป็นนวัตกรรมวิธีที่ช่วยให้การเข้าถึงข้อมูลไม่ว่าจะเป็นข้อมูลภาษาหรือข้อมูลอื่นใดทำได้อย่างมีประสิทธิภาพมากขึ้น

นอกจากข้อมูลบริบท บางโครงการก็อาจต้องการใช้ข้อมูลที่วิเคราะห์เพิ่มเติมใส่ลงไปในตัวงานนั้นด้วย เช่น การกำกับหมวดคำ การระบุชื่อเฉพาะ เพราะในช่วงเริ่มต้น คอมพิวเตอร์ยังประมวลผลด้วยตัวเองไม่ได้ ต้องอาศัยข้อมูลต่าง ๆ ที่มีการระบุไว้อย่างชัดเจนจึงจะค้นข้อมูลตรงตามเงื่อนไขได้ เมื่อต้องมีการกำกับข้อมูลอื่น ๆ เพิ่มเติมลงไป แต่ละโครงการก็อาจกำหนดวิธีการกำกับข้อมูล (Annotation) ตามที่ออกแบบเอง ปัญหาที่ตามมาคือการใช้งานแลกเปลี่ยนข้อมูลร่วมกันเป็นไปได้ยาก อีกทั้งเครื่องมือหรือโปรแกรมที่จะทำงานค้นข้อมูลก็ไม่สามารถใช้ร่วมกันได้ จึงมีโครงการที่คิดกำหนดมาตรฐานการกำกับข้อมูลต่าง ๆ คือโครงการ Text Encoding Initiative (TEI Consortium, eds., 2024) ที่ในปัจจุบัน นิยมเลือกใช้ภาษา XML เป็นมาตรฐานสำหรับการกำกับข้อมูล (Markup) โดยมีการกำหนดแท็กและโครงสร้างของข้อมูลให้เป็นไปในทิศทางเดียวกัน เพื่อส่งเสริมการแลกเปลี่ยนและการประมวลผลข้อมูลอย่างเป็นระบบ ตัวอย่างโครงการที่นำแนวทางการกำกับข้อมูลตามมาตรฐานของ Text Encoding Initiative (TEI) มาใช้ ได้แก่ โครงการ British National Corpus โครงการ Emily Dickinson Archive และโครงการ Victorian Women Writers' Project ซึ่งล้วนเป็นโครงการด้านมนุษยศาสตร์ดิจิทัลที่เน้นการจัดการและนำเสนอข้อมูลตัวบทอย่างมีโครงสร้าง

การนำเทคโนโลยีคอมพิวเตอร์เข้ามาจัดการข้อมูลทางมนุษยศาสตร์เป็นนวัตกรรมวิธีที่เข้ามาและส่งผลให้การเข้าถึงข้อมูลและจัดการข้อมูลสะดวกและง่ายมากขึ้น สามารถศึกษาข้อมูลได้ครบถ้วนครอบคลุมทุกตัวอย่างการใช้ได้ แทนที่จะต้องอาศัยวิธีการอ่านและหาตัวอย่างที่ต้องการเอาเอง ยิ่งเมื่อข้อมูลมีขนาดใหญ่มากขึ้นเรื่อย ๆ การศึกษาในรูปแบบเดิมที่จะต้องอ่านอย่างละเอียดทั้งหมดก็เป็นไปได้ยากขึ้น จำต้องอาศัยการมองข้อมูลจากภาพรวมหรือมุมกว้างหรือที่เรียกว่า Distant Reading (Ossom-Williamson & Rambsy, 2021; Underwood, 2016)

2.4 สถิติกับงานมนุษยศาสตร์

วิธีการสถิติแม้เป็นวิธีการพื้นฐานที่ใช้กับศาสตร์อื่น ๆ โดยเฉพาะศาสตร์ทางด้านวิทยาศาสตร์ แต่สำหรับมนุษยศาสตร์ที่พื้นฐานการศึกษาเป็นการศึกษาเชิงคุณภาพ ใช้การวิเคราะห์วิจารณ์เพื่อเข้าถึงคุณค่าและความคิดมนุษย์ผ่านงานสร้างสรรค์ต่าง ๆ แต่เมื่องานมนุษยศาสตร์เริ่มมีข้อมูลจำนวนมากขึ้น การใช้สถิติจึงเป็นนวัตกรรมวิธีที่ถูกนำมาใช้เพื่อช่วยประมวลผลข้อมูล เพราะการดูภาพรวมข้อมูลขนาดใหญ่ สถิติแบบพรรณนา

อย่างค่าเฉลี่ย ส่วนเบี่ยงเบนมาตรฐาน ความถี่ และการกระจายตัว ข้อมูลสามารถใช้อธิบายลักษณะของข้อมูลนั้น ๆ ได้ นอกจากนี้ การใช้วิธีการทางสถิติยังช่วยแสดงให้เห็นความสัมพันธ์ระหว่างตัวแปรต่าง ๆ ในข้อมูล ตัวอย่างโครงการมนุษยศาสตร์ดิจิทัลที่ใช้ข้อมูลเชิงสถิติช่วยให้เห็นมิติต่างๆ ของข้อมูล เช่น โครงการ Digital Tolkien Project (Tauber, n.d.) และ Tolkien's books analysed (LOTR Project, n.d.) ที่นอกจากให้ค้นหาข้อมูลในหนังสือได้แล้ว ยังแสดงผลข้อมูลพื้นฐานทางสถิติ เช่น จำนวนคำในหนังสือแต่ละเล่ม แต่ละบท ตำแหน่งหน้าหนังสือที่พบคำต่าง ๆ ในหนังสือแต่ละเล่ม จำนวนคำใหม่ในแต่ละบท การปรากฏร่วมกันของตัวละครต่าง ๆ ในหน้าเดียวกันของหนังสือแต่ละเล่ม ข้อมูลต่าง ๆ นี้จะช่วยให้ผู้อ่านเห็นภาพมุมกว้างของหนังสือในแบบที่การอ่านตามปกติจะไม่เห็น

วิธีการทางสถิติอื่น ๆ ที่เกี่ยวข้องกับการอธิบายความสัมพันธ์ของตัวแปรต่าง ๆ หรือใช้สร้างแบบจำลองอธิบายความสัมพันธ์ของตัวแปร เป็นวิธีการทางสถิติที่ใช้ในการศึกษาของศาสตร์อื่น ๆ ก็สามารถนำมาใช้ในงานมนุษยศาสตร์เพื่อการศึกษาวิจัยได้ สิ่งที่น่าสนใจที่นักมนุษยศาสตร์ต้องทำความเข้าใจคือลักษณะข้อมูลที่มีในงานของตัวเองว่าเป็นข้อมูลลักษณะไหน (Aroonmanakun, 2022) วิธีการทางสถิติแบบต่าง ๆ ที่เหมาะสมกับการนำมาใช้อธิบายข้อมูลที่มีว่ามีวิธีการใดได้บ้างและคำนวณอย่างไร (Aroonmanakun, 2023a, 2023b) ซึ่งบางครั้งก็จำเป็นต้องแปลงหรือสังเคราะห์ข้อมูลให้อยู่ในรูปที่เหมาะสมสำหรับนำไปคำนวณในทางสถิติต่อไปได้ เช่น การแปลงตัวบทเป็นชุดตัวเลขหรือเวกเตอร์ที่ใช้แทนข้อมูลตัวบทนั้น (Aroonmanakun, 2022)

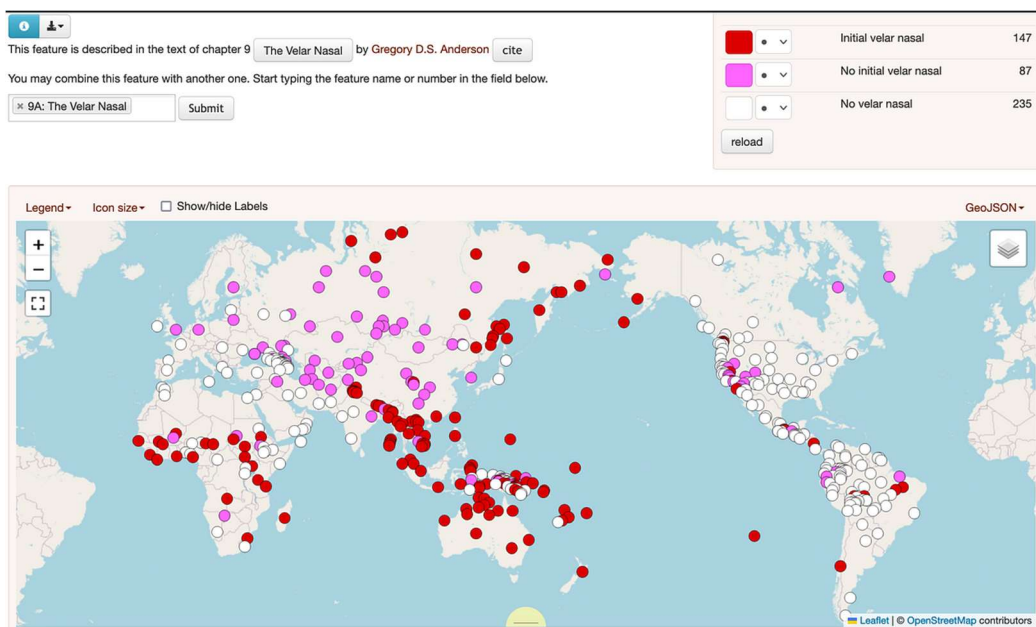
เมื่อสามารถสังเคราะห์ข้อมูลให้อยู่ในรูปของเวกเตอร์ได้ ก็ทำให้ข้อมูลที่ได้เป็นข้อมูลแบบหลายตัวแปร ซึ่งสามารถใช้วิธีการทางสถิติแบบต่าง ๆ ที่เหมาะกับงานนั้นได้ งานของ Biber (1993) เป็นตัวอย่างงานที่อาศัยการแปลงข้อมูลภาษาผ่านการวิเคราะห์หลักคุณทางภาษา (Linguistic features) จนได้ชุดข้อมูลหลักคุณทางภาษาหรือเวกเตอร์ที่นำมาวิเคราะห์ความสัมพันธ์ภายในด้วยวิธีการทางสถิติที่เรียกว่า Multidimensional Analysis โดย Biber เลือกใช้สถิติ Factor Analysis เพื่อลดขนาดข้อมูลให้เหลือจำนวนมิติหรือ dimension น้อยลง และจะเห็นว่าหลักคุณทางภาษาใดที่เข้าพวกกันและมีผลเชิงบวกหรือลบกับแต่ละมิตินั้น Biber ใช้วิธีการนี้เพื่อวิเคราะห์ประเภทตัวบทหรือ genre ทำให้เห็นความแตกต่างระหว่างตัวบทประเภทต่าง ๆ อย่างเป็นรูปธรรมได้

นอกจากงานของ Biber งานด้าน Stylometry ที่สนใจศึกษาเรื่อง Authorship ของงานเขียนคือหาคำตอบว่างานที่สงสัยนั้นเป็นงานเขียนของใคร ก็สามารถใช้อ้างอิงสังเคราะห์มาเป็นเวกเตอร์ได้ ตัวอย่างงานของ Binongo (2003) ต้องการหาคำตอบว่าหนังสือพอมด Oz เล่มที่ 15 นั้นใครเป็นผู้แต่งกันแน่ระหว่างนักเขียนสองคน เขาหาคำตอบโดยดูชุดคำไวยากรณ์ที่พบมากที่สุด 50 คำเป็นตัวแทนของหนังสือ คือใช้คำไวยากรณ์ 50 คำเพื่อสร้างเวกเตอร์แทนหนังสือแต่ละเล่มของนักเขียนทั้งสองคน จากนั้นใช้สถิติ Principal Component Analysis เพื่อลดทอนข้อมูล 50 มิติให้เหลือ 2 มิติ แล้วจึงเห็นว่าผลที่ได้จากหนังสือของนักเขียนสองคนนี้แตกต่างกัน และเมื่อนำหนังสือที่สงสัยมาสังเคราะห์ข้อมูลชุดคำไวยากรณ์นำมาเปรียบเทียบกับผลที่ได้ของนักเขียนทั้งสองก็จะเห็นว่าใครควรเป็นผู้เขียนหนังสือเล่มนี้

วิธีการทางสถิติอีกแบบที่นิยมใช้ในการวิเคราะห์ตัวบทคือ Topic Modeling (Jelodar et al., 2017) ซึ่งอาศัยการสังเคราะห์ข้อมูลเวกเตอร์จากตัวบท อาศัยการปรากฏของคำเนื้อหาในแต่ละส่วนของตัวบทเป็นข้อมูลพื้นฐานในการสร้างข้อมูลเวกเตอร์จากตาราง term-document และใช้อัลกอริทึมอย่าง Latent Dirichlet Allocation (LDA) ซึ่งเป็นหนึ่งใน unsupervised learning เพื่อใช้จำแนกหัวข้อ (Topic) ในชุดข้อมูลนั้น แนวคิดหลักของ LDA คือการสันนิษฐานว่าแต่ละเอกสารในชุดข้อมูลประกอบด้วยหัวข้อต่าง ๆ ซึ่งแต่ละหัวข้อจะมีกลุ่มคำศัพท์เฉพาะของหัวข้อนั้น เทคนิคนี้จะทำการวิเคราะห์ว่าคำแต่ละคำในเอกสารมีความเกี่ยวข้องกับหัวข้อใดด้วยน้ำหนักเท่าใด จากนั้นจึงสามารถจัดกลุ่มเอกสารตามหัวข้อที่เกี่ยวข้องได้ โมเดล LDA จึงเป็นเครื่องมือที่มีประโยชน์มากในการวิเคราะห์ข้อมูลที่เป็นข้อความ โดยเฉพาะในการค้นหาและจำแนกหัวข้อหลัก ๆ จากชุดข้อมูลข้อความที่ใหญ่และซับซ้อน

2.5 การแสดงภาพนิทัศน์จากข้อมูล

ในการแสดงผลการวิเคราะห์ข้อมูล นอกเหนือจากการใช้วิธีการทางสถิติเพื่อคำนวณผลของข้อมูลออกมาเป็นตัวเลข การแสดงผลด้วยภาพก็เป็นอีกวิธีการหนึ่งของการแสดงผล การแสดงผลเป็นแผนภูมิและกราฟพื้นฐาน เช่น กราฟเส้น กราฟแท่ง กราฟวงกลม ช่วยให้เห็นภาพข้อมูลชัดเจน นอกจากกราฟต่าง ๆ การแสดงผลภาพยังพัฒนาต่อมาเป็นการแสดงที่ซับซ้อนขึ้น เช่น การแสดงข้อมูลเชื่อมโยงกับแผนที่ดังตัวอย่างโครงการ World Atlas of Language Structures (Dryer & Haspelmath, 2013) ที่แสดงว่าภาษาที่สนใจมีพุดอยู่ที่ใดให้เห็นผ่านแผนที่โลก หรือจะเลือกแสดงเฉพาะเสียงในภาษาว่ามีภาษาใดใช้เสียงนั้นและอยู่ที่ส่วนใดของโลกก็ได้



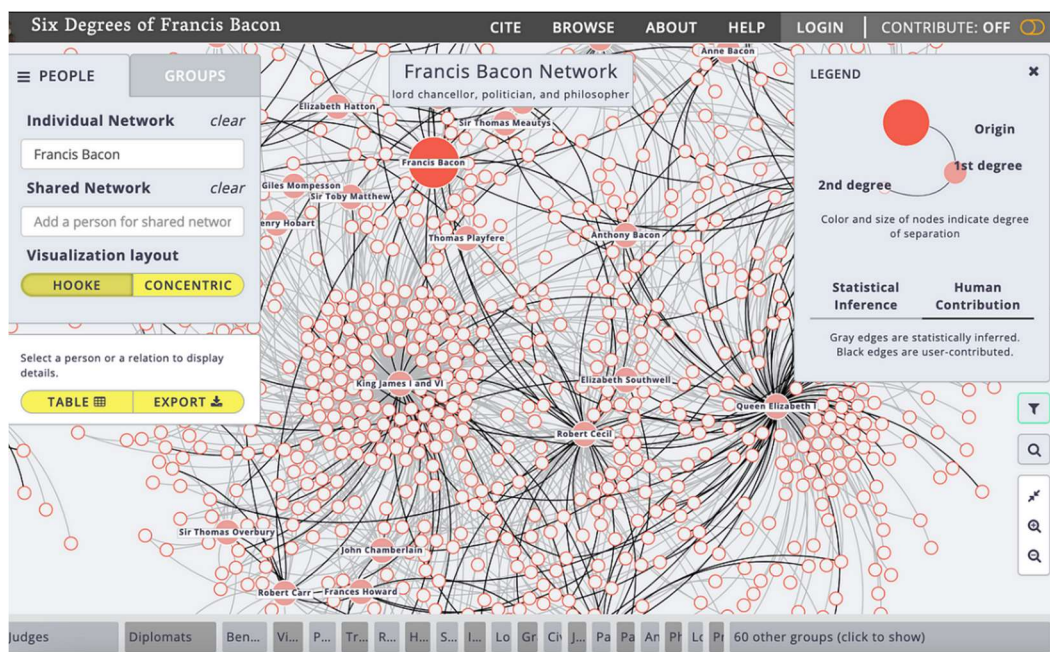
ภาพที่ 4 แผนที่แสดงตำแหน่งภาษาที่พบในที่ต่าง ๆ⁶

⁶ <https://wals.info/>

นอกจากการแสดงผลผ่านแผนที่ตามช่วงเวลา (Temporal map) ดังตัวอย่างโครงการ Shakespeare's Circuits (n.d.) และโครงการ American Panorama (n.d.) ตามที่กล่าวมาแล้ว ยังมี การแสดงข้อมูลเชิงปริภูมิ (3D space) เช่น โครงการ Digital Karnak (n.d.) ที่สร้างแบบจำลองสามมิติเพื่อให้ เห็นสภาพวิหารในยุคอียิปต์โบราณและนำเสนอเป็นวิถีทัศน์ให้เข้าเยี่ยมชมได้

การแสดงภาพความสัมพันธ์ผ่านเครือข่าย (Network) เช่น โครงการ Six Degrees of Francis Bacon (n.d.) ที่แสดงภาพนิทัศน์การติดต่อระหว่างบุคคลที่มีชื่อเสียงอย่าง Francis Bacon, William Shakespeare, และคนอื่น ๆ ในช่วงศตวรรษที่ 16 และ 17 โดยแสดงให้เห็นเป็นภาพโยงการติดต่อระหว่าง บุคคลนั้นกับบุคคลอื่น ๆ ซึ่งสามารถกดดูลักษณะการติดต่อและช่วงเวลาการติดต่อได้ (ภาพที่ 5)

การแสดงก้อนคำ (Word cloud) เป็นอีกวิธีการหนึ่งของการแสดงผลจำนวนการปรากฏของคำ ในข้อมูล ยิ่งคำนั้นพบมากในข้อมูล คำที่แสดงก็จะมีขนาดใหญ่ตามไปด้วย การแสดงก้อนคำจึงเป็นอีกนวัตวิธ ีที่นิยมใช้แสดงให้เห็นเนื้อหาคำในข้อมูลที่ศึกษา มีเครื่องมือจำนวนมากที่ช่วยสร้างก้อนคำได้ เช่น WordClouds.com⁷, WordArts⁸, TagCrowd⁹ เป็นต้น ภาพที่ 6 เป็นตัวอย่างภาพก้อนคำจากหนังสือ



ภาพที่ 5 ภาพแสดงเครือข่ายการติดต่อของ Francis Bacon¹⁰

⁷ <https://www.wordclouds.com/>

⁸ <https://wordart.com/create>

⁹ <https://tagcrowd.com/>

¹⁰ http://www.sixdegreesoffrancisbacon.com/?ids=10000473&min_confidence=60&type=network



ภาพที่ 6 ภาพแสดงก้อนคำจากหนังสือ “A Companion to Digital Humanities”

การแสดงผลภาพนิทัศน์เหล่านี้จึงเป็นนวัตกรรมวิธีที่เข้ามาช่วยในการศึกษาทางมนุษยศาสตร์ การแสดงผลข้อมูลเป็นภาพนิทัศน์ช่วยให้ข้อมูลที่ซับซ้อนหรือมีปริมาณมากเป็นสิ่งที่สามารถเข้าใจได้ง่ายขึ้น อีกทั้งทำให้สามารถสำรวจข้อมูลได้หลายมิติ เช่น การแสดงความสัมพันธ์ของข้อมูลทางเวลา พื้นที่ หรือแม้แต่ความสัมพันธ์ทางสังคม ซึ่งมีประโยชน์ในการวิเคราะห์ข้อมูลที่ซับซ้อน

2.6 แบบจำลองตัวแทน

นวัตกรรมวิธีอีกอย่างคือแบบจำลองตัวแทนหรือ Agent-Based Model (ABM) ซึ่งเป็นแบบจำลองทางคอมพิวเตอร์ (Computer simulation) ใช้จำลองสถานการณ์อย่างหนึ่งอย่างใดที่มีตัวแทนต่าง ๆ ในแบบจำลองนั้น แต่ละตัวแทนมีพฤติกรรมตามกฎที่ระบุในแบบจำลองว่า เมื่อพบสภาพแวดล้อมหนึ่งหรือมีปฏิสัมพันธ์กับตัวแทนอื่นจะมีพฤติกรรม ให้กระทำหรือไม่กระทำอะไร ABM เป็นเทคโนโลยีที่หลายศาสตร์ใช้ในการศึกษาพฤติกรรมของมนุษย์ คือใช้จำลองสถานการณ์ให้เห็นว่าภายใต้เงื่อนไขที่กำหนดจะเกิดเหตุการณ์อะไรขึ้นได้บ้าง เช่น พฤติกรรมการบริโภคของผู้ซื้อ

งานทางมนุษยศาสตร์ดิจิทัล ก็มีผู้เสนอให้ใช้ ABM ศึกษางานทางมนุษยศาสตร์ได้ เช่น Gavin (2014) เสนอให้ใช้ ABM กับการศึกษาเหตุการณ์ทางประวัติศาสตร์ ABM ช่วยให้ภาพความเข้าใจว่า ถ้าหากเกิดสิ่งนี้แล้ว ผลที่จะเกิดตามมาจะเป็นอย่างไร การศึกษาทางประวัติศาสตร์แต่เดิมเราใช้ข้อมูลเชิงสถิติคือข้อมูลเท่าที่มีแล้วพยายามอธิบายว่าเกิดอะไรขึ้นโดยสันนิษฐานว่าเพราะเหตุใด แต่หากใช้แบบจำลองตัวแทนแล้ว เราสามารถจำลองดูผลที่จะเกิดขึ้นได้จากปฏิสัมพันธ์ของตัวแทนต่าง ๆ ในแบบจำลองนั้น ทำให้ทดลองสร้างแบบจำลองเพื่ออธิบายเหตุการณ์ที่เกิดขึ้นได้ดีกว่า Gasmi et. al. (2014) เสนอการใช้แบบจำลองตัวแทนเพื่อจำลองเหตุการณ์ในอดีตเช่นกัน โดยทดสอบว่าผลของการจัดการจากตัวแทนหรือ agent ต่าง ๆ ที่

เกี่ยวข้องในช่วงเวลา นั้นส่งผลต่อเหตุการณ์น้ำท่วมใหญ่ในเวียดนามเมื่อปีค.ศ. 1926 อย่างไร และสามารถเข้าใจว่าถ้าหากการจัดการเปลี่ยนเป็นอีกแบบ ผลที่จะเกิดขึ้นจะเปลี่ยนไปหรือไม่

2.7 การประมวลผลภาษา

พัฒนาการของงานด้านการประมวลผลภาษา (Natural Language Processing) เป็นอีกปัจจัยหนึ่งที่ส่งผลต่อพัฒนาการของนวัตกรรมที่นำมาใช้ทางมนุษยศาสตร์ดิจิทัล การจัดเก็บข้อมูลภาษาและประมวลผลภาษาในงานต่าง ๆ ต้องอาศัยเครื่องมือการประมวลผลภาษาพื้นฐานหลายอย่าง เช่น การวิเคราะห์หัววิจิภาค เพื่อแยกส่วนที่เป็นวิภัติปัจจัยออกจากแก่นคำ (Stemming) ทำให้การนับจำนวนคำไม่ว่าจะมีรูปผันไปอย่างไร ก็นับเป็นคำเดียวกันได้ หรือการวิเคราะห์หมวดคำในข้อมูลให้รู้ว่คำนั้นควรเป็นหมวดคำใด เป็นคำกริยา หรือคำนาม เป็นต้น (Part of speech tagging) ข้อมูลเหล่านี้จะช่วยให้การค้นข้อมูลได้ถูกต้องตามต้องการมากขึ้น โปรแกรม Concordance ทั่วไปสามารถจัดการข้อมูลเหล่านี้ได้ บางโปรแกรมมีการประมวลผลที่ซับซ้อนขึ้น เช่น วิเคราะห์โครงสร้างประโยคและความสัมพันธ์ภายใน วิเคราะห์ประเภททางความหมายของคำต่าง ๆ วิเคราะห์หน่วยภาษาที่เป็นชื่อบุคคล องค์กร สถานที่ หรือที่เรียกว่าการทำ Named entity recognition เป็นต้น ความสามารถในการประมวลผลภาษาเหล่านี้อาจผนวกเป็นส่วนหนึ่งของโปรแกรมที่ใช้ เช่น LancsBox มี POS tagging สำหรับใช้กับบางภาษา หรืออาจพัฒนาเป็นโมดูลให้เรียกใช้งานเอง เช่น NLTK (2009) เป็น Python library ที่ใช้ประมวลผลภาษาอังกฤษได้ดี มีหลากหลายฟังก์ชันให้เลือกใช้ เช่น การทำ Tokenization, POS tagging, Named entity recognition, Syntactic parse เป็นต้น

The image shows the NLTK Python library interface. On the left is a sidebar with navigation links: Search, NLTK Documentation, API Reference, Example Usage, Module Index, Wiki, FAQ, Open Issues, NLTK on GitHub, Installation, Installing NLTK, Installing NLTK Data, More, Release Notes, Contributing to NLTK, and NLTK Team. The main content area is titled 'Identify named entities:' and contains a code snippet for using NLTK's named entity chunking. Below this is a section titled 'Display a parse tree:' with a code snippet for importing the treebank and displaying a parse tree. At the bottom is a visual representation of a parse tree for the sentence 'Pierre Venken 61 years old will join the board as a nonexecutive director Nov. 29'.

```

Identify named entities:

>>> entities = nltk.chunk.ne_chunk(tagged)
>>> entities
Tree('S', [(('At', 'IN'), ('eight', 'CD')), ('o'clock', 'JJ'), ('on', 'IN'), ('Thursday', 'NNP'), ('morning', 'NN'), ('PERSON', [(('Arthur', 'NNP'))]), ('did', 'VBD'), ('n't', 'RB'), ('feel', 'VB'), ('very', 'RB'), ('good', 'JJ'), ('.', '.')])

Display a parse tree:

>>> from nltk.corpus import treebank
>>> t = treebank.parsed_sents('wsj_0001.mrg')[0]
>>> t.draw()

```

ภาพที่ 7 Python library NLTK สำหรับประมวลผลภาษาอังกฤษ¹¹

¹¹ <https://www.nltk.org/>

The Stanford Natural Language Processing Group (n.d.) เป็นกลุ่มนักวิจัยอีกกลุ่มที่พัฒนาเครื่องมือการประมวลผลพื้นฐานให้สามารถนำไปใช้งานได้ ภาษาที่ใช้พัฒนาเป็นหลักคือภาษา Java การพัฒนาเครื่องมือประมวลผลภาษาที่มีมากขึ้นนี้เองที่ทำให้ผู้ที่ต้องการข้อมูลที่มีการวิเคราะห์ทางภาษาศาสตร์สำหรับใช้ในการศึกษาสามารถได้ข้อมูลที่มีมากกว่ารูปเขียนที่เห็น และเป็นนวัตกรรมวิธีที่มนุษยศาสตร์ดิจิทัลสามารถนำมาใช้ประโยชน์ได้

2.8 สื่อสังคมออนไลน์

การสื่อสารยุคใหม่เป็นการสื่อสารดิจิทัล โดยเฉพาะการสื่อสารผ่านสื่อสังคมออนไลน์กลายเป็นช่องทางหลักอย่างใหม่ในการสื่อสารระหว่างมนุษย์ ข้อมูลจำนวนมากมหาศาลในการสื่อสารที่เกิดขึ้นเป็นข้อมูลมหัต (Big data) ที่สามารถนำมาศึกษาเพื่อให้เข้าใจมนุษย์ในภาพรวมรวมถึงความต่างทางวัฒนธรรมได้ นวัตกรรมที่ช่วยในการศึกษาข้อมูลมหัตทั้งที่เป็นตัวภาษาหรือภาพทำให้สามารถสร้างภาพนิทัศน์แบบต่าง ๆ ให้เข้าใจลักษณะต่าง ๆ ของการสื่อสารของคนแต่ละกลุ่มได้ ตัวอย่างโครงการ Phototrails (n.d.) และ Hochman & Manovich (2013) ใช้ภาพจากอินสตาแกรม (Instagram) จำนวน 2.3 ล้านภาพที่ส่งจากผู้ใช้ในภูมิภาคต่าง ๆ ทั่วโลกเพื่อศึกษาลักษณะความต่างของการใช้อินสตาแกรมของผู้ใช้จาก 13 เมืองทั่วโลก ความแตกต่างของรูปแบบและโครงสร้างข้อมูลภาพเหล่านี้ใช้บ่งลักษณะเฉพาะคนแต่ละกลุ่มซึ่งสะท้อนถึงความต่างทางวัฒนธรรมได้ เป็นการศึกษาที่เรียกว่า Cultural Analytics ในโครงการใช้รูปภาพที่ได้มาสร้างภาพนิทัศน์ด้วยเทคนิคแบบต่าง ๆ ได้แก่ Radial, Montage, Photoplot, และ Photo and Line และเห็นความต่างของการใช้อินสตาแกรมในแต่ละเมือง

โครงการ Selfsity (n.d.) รวบรวมภาพถ่ายตัวเอง (Selfies) ผ่านอินสตาแกรม จำนวน 656,000 ภาพจากคนในเมืองต่าง ๆ แห่งละ 20,000–30,000 ภาพ แล้วสร้างภาพนิทัศน์แบบต่าง ๆ เพื่อแสดงถึงความต่างทางวัฒนธรรมของแต่ละแห่ง เช่น การยิ้มในภาพถ่ายตัวเองของแต่ละเมืองก็มีความต่างกัน และอายุและเพศของผู้โพสต์ภาพตัวเองในแต่ละเมืองที่ต่างกัน

ข้อมูลสื่อสังคมออนไลน์อีกประเภทคือ ทวิตเตอร์ (Twitter) หรือ X ซึ่งเป็นข้อความสั้น ๆ ที่นิยมส่งเพื่อแสดงความคิดเห็นเรื่องต่าง ๆ มักใช้เป็นแหล่งข้อมูลเพื่อศึกษากระแสความคิดที่เกิดขึ้นในช่วงเวลาใดเวลาหนึ่ง โครงการ Digital Narratives of Covid-19 (n.d.) เป็นโครงการรวบรวมข้อมูลทวิตเตอร์ ที่เกี่ยวข้องกับ Covid-19 ที่เป็นภาษาอังกฤษและภาษาสเปนในช่วงการระบาด และเผยแพร่ข้อมูลและโปรแกรมที่ใช้กับคลังข้อมูลนี้ มีการใช้โปรแกรม Voyant¹² เพื่อวิเคราะห์คลังข้อมูลและใช้ Topic modeling วิเคราะห์ข้อมูลว่าคำที่บ่งบอก หัวข้อที่สำคัญสามหัวข้อมีอะไรบ้าง

ทวิตเตอร์ นอกจากเป็นข้อมูลการสื่อสารในสื่อสังคมออนไลน์ ยังเป็นช่องทางแสดงออกทางวรรณกรรม Twitterature เป็นคำประสมระหว่าง Twitter และ Literature อันเป็นการบอกถึงการใช้อินเทอร์เน็ตเป็นช่องทางใหม่ใช้แสดงออกซึ่งงานสร้างสรรค์ เช่น คำคม บทกวี นิยาย Project Twitter

¹² <https://voyant-tools.org/>

Literature (n.d.) เป็นโครงการมนุษยศาสตร์ดิจิทัลที่รวบรวมข้อมูลงานเขียนผ่านทวิตเตอร์ โดยใช้คำสำคัญ 35 คำที่บอกความเป็นงานเขียนและชีวิตสังคมนักเขียนในการสกัดข้อมูลนี้

3. นวัตกรรมและผลต่อการศึกษาทางมนุษยศาสตร์

พัฒนาการของการศึกษาทางมนุษยศาสตร์ดิจิทัลได้รับประโยชน์จากพัฒนาทางคอมพิวเตอร์โดยตรง เมื่อการพัฒนาเทคโนโลยีคอมพิวเตอร์ทำให้คอมพิวเตอร์ทำงานต่าง ๆ ได้ดีขึ้น ก็ทำให้นวัตกรรมที่ใช้ในงานมนุษยศาสตร์ดิจิทัลมีความซับซ้อนและความสามารถมากขึ้น เริ่มต้นจากการใช้เทคโนโลยีคอมพิวเตอร์เพื่อเก็บข้อมูลและประมวลผลข้อมูล และขยายขีดความสามารถในการจัดเก็บข้อมูลที่ไม่ใช่เป็นเพียงตัวบท แต่รวมภาพ เสียง วิดีทัศน์ และยังสามารถเชื่อมโยงข้อมูลกับการแสดงผลผ่านแผนที่ แผนภูมิแบบต่าง ๆ การใช้เทคนิคภาพนิทัศน์แบบต่าง ๆ เพื่อแสดงลักษณะข้อมูลที่หลากหลายและซับซ้อนได้ ตลอดจนการใช้วิธีการทางสถิติ การเรียนรู้ด้วยเครื่อง ทำให้เทคนิคการวิเคราะห์ตัวบท (Text analytics) ต่าง ๆ เช่น การจำแนกประเภทเอกสาร (Text classification) (Dhingra et al., 2022; Pintas et al., 2021) การระบุอารมณ์ข้อความ (Sentiment analysis) (Yadav & Vishwakarma, 2019; Wankhade et al., 2022) การหาความคล้ายคลึงทางความหมาย (Semantic similarity) ฯลฯ เป็นเครื่องมือที่สามารถนำมาใช้ประโยชน์ในโครงการมนุษยศาสตร์ดิจิทัลต่าง ๆ ได้ เทคโนโลยีปัญญาประดิษฐ์โดยเฉพาะโมเดลภาษาขนาดใหญ่ (Large language model) เป็นนวัตกรรมที่น่าจับตามองว่าจะมีผลต่อการศึกษาทางมนุษยศาสตร์ดิจิทัลมากเพียงใด เพราะเครื่องมืออย่าง ChatGPT ได้แสดงให้เห็นถึงความรู้ความเข้าใจทางภาษาที่ลึกซึ้ง สามารถตอบคำถามและทำงานวิเคราะห์ทางภาษาไม่ว่าจะเป็นเรื่องความหมายคำ มโนอุปลักษณ์ หรือวาทกรรมวิเคราะห์เชิงวิพากษ์ได้

ความเกี่ยวข้องกับศาสตร์ต่าง ๆ นอกเหนือจากมนุษยศาสตร์ทำให้มนุษยศาสตร์ดิจิทัลมีลักษณะที่เป็นการศึกษาข้ามศาสตร์ (Transdisciplinary) คือการศึกษาที่หลอมรวมความรู้ศาสตร์ต่าง ๆ ที่เกี่ยวข้องบูรณาการเป็นการศึกษาในลักษณะใหม่ ศูนย์วิจัยทางด้านมนุษยศาสตร์ดิจิทัลจึงมักให้ข้อมูลเครื่องมือต่าง ๆ ที่เกี่ยวข้องและสามารถนำมาใช้ประโยชน์ได้ เช่น มหาวิทยาลัยเท็กซัส ให้ข้อมูลทรัพยากรและเครื่องมือที่เป็นประโยชน์ต่องานมนุษยศาสตร์ดิจิทัล โดยจัดแบ่งเป็นหมวดต่าง ๆ ได้แก่ การจัดการข้อมูล (Data management) การวิเคราะห์ตัวบทและการทำเหมืองข้อมูล (Text analysis and data mining) ระบบสารสนเทศทางภูมิศาสตร์และแผนที่ (GIS and mapping) การแสดงภาพนิทัศน์ (Visualization) และการเผยแพร่ผลงาน (Publishing)

Digital Humanities Tools and Resources

Search this Guide

Text Analysis & Data Mining

Home

- Introductory Readings

Methods

- Data Management
- Text Analysis & Data Mining**
- GIS & Mapping
- Visualization
- Publishing

Resources

- Creating and Publishing Projects
- DH at UT Austin
- DH Beyond UT Austin

Introduction

Computational text analysis tools allow scholars to read bodies of text in new ways by using machine learning to pick up on word frequency patterns in texts. This process, often called "distant reading" or "topic modeling," can complement traditional close reading by providing insights into patterns such as word usage, psychological tendencies, and language commonly associated with historical events, etc. that a human might not notice without the assistance of computational tools. To explore the capabilities of text analysis programs, please refer to the list of commonly used tools below.

Related LibGuides

- Text Analysis
- Working with APIs

Books

Data Mining and Learning Analytics by Samira ElAtia (Editor); Osmar R. Zaiane (Editor); Donald Ipperciel (Editor)
ISBN: 1118998235
Publication Date: 2016-09-26

Tools

- Voyant

Voyant is a web-based, dashboard-style tool that allows users to upload a corpora and visualize patterns in various ways. For instance, users can experiment with

ภาพที่ 8 หน้าเว็บของศูนย์มนุษยศาสตร์ดิจิทัลให้ข้อมูลเกี่ยวกับเครื่องมือต่าง ๆ

นวัตกรรมมาพร้อมกับการเปลี่ยนแปลงของธรรมชาติข้อมูลในงานมนุษยศาสตร์ จากเดิมที่ข้อมูลหลักเป็นสิ่งพิมพ์กลายมาเป็นข้อมูลดิจิทัล ทั้งที่เป็นการแปลงข้อมูลเดิมเป็นดิจิทัลหรือนำเข้าข้อมูลใหม่ที่เป็นดิจิทัลแต่กำเนิดอย่างสื่อสังคมออนไลน์ และทำให้การทำงานมนุษยศาสตร์จากเดิมที่ปกติทำงานเป็นเอกเทศมาเป็นงานข้ามศาสตร์ที่ต้องร่วมมือกับผู้อื่น และมีช่องทางติดต่อแลกเปลี่ยนความเห็นกันผ่านช่องทางออนไลน์มากขึ้น นวัตกรรมใหม่ ๆ เหล่านี้ที่เข้ามามีผลต่อการศึกษามนุษยศาสตร์ แต่ก็ไม่ได้เปลี่ยนแปลงคุณค่าและความสนใจในสาขาที่มีแต่เดิม คือการศึกษาเพื่อให้เข้าใจมนุษย์มากขึ้น ไม่ได้เปลี่ยนธรรมชาติของมนุษยศาสตร์ที่อาศัยการคิดเชิงวิพากษ์ (Critical thinking) การตีความจากข้อมูล (Interpretive) ความหลากหลายของความคิดเห็นในเรื่องเดียวกัน (Pluralism) อันนำไปสู่การถกเถียงเรื่องนั้น ๆ (Debate) เพียงแต่สิ่งที่ใช้เป็นข้อมูลการศึกษาจากเดิมที่เป็นงานสร้างสรรค์ต่าง ๆ ทั้งที่เป็นงานเขียน งานศิลปะต่าง ๆ การสื่อสารปฏิสัมพันธ์ ความคิดความเชื่อของมนุษย์ในสังคมวัฒนธรรมได้เปลี่ยนเป็นข้อมูลดิจิทัลที่มีจำนวนมากมายมหาศาล เป็นข้อมูลมหัศจรรย์ที่ต้องใช้นวัตกรรมใหม่ และเปิดมิติการศึกษาลักษณะใหม่ ๆ ขึ้น

จากการศึกษางานเป็นเรื่อง ๆ มาสู่การศึกษาจากข้อมูลขนาดใหญ่ ทำให้การศึกษาทางมนุษยศาสตร์ขยายจากการศึกษางานเชิงปัจเจกมาสู่การศึกษางานหลากหลาย ทำให้เห็นและเข้าใจมนุษย์ในภาพรวมมากขึ้น เข้าใจความคิดความเชื่อวัฒนธรรมในภาพรวมได้ มนุษยศาสตร์ดิจิทัลจึงได้ขยายขอบฟ้าการศึกษาครอบคลุมทั้งงานทางมนุษยศาสตร์และงานทางสังคมศาสตร์ในแขนงต่าง ๆ ด้วย แต่ทั้งหมดก็ยังตอบจุดมุ่งหมายเดิมทางมนุษยศาสตร์คือเพื่อให้เราเข้าใจมนุษย์มากขึ้น และยังเข้าใจถึงพฤติกรรมและปฏิสัมพันธ์ของมนุษย์ในสังคม

4. นวัตกรรมกับการศึกษามนุษยศาสตร์ในประเทศไทย

ตอนที่ 2 และ 3 ได้ให้ภาพรวมของนวัตกรรมที่เข้ามามีผลต่อการศึกษาทางมนุษยศาสตร์ดิจิทัลในบริบทโลก คำถามต่อไปคือในบริบทประเทศไทยเองมีความสนใจใช้นวัตกรรมเหล่านี้แล้วมากน้อยเพียงใด ที่ผ่านมารวบรวมข้อมูลภาษาเพื่อนำมาใช้ประโยชน์ในการศึกษาหรือการสร้างคลังข้อมูลภาษาไทยเป็น นวัตกรรมที่ถูกนำมาใช้ประโยชน์แล้วระยะหนึ่ง คลังข้อมูลภาษาไทยแห่งชาติ¹³ เป็นตัวอย่างของคลังข้อมูลที่มีผู้สนใจใช้เพื่อดึงข้อมูลการใช้ภาษาไทยที่เกิดขึ้นจริงมาใช้ศึกษาวิจัยเพื่อตอบคำถามเรื่องต่าง ๆ การเก็บข้อมูลในรูปของหอเอกสารดิจิทัลก็เป็นอีกหนึ่งนวัตกรรมที่มีการนำมาใช้ประโยชน์ หอสมุดดิจิทัลวชิรญาณ¹⁴เป็นแหล่งรวบรวมหนังสือภาษาไทยเก่าหลากหลายประเภทให้ผู้ใช้สามารถอ่านหรือเข้าถึงข้อมูลในหนังสือได้ มีข้อมูลร่องรอยการอ่านต่าง ๆ หนังสือจดหมายเหตุและหนังสือเนื้อหาวิชาต่าง ๆ ส่วนการประมวลผลภาษาไทยเป็นอีกหนึ่งแขนงที่มีการพัฒนาและสามารถนำมาใช้ประโยชน์ได้มากขึ้น โดยเฉพาะช่วงหลัง ๆ ที่มีผู้สนใจศึกษาการประมวลผลภาษาไทยมากขึ้น มีโมดูลการประมวลผลภาษาไทยที่เปิดให้ใช้ได้ทั่วไป เช่น PyThaiNLP¹⁵, TLTk¹⁶ เป็นต้น

สำหรับการแสดงภาพนิทัศน์ในงานมนุษยศาสตร์ที่ผ่านมายังไม่มีการใช้งานแพร่หลาย มีบางงานที่ต้องอาศัยการแสดงผลผ่านแผนที่ เช่น งานด้าน GIS ที่คณะอักษรศาสตร์นำเสนอข้อมูลภูมิศาสตร์คำศัพท์ที่รวบรวมจากการเก็บข้อมูลภาษาถิ่น (Teerarajanarat & Tingsabadh, 2011a, 2011b) อย่างไรก็ตาม ยังมีงานทางมนุษยศาสตร์ไทยอีกมากที่สามารถใช้ประโยชน์จากนวัตกรรมทางมนุษยศาสตร์ดิจิทัล “โครงการนวัตกรรมในมนุษยศาสตร์ดิจิทัล: การพัฒนาและการประยุกต์ใช้” ในชุดโครงการ “วิถีดิจิทัลในมนุษยศาสตร์สู่การพัฒนาทุนมนุษย์อย่างยั่งยืน” เป็นโครงการที่ได้รับการสนับสนุนจาก หน่วยบริหารและจัดการทุนด้านการเพิ่มความสามารถในการแข่งขันของประเทศ (บพข.) และ สำนักงานสภานโยบายการอุดมศึกษา วิทยาศาสตร์ วิจัยและนวัตกรรมแห่งชาติ (สอวช.) เพื่อให้ทำงานวิจัยบุกเบิกด้านนี้ โครงการประกอบด้วย 5 โครงการย่อยเป็นตัวอย่างที่ทำให้เห็นการใช้นวัตกรรมเพื่อพัฒนางานด้านมนุษยศาสตร์ดิจิทัลในประเทศไทยในด้านต่าง ๆ (Aroonmanakun et al., 2024)

โครงการย่อย “การพัฒนามาตรฐานและแพลตฟอร์มข้อมูลเพื่อการเชื่อมโยงข้อมูลด้านวัฒนธรรม” เป็นโครงการที่ทดลองศึกษาการเชื่อมโยงฐานข้อมูลด้านวัฒนธรรมโดยเน้นที่สาขามานุษยวิทยา 31 ฐานข้อมูลในศูนย์มานุษยวิทยาสรณ (องค์การมหาชน) โดยออกแบบออนโทโลยีด้านมานุษยวิทยา (Anthropology ontology) เพื่อให้สามารถเชื่อมโยงข้อมูลเดิมในแต่ละฐานเข้าด้วยกันได้ งานวิจัย Buranarach et al., (2022) นำเสนอการบูรณาการข้อมูลที่ถูกจัดเก็บไว้ในฐานข้อมูลที่หลากหลาย การบูรณาการข้อมูลนี้ต้องจัดการปัญหาความหลากหลายของข้อมูล (Data heterogeneity) ทั้งในแง่ของโครงสร้างข้อมูล (Data schema) และค่าข้อมูล (Data values) จากแหล่งข้อมูลที่แตกต่างกัน และต้องจัดการกับปัญหาความหลากหลายในรูปแบบการ

¹³ <http://www.arts.chula.ac.th/ling/tnc/>

¹⁴ <https://vajirayana.org/>

¹⁵ <https://pythainlp.org/>

¹⁶ <https://pypi.org/project/tltk/>

นำเสนอข้อมูล (Data representation) และปัญหาด้านความหมาย (Semantic issues) ที่ข้อมูลเดียวกัน อาจปรากฏในหลายฐานข้อมูลด้วยรูปแบบที่แตกต่างกัน งานวิจัยนี้จึงได้นำเสนอกรอบแนวคิดสำหรับการ บูรณาการเมทาดาตา เพื่อสนับสนุนคลังข้อมูลดิจิทัลในด้านมานุษยวิทยาสังคมและวัฒนธรรมของศูนย์ มานุษยวิทยาสรินธร กรอบแนวคิดนี้ถูกออกแบบโดยอ้างอิงจากโมเดลวงจรชีวิตของเมทาดาตา (Metadata lifecycle model: MLM) โดยใช้การจับคู่โครงสร้างข้อมูลแบบไม่เป็นขั้นตอน (Non-procedural schema mappings) เพื่อแสดงความสัมพันธ์ของข้อมูลในโครงสร้างที่หลากหลาย ต่อมาในงานวิจัยของ Ngootip et al. (2023) ได้นำเสนอการพัฒนาแพลตฟอร์มข้อมูลเชื่อมโยงแบบเปิด (Linked Open Data - LOD) เพื่อจัด ระเบียบและอำนวยความสะดวกในการเข้าถึงความรู้เกี่ยวกับนิทานพื้นบ้านและกลุ่มชาติพันธุ์ในประเทศแถบ อนุภูมิภาคกลุ่มแม่น้ำโขง แพลตฟอร์มดังกล่าวถูกพัฒนาขึ้นโดยใช้ Google OpenRefine เพื่อสร้างการเชื่อมโยง กับแหล่งข้อมูลภายนอก และไฟล์ RDF (รูปแบบ N-Triples) ถูกนำไปใช้งานบนเครื่องแม่ข่ายเพื่อเป็นจุด บริการสำหรับการสืบค้นข้อมูลเปิดเชื่อมโยง แพลตฟอร์ม LOD ที่พัฒนาขึ้นนี้แสดงให้เห็นถึงศักยภาพสำคัญใน การขยายการใช้งานไปยังเนื้อหาในโดเมนต่าง ๆ พร้อมทั้งเป็นทรัพยากรที่มีคุณค่าในการเข้าถึงและสำรวจ มรดกทางวัฒนธรรมอันหลากหลายของนิทานพื้นบ้านในประเทศแถบอนุภูมิภาคกลุ่มแม่น้ำโขง

งานวิจัยโครงการนี้จึงเป็นกรณีศึกษาของการบูรณาการเมทาดาตาในคลังข้อมูลดิจิทัลของศูนย์ มานุษยวิทยาสรินธร (SAC) โดยออกแบบโครงสร้างเมทาดาตากลาง (SAC common metadata schema) เพื่อรองรับการจับคู่ข้อมูลจากคลังข้อมูลดิจิทัล 31 แห่ง นอกจากนี้ ยังได้พัฒนาระบบ "One Search" ของ SAC เพื่อเป็นตัวอย่างการนำกรอบแนวคิดไปใช้จริง ผลการประเมินพบว่า กรอบแนวคิดการบูรณาการ เมทาดาตานี้สามารถสนับสนุนผู้เชี่ยวชาญด้านมานุษยวิทยาสังคมและวัฒนธรรมในการค้นหาข้อมูลแบบรวม จากหลายคลังข้อมูลได้อย่างมีประสิทธิภาพ ระบบต้นแบบสามารถสืบค้นข้อมูลทั้งหมดผ่านทางเว็บ (<https://onedb.sac.or.th/>) ในโครงการยังพัฒนาระบบการแนะนำการสืบค้นโดยช่วยแนะนำรายการที่มี คล้ายคลึงกันไปด้วย ความน่าสนใจและความท้าทายของงานนี้อยู่ที่การพยายามเชื่อมโยงฐานข้อมูลต่าง ๆ ซึ่งมี ข้อมูลต่างกันให้สามารถเชื่อมโยงกันได้ผ่านการใช้ออนโทโลยี ทำให้การสืบค้นเดียวสามารถเข้าถึงฐานข้อมูล ทั้งหมดได้ งานวิจัยนี้นับเป็นตัวอย่างที่ดีของการพัฒนาด้านออนโทโลยีและการประยุกต์ใช้เมทาดาตาในการ เชื่อมโยงฐานข้อมูลด้านมนุษยศาสตร์ หรือแม้แต่ฐานข้อมูลในสาขาอื่น ๆ ทั้งนี้ ชี้ให้เห็นถึงความสำคัญของการ กำหนดมาตรฐานกลางในการระบุเขตข้อมูล (Field) ของฐานข้อมูลตั้งแต่ขั้นตอนเริ่มต้นของโครงการ เนื่องจาก จะเอื้อต่อการเชื่อมโยงข้อมูลระหว่างฐานต่าง ๆ ในอนาคตอย่างมีประสิทธิภาพ นอกจากนี้ ระบบยังมีการ ประมวลผลภาษาไทยโดยใช้เทคนิคการหาความคล้ายคลึงของเอกสารผ่านเวกเตอร์ที่ได้จากการคำนวณ ค่าความถี่-กลับความถี่เอกสาร (TF/IDF) ซึ่งช่วยในการวิเคราะห์และจัดกลุ่มข้อมูลเชิงความหมายได้อย่างมี ประสิทธิภาพ

โครงการย่อย “ระบบแสดงข้อมูลบรรณานุกรมและเนื้อหาหนังสือภาษาไทยเป็นภาพเชิงตอบโต้ เพื่อ การสำรวจและวิเคราะห์ข้อมูลขั้นต้น” เป็นโครงการที่สนใจนวัตกรรมการแสดงภาพนิทัศน์ที่เหมาะสมสำหรับ

นำมาใช้ในระบบสืบค้นข้อมูลบรรณานุกรม ในโครงการได้ศึกษาการแสดงผลภาพนิทัศน์รูปแบบต่าง ๆ ที่มีการใช้เพื่อช่วยแสดงผลข้อมูลจากห้องสมุด แต่ด้วยเหตุที่ฐานข้อมูลห้องสมุดในประเทศไทยไม่ได้มีข้อมูลหลากหลายมากพอที่จะนำมาใช้แสดงผลได้ ในงานวิจัย Ruchikachorn et al. (2024) จึงนำเสนอการพัฒนาระบบการสร้างภาพข้อมูลแบบโต้ตอบที่เน้นผู้ใช้งาน (User-centric interactive visualization system) ซึ่งออกแบบมาเพื่อจัดการข้อมูลบรรณานุกรมโดยเฉพาะ การพัฒนาระบบนี้ใช้แนวทาง Design Study "Lite" ซึ่งเป็นกระบวนการที่ปรับปรุงต้นแบบโดยทำงานร่วมกับผู้เชี่ยวชาญในสาขาโดยตรง ต้นแบบที่พัฒนาขึ้นช่วยให้ผู้ใช้สามารถเจาะลึกข้อมูลผ่านลำดับของแผนภูมิแท่งที่ปรับแต่งได้ ผู้ใช้สามารถปรับแต่งการสร้างภาพข้อมูลและจัดลำดับตามความต้องการเฉพาะของตน การทดสอบระบบนี้ดำเนินการกับผู้เข้าร่วม 10 คน ซึ่งได้รับมอบหมายให้ทำงาน 5 งาน และวิเคราะห์พฤติกรรมของพวกเขาเพื่อค้นหาความสัมพันธ์ระหว่างประเภทของงานและกลยุทธ์ในการพัฒนาลำดับของแผนภูมิ ผลการวิจัยชี้ให้เห็นถึงปัจจัยบางประการที่มีศักยภาพในการเปิดประเด็นวิจัยใหม่ ๆ เกี่ยวกับลำดับการสร้างภาพข้อมูล และได้เผยแพร่ที่ <https://github.com/puripant/library-dash-public>

โครงการย่อย “เหนือกว่าเวิร์ดคลาวด์: การแสดงตัวบทให้เป็นภาพเชิงความหมาย” เป็นอีกโครงการที่สนใจนวัตกรรมวิธีการแสดงผลภาพนิทัศน์ โดยต้องการปรับปรุงการแสดงผลก่อนคำในรูปแบบใหม่ ให้สะท้อนความเข้าใจในข้อมูลมากขึ้น ความท้าทายจึงอยู่ที่สองเรื่องใหญ่ เรื่องแรกเป็นการแสดงผลภาษาไทยที่ต้องการการประมวลผลเฉพาะ ส่วนเรื่องที่สองเป็นการหาวัดวิธีใหม่ที่จะช่วยให้เข้าใจตัวบทมากขึ้น Rutherford et al. (2024) กล่าวถึงความพยายามในการปรับปรุงการแสดงผลภาพก่อนคำที่ผ่านมาว่ามีการพยายามจัดกลุ่มคำตามความหมายโดยใช้ Topic modeling แต่ก็ได้สนใจเรื่องการจัดวางตำแหน่งและสีของการแสดงผล อีกทั้งไม่ได้ทดสอบจริงว่าสิ่งที่แสดงนั้นได้ผลดีเพียงใด ในงานวิจัยนี้จึงได้ทดลองใช้หลายอัลกอริทึม ได้แก่ k-means, t-SNE และ LDA Mallet เพื่อสร้างและเปรียบเทียบการแสดงผลก่อนคำที่มีการจัดกลุ่มตามความหมาย (Semantic grouping) การลงสีตามความหมาย (Semantic coloring) และการปรับขนาดฟอนต์ตามความหมาย (Semantic font-sizing) อาศัยทฤษฎีเกี่ยวกับการรับรู้ทางสายตาของมนุษย์ (Visual perception) โดยเฉพาะอย่างยิ่งหลักการการรับรู้แบบเกสตัลต์ (Gestalt principles) มาใช้เพื่อให้การแสดงผลภาพนิทัศน์มีความหมายมากกว่าเดิม ในงานวิจัยมีการทดลองกับกลุ่มตัวอย่างเพื่อวัดประสิทธิภาพของสิ่งที่พัฒนานี้ ผลการวิจัยพบว่า ผู้เข้าร่วมชื่นชอบภาพก่อนคำแบบการจัดกลุ่มและการลงสีในกลุ่มย่อยที่ถูกควบคุมด้วย k-means ซึ่งได้รับอิทธิพลจากหลักการ Gestalt ในเรื่องของความใกล้เคียงและความคล้ายคลึงที่ช่วยในการจัดระเบียบการรับรู้ ความน่าสนใจของงานนี้จึงไม่ใช่เพียงใช้กับภาษาไทยได้ แต่ยังประยุกต์วิธีการประมวลผลภาษาเพื่อจัดกลุ่มคำตามความหมายและแสดงด้วยสีและตำแหน่งที่ใกล้เคียงกันในรูปแบบที่ผู้ใช้พอใจ วิธีการนี้นอกจากจะใช้กับภาษาไทยแล้วยังใช้กับภาษาอื่นเช่นภาษาอังกฤษได้ด้วย โมเดลการแสดงผลก่อนคำนี้เผยแพร่ที่ <https://pypi.org/project/swordcloud/>

โครงการย่อย “การศึกษาเปรียบเทียบการวิเคราะห์ด้วยวิธีการต่าง ๆ ทางภาษาศาสตร์คอมพิวเตอร์” สนใจการนำนวัตกรรมวิธีมาใช้กับตัวบทภาษาไทย ในโครงการได้สำรวจวิธีการต่าง ๆ ทางภาษาศาสตร์คอมพิวเตอร์ที่สามารถนำมาใช้ทำงานกับตัวบท และกำหนดเป้าเพื่อหาวิธีที่ใช้บอกระดับความยากง่ายของตัวบทภาษาไทย ซึ่งทรัพยากรพื้นฐานที่ควรมีคือคำศัพท์และลักษณะทางภาษาที่สะท้อนความยากง่ายของตัวบท แต่เนื่องจากข้อมูลเหล่านี้ยังไม่มีเผยแพร่ ในงานวิจัยนี้จึงได้พัฒนารายการคำศัพท์พื้นฐานภาษาไทยมาใช้เบื้องต้น โดยนำข้อมูลจากคลังข้อมูลภาษาไทยแห่งชาติและคลังข้อมูลหนังสืออ่านภาษาไทย 4 ระดับขึ้นมาเทียบเคียงและสร้างรายการคำศัพท์พื้นฐาน 1,404 คำ ส่วนรายการลักษณะทางภาษาเนื่องจากยังไม่มีเครื่องมือที่วิเคราะห์ลักษณะภาษาไทยได้ ในงานนี้จึงได้พัฒนาโมดูลสำหรับการวิเคราะห์ข้อความ (Text analysis) เพื่อแจกแจงลักษณะต่าง ๆ ของตัวบทออกมาให้ได้ก่อน ในกระบวนการนี้ต้องอาศัยการวิเคราะห์ลักษณะทั้งระดับพยางค์ (SylAna) ระดับคำ (WordAna) และระดับวากยสัมพันธ์ที่ต้องอาศัยการวิเคราะห์โครงสร้างพยางค์ ทั้งหมดนี้ยังไม่มีเครื่องมือใดให้ใช้จึงต้องพัฒนาโมดูลเหล่านี้ขึ้นมาก่อน เมื่อสามารถหาข้อมูลลักษณะทางภาษาในระดับต่าง ๆ ได้แล้ว จึงใช้ลักษณะที่ได้เป็นเวกเตอร์แทนตัวบทต่าง ๆ เพื่อนำมาประมวลผลเรื่องระดับความยากง่าย อาศัยตัวบทที่พบในหนังสือเรียนภาษาไทยแต่ละระดับขึ้นมาใช้สร้างโมเดลให้เข้าใจลักษณะตัวบทในระดับชั้นต่าง ๆ และใช้โมเดลที่สร้างขึ้นทำนายระดับชั้นของตัวบทอื่น ๆ ต่อไป ความน่าสนใจของงานนี้นอกจากอยู่ที่การพัฒนาทรัพยากรและเครื่องมือพื้นฐานเพื่อแปลงตัวบทให้เป็นเวกเตอร์ของลักษณะทางภาษาเพื่อสร้างโมเดลภาษาบอกระดับชั้น เครื่องมือที่พัฒนาขึ้นนี้จะพื้นฐานสำคัญสำหรับการประยุกต์ใช้งานในด้านอื่น ๆ ต่อไป โดยโมดูลต่าง ๆ ได้ถูกรวมเป็นส่วนหนึ่งของไลบรารี TLTK และเผยแพร่ผ่านเว็บไซต์ <https://pypi.org/project/tltk/> ตัวอย่างการทดลองสร้างโมเดลสำหรับจำแนกประเภทของตัวบทที่โครงการดำเนินการแสดงให้เห็นว่าการพัฒนาโปรแกรมเพื่อแปลงตัวบทให้เป็นเวกเตอร์ของลักษณะทางภาษา นับเป็นเครื่องมือที่มีศักยภาพในการศึกษามนุษยศาสตร์ดิจิทัลสำหรับภาษาไทย อย่างไรก็ตาม ลักษณะทางภาษาที่สามารถวิเคราะห์ได้ในปัจจุบันยังคงจำกัดเฉพาะลักษณะพื้นฐานเท่านั้น จึงควรมีการพัฒนาต่อยอดเพื่อรองรับการวิเคราะห์ลักษณะเชิงลึกเพิ่มเติม เช่น ลักษณะทางความหมาย (Semantic features) และลักษณะทางปริจเฉท (Discourse features) เพื่อให้สามารถประยุกต์ใช้ในการวิเคราะห์ตัวบทภาษาไทยได้อย่างครอบคลุมยิ่งขึ้น

ตัวอย่างโครงการต่าง ๆ ที่กล่าวมาทำให้เห็นว่าการเพิ่มความสามารถในการประมวลผลภาษาไทยเป็นเรื่องสำคัญที่ต้องการการพัฒนาอย่างต่อเนื่อง การวิเคราะห์โครงสร้างประโยคจำเป็นต้องมีคลังข้อมูลที่แสดงความสัมพันธ์แบบพยางค์ของคำในภาษาไทย Aroonmanakun & Champai boon (2021) ได้กำหนดแนวทางการวิเคราะห์โครงสร้างพยางค์ในภาษาไทยและเริ่มกำกับโครงสร้างต้นไม้พยางค์จากข้อมูลส่วนหนึ่งของคลังข้อมูลภาษาไทยแห่งชาติ ข้อมูลนี้ได้นำมาใช้ร่วมกับโปรแกรม MaltParser¹⁷ ซึ่งเป็น Data-driven dependency parsing (ใช้ ภาษา Java) แล้วนำมาผนวกเป็นโมดูล MaltParser ใน TLTK ต่อมา Sriwirote

¹⁷ <http://www.maltparser.org/>

et al. (2025) ได้เพิ่มเติมคลังข้อมูลต้นไม้มิ่งพามาจากตัวบทอื่น ๆ และพัฒนาโปรแกรมแจกส่วนประโยคโดยใช้แบบจำลองพื้นฐานแบบ Transformer สำหรับวิเคราะห์โครงสร้างพยางค์ในภาษาไทย

การใช้ประโยชน์จากคลังข้อมูลแบบต่าง ๆ เป็นตัวอย่างของการใช้นวัตกรรมเพื่อการศึกษาภาษา Aroonmanakun & Aroonmanakun (2023) ใช้คลังข้อมูลภาษาไทย คลังข้อมูลภาษาอังกฤษ และคลังข้อมูลการแปลภาษาอังกฤษ-ไทย เพื่อศึกษาคำพ้องความหมายในภาษาอังกฤษ *little* และ *small* และคำเทียบเท่าในภาษาไทย *น้อย* และ *เล็ก* ผลการศึกษาพบว่า *little* และ *small* กับคำในภาษาไทย *น้อย* และ *เล็ก* ไม่ได้มีความหมายเหมือนกันโดยสมบูรณ์ นอกจากนี้ ยังพบว่า *little* และ *small* สามารถแปลเป็นภาษาไทยด้วยคำอื่น ๆ นอกเหนือจาก *น้อย* และ *เล็ก* ซึ่งสะท้อนถึงความซับซ้อนของความหมายที่คำเหล่านี้สามารถเชื่อมโยงไปสู่ความหมายอื่น ๆ ได้ งานวิจัยนี้ชี้ให้เห็นว่าข้อมูลจากพจนานุกรมเพียงอย่างเดียวไม่เพียงพอ ผู้เรียนควรเรียนรู้วิธีใช้คลังข้อมูลภาษาเป็นแหล่งข้อมูลเพิ่มเติมเพื่อเพิ่มพูนความรู้ด้านคำศัพท์ในทั้งสองภาษา

โครงการย่อย “การพัฒนาแบบจำลองตัวแทนอัจฉริยะเพื่อพยากรณ์กระบวนการตัดสินใจและพฤติกรรมของมนุษย์ กรณีศึกษา นักท่องเที่ยวในประเทศไทย” เป็นโครงการที่แสดงให้เห็นถึงการใช้แบบจำลองตัวแทนในบริบทสังคมไทย ในโครงการได้ทดลองสร้างโมเดลเพื่อทำนายพฤติกรรมของนักท่องเที่ยวที่มีต่อตัวแปรต่าง ๆ ในการทดลองโมเดล Panbumrungkij & Vannameteetee (2023) ได้พัฒนาแบบจำลองตัวแทนโดยใช้โปรแกรม NetLogo ร่วมกับระบบสารสนเทศภูมิศาสตร์ เพื่อศึกษาพฤติกรรมการเลือกสถานที่ท่องเที่ยวและการกระจายตัวของนักท่องเที่ยวรายวันในจังหวัดชลบุรี โดยมีการจำลองนักท่องเที่ยวจำนวน 3,200 คน แบ่งเป็น 12 กลุ่มตามเพศและช่วงอายุ โดยแต่ละกลุ่มมีความสนใจในแหล่งท่องเที่ยวแตกต่างกัน แหล่งท่องเที่ยวถูกแบ่งออกเป็น 6 ประเภท รวมทั้งสิ้น 59 แห่ง แบบจำลองทำให้เห็นกระบวนการตัดสินใจของตัวแทนในการเลือกแหล่งท่องเที่ยวประเภทต่าง ๆ และการกระจายตัวของนักท่องเที่ยวในพื้นที่ต่าง ๆ ไม่สม่ำเสมอ นอกจากรายวันในจังหวัดชลบุรีแล้ว ในโครงการได้เลือกศึกษาอีก 5 จังหวัดรวมเป็น 6 จังหวัดเพื่อใช้เป็นตัวแทน 6 ภูมิภาค คือ พระนครศรีอยุธยา (ภาคกลาง) ชลบุรี (ภาคตะวันออก) กาญจนบุรี (ภาคตะวันตก) อุตรดิตถ์ (ภาคเหนือ) ภูเก็ต (ภาคใต้) และ เชียงราย (ภาคเหนือ) อย่างไรก็ตาม การประมวลผลของจังหวัดเชียงราย ไม่สามารถดำเนินการได้ เนื่องจากการนำเข้าข้อมูลเชิงพื้นที่ของจังหวัดเชียงรายไม่สมบูรณ์และติดปัญหาระหว่างการประมวลผล การศึกษาครั้งนี้ จึงประมวลผลโปรแกรมการประมวลผล NetLogo จำนวนทั้งสิ้น 5 จังหวัด และได้ข้อสรุปโดยรวมว่า ขนาดและรูปร่างของจังหวัดมีผลโดยตรงต่อการกระจายตัวและความหนาแน่นของตัวแทนนักท่องเที่ยว โดยเฉพาะในจังหวัดที่มีพื้นที่กว้างและแหล่งท่องเที่ยวกระจุกกระจาย ตัวแทนจะมีข้อจำกัดด้านเวลาเดินทางทำให้ไม่สามารถไปถึงแหล่งท่องเที่ยวทั้งหมดได้ ขณะที่แหล่งท่องเที่ยวที่อยู่ใกล้กันจะดึงดูดตัวแทนให้ใช้เวลาอยู่นานขึ้น การกระจายตัวของประเภทแหล่งท่องเที่ยวที่ไม่เท่าเทียมกันในแต่ละจังหวัดทำให้แหล่งที่มีตัวเลือกน้อยมีความหนาแน่นของตัวแทนสูงกว่า แม้ไม่ใช่แหล่งยอดนิยม ตัวแทนจะถูกบังคับให้เลือกตามโอกาสและจำนวนสถานที่ที่จำกัดในประเภทนั้น ขณะที่แหล่งธรรมชาติมักได้รับความนิยมเป็นพิเศษ ยกเว้นจังหวัดที่โดดเด่นด้านประวัติศาสตร์ เช่น

พระนครศรีอยุธยา ตำแหน่งการปล่อยตัวของนักท่องเที่ยว เช่น บริเวณทางเข้าหลัก ก็มีผลต่อการกระจุกตัวของนักท่องเที่ยวเช่นกัน รวมถึงข้อจำกัดด้านการเข้าถึงแหล่งท่องเที่ยว เช่น ภูมิประเทศที่ยากลำบาก ทำให้นักท่องเที่ยวเข้าถึงบางพื้นที่ได้ยาก สุดท้าย ปัจจัยด้านภูมิประเทศและการเข้าถึงมักมีบทบาทสำคัญกว่าความต้องการส่วนตัวในการตัดสินใจเลือกแหล่งท่องเที่ยวของนักท่องเที่ยว

ความท้าทายของโครงการวิจัยนี้นอกจากอยู่ที่การประยุกต์ใช้แบบจำลองตัวแทนเพื่อศึกษาพฤติกรรมนักท่องเที่ยว ยังอยู่ที่ข้อมูลป้อนเข้าในระบบซึ่งต้องเป็นข้อมูลที่สอดคล้องกับความจริง เนื่องจากข้อมูลการท่องเที่ยวทุกอย่างไม่ได้มีพร้อมใช้ ในโครงการจึงต้องเก็บข้อมูลด้วยการทำแบบสอบถามไปตามพื้นที่เป้าหมายเพื่อรวบรวมข้อมูลต่าง ๆ ที่ต้องการใช้ในการทำงานของแบบจำลองตัวแทน การเก็บข้อมูลการท่องเที่ยวอย่างสม่ำเสมอจึงเป็นพื้นฐานที่จำเป็นที่ภาครัฐควรใส่ใจเก็บข้อมูลต่าง ๆ เพื่อให้มีข้อมูลเพียงพอในการใช้ประโยชน์ได้เต็มที่

นอกจากนวัตวิธิต่าง ๆ ที่กล่าวมา เทคโนโลยีปัญญาประดิษฐ์ก็เป็นอีกหนึ่งนวัตกรรมที่จะมีผลต่อการศึกษาทางมนุษยศาสตร์ดิจิทัลต่อไป Aroonmanakun (2023) กล่าวถึงความสามารถของปัญญาประดิษฐ์หรือ AI ในงานด้านภาษาและการศึกษา และได้ทดลองใช้ AI ช่วยเขียนบทความวิชาการซึ่งได้ผลมาเป็นเนื้อหาและรูปแบบที่น่าเชื่อถือ แต่มีปัญหาเรื่องข้อมูลที่อาจไม่ใช่ข้อมูลจริง การตรวจสอบภาษาว่าเขียนด้วย AI หรือไม่นั้นไม่ใช่เรื่องง่ายและตรงไปตรงมา อีกทั้ง AI มีพัฒนาการความสามารถเพิ่มขึ้นอย่างรวดเร็วตลอดเวลา ความสำคัญจึงไม่ใช่เรื่องการอนุญาตให้ผู้เรียนใช้หรือไม่ใช้ AI แต่อยู่ที่จะทำอย่างไรให้ AI เป็นประโยชน์อย่างแท้จริงต่อระบบการศึกษา และด้วยความสามารถทางภาษาของ AI เอง การนำมาใช้เพื่อประโยชน์ในงานมนุษยศาสตร์ดิจิทัลจึงไม่ใช่เรื่องไกลตัว

5. บทสรุป

พัฒนาการของเทคโนโลยีคอมพิวเตอร์นำมาซึ่งการศึกษาทางมนุษยศาสตร์รูปแบบใหม่ที่เรียกว่ามนุษยศาสตร์ดิจิทัล การศึกษาวิจัยทางมนุษยศาสตร์ดิจิทัลไม่ได้เป็นเพียงการนำคอมพิวเตอร์มาช่วยประมวลผลข้อมูล แต่ได้เปิดมิติใหม่ในการศึกษาวิจัยทางมนุษยศาสตร์ ทำให้สามารถศึกษาจากข้อมูลจำนวนมากได้อย่างเป็นกาวีสัยและเปิดแง่มุมใหม่ ๆ ในการศึกษาทางมนุษยศาสตร์ บทความนี้ได้แสดงให้เห็นถึงนวัตวิธิต่าง ๆ ที่มีใช้ในโครงการด้านมนุษยศาสตร์ดิจิทัลในบริบทโลก และนำเสนอตัวอย่างโครงการวิจัยในไทยที่ได้ใช้นวัตวิธิต่าง ๆ ในการศึกษาวิจัย โครงการเหล่านี้เป็นเพียงตัวอย่างส่วนหนึ่งที่แสดงให้เห็นถึงความหลากหลายในงานด้านมนุษยศาสตร์ดิจิทัลในบริบทสังคมไทย ยังมีงานทางมนุษยศาสตร์ดิจิทัลอีกจำนวนมากที่มีการพัฒนาขึ้นในบริบทสังคมไทย การสำรวจสถานภาพการศึกษามนุษยศาสตร์ดิจิทัลที่ครอบคลุมจะช่วยให้เห็นทิศทาง แนวโน้ม และหัวข้อวิจัยที่ควรศึกษาต่อได้ การพัฒนางานด้านมนุษยศาสตร์ดิจิทัลยังมีความจำเป็นต้องได้รับการสนับสนุนและผลักดันให้เป็นที่ยอมรับและเข้าใจในแวดวงวิชาการไทย จึงจะสามารถต่อยอดงานสร้างสรรค์ใหม่ ๆ ได้

6. กิตติกรรมประกาศ

บทความนี้เป็นบทความสังเคราะห์ผลที่ได้จากโครงการย่อย 5 โครงการที่ได้รับงบประมาณสนับสนุนจากกองทุนส่งเสริมวิทยาศาสตร์วิจัยและนวัตกรรม โดยหน่วยบริหารและจัดการทุนด้านการพัฒนากำลังคน และทุนด้านการพัฒนาสถาบันอุดมศึกษา การวิจัยและการสร้างนวัตกรรม สัญญาเลขที่ B05F640234

เอกสารอ้างอิง

- AdminO4DH. (2019, October 20). **Ontology for digital humanities**. O4DH. <http://o4dh.com/>
- American Panorama. (n.d.). **American Panorama: An atlas of United States history**. <https://dsl.richmond.edu/panorama>
- Aroonmanakun, V., & Aroonmanakun, W. (2023). A comparison of ‘little’ and ‘small’ in English and Thai: A corpus-based study. **LEARN Journal**, 16(1), 296-319.
- Aroonmanakun, W. (2022). **Language data and statistics 1: Language data**. <https://awirote.medium.com/ข้อมูลภาษากับสถิติ-1-ข้อมูลภาษา-bfc1c513106a>
- Aroonmanakun, W. (2023a). **Language data and statistics 2: Basic statistics**. <https://awirote.medium.com/ข้อมูลภาษากับสถิติ-2-สถิติพื้นฐาน-9c7237ad4147>
- Aroonmanakun, W. (2023b). **Language data and statistics 3: Data usage**. <https://awirote.medium.com/ข้อมูลภาษากับสถิติ-3-การใช้ข้อมูล-7143183c0aff>
- Aroonmanakun, W. (2023e). AI and the future of higher education. **Journal of New Horizon in Higher Education**, 1(1), 19–37.
- Aroonmanakun, W., & Champaiboon, T. (2021). **Thai data annotation guidelines based on the Universal Dependencies framework** (Version 2) [In Thai].
- Aroonmanakun, W., Buranarach, M., Chansanam, W., Rungcharoensuksri, S., Buranasing, W., Ruchikachorn, P., Rutherford, A., Panbumrungskij, T., & Vannamettee, E. (n.d.). **Innovative methods in digital humanities: Development and application** [Research report]. Program Management Unit for Human Resources & Institutional Development, Research and Innovation (PMU-B) and Office of National Higher Education Science Research and Innovation Policy Council (NXPO). [In Thai].
- Biber, D. (1993). The multi-dimensional approach to linguistic analyses of genre variation: An overview of methodology and findings. **Computers and the Humanities**, 26, 331–345.
- Biber, D. (2011). Corpus linguistics and the study of literature: Back to the future? **Scientific Study of Literature**, 1, 15–23.
- Binongo, J. N. G. (2003). Who wrote the 15th book of Oz? An application of multivariate analysis to authorship attribution. **CHANCE**, 16(2), 9–17. <https://doi.org/10.1080/09332480.2003.10554843>
- Buranarach, M., Buranasing, W., Rungcharoensuksri, S., Sarawasee, P., Ngootip, T., & Chansanam, W. (2022). Metadata integration framework for data integration of socio-cultural anthropology digital repositories: A case study of Princess Maha Chakri Sirindhorn Anthropology Centre. **Informatics**, 9(2), 38.

- Culturomics. (2024). In **Wikipedia**. <https://en.wikipedia.org/wiki/Culturomics>
- Dhingra, M., Dhabliya, D., Dubey, M. K., Gupta, A., & Reddy, D. H. (2022). A review on comparison of machine learning algorithms for text classification. **2022 5th International Conference on Contemporary Computing and Informatics (IC3I)**, 1818–1823.
- Digital Karnak. (n.d.). **Digital Karnak – ArchVR at UCLA**. <https://archvr.ucla.edu/projects/digital-karnak/>
- Digital Narratives of Covid-19. (n.d.). **Digital Narratives of Covid-19**. <https://covid.dh.miami.edu/>
- Divisions, S. C. R. (n.d.). **LibGuides: African American women writers of the 19th century**. <https://libguides.nypl.org/african-american-women-writers-of-the-19th-century>
- Dryer, M. S., & Haspelmath, M. (Eds.). (2013). **WALS Online (v2020.3)** [Data set]. Zenodo. <https://doi.org/10.5281/zenodo.7385533>
- Gasmi, N., Grignard, A., Drogoul, A., Gaudou, B., Taillandier, P., Tessier, O., & Vo, D. A. (2014, May 5). **Reproducing and exploring past events using agent-based geo-historical models**. <https://hal.science/hal-01394659>
- Gavin, M. (2014). Agent-based modeling and historical simulation. **Digital Humanities Quarterly (DHQ)**, 8(4). <http://www.digitalhumanities.org/dhq/vol/8/4/000195/000195.html>
- Hochman, N., & Manovich, L. (2013). Zooming into an Instagram city: Reading the local through social media. **First Monday**, 18(7). <https://doi.org/10.5210/fm.v18i7.4711>
- Jane Austen Fiction Manuscripts. (n.d.). **Jane Austen Fiction Manuscripts**. <https://janeausten.ac.uk/>
- Jelodar, H., Wang, Y., Yuan, C., & Feng, X. (2017). Latent Dirichlet Allocation (LDA) and topic modeling: Models, applications, a survey. **Multimedia Tools and Applications**, 78, 15169–15211. <https://doi.org/10.1007/s11042-018-6894-4>
- LOTR Project. (n.d.). **An interactive analysis of Tolkien's works**. <http://lotrproject.com/statistics/books/>
- Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Brockman, W., The Google Books Team, Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M. A., & Aiden, E. L. (2010). Quantitative analysis of culture using millions of digitized books. **Science**. 331(6014), 176-182. <https://doi.org/10.1126/science.1199644>
- Ngootip, T., Manorom, P., Chansanam, W., & Buranarach, M. (2023). Developing a linked open data platform for folktales in the Greater Mekong Subregion. **Emerging Science Journal**, 7(6), 1937–1953. <https://doi.org/10.28991/esj-2023-07-06-06>
- NLTK. (2009). **Natural Language Toolkit — NLTK 3.4.4 documentation**. <https://www.nltk.org/>
- Ossom-Williamson, P., & Rambsy, K. (2021). 1.4 What is distant reading? In **The data notebook**. Mavs Open Press. <https://uta.pressbooks.pub/datanotebook/chapter/1-4-distant-reading>
- Panbumrungkij, T., & Vannamettee, E. (2023). A study of tourist behavior in destination selection and distribution of tourists using an agent-based model: A case study of Chon Buri Province [In Thai]. **Journal of Letters**, 52(2), 1–30.
- Pelagios Network. (n.d.). **Pelagios Network**. <https://pelagios.org/>
- Phototrails. (n.d.). **Phototrails**. <http://phototrails.info/>

- Pintas, J. T., Fernandes, L. A., & Garcia, A. C. (2021). Feature selection methods for text classification: A systematic literature review. **Artificial Intelligence Review**, 54(8), 6149–6200. <https://doi.org/10.1007/s10462-021-09970-6>
- Project Twitter Literature. (n.d.). **Project Twitter Literature**. <https://twitlit.github.io/index.html>
- Recogito. (n.d.). **Semantic annotation without the pointy brackets**. <https://recogito.pelagios.org/>
- Ruchikachorn, P., Sumanakul, S., Punyasiri, T., & Wilainuch, N. (2024). **A study on dashboard creation process based on bibliographic data** [Manuscript submitted for publication].
- Rutherford, A., Ruechai, K., Angkavichai, N., & Ruchikachorn, P. (2024). **From aesthetics to functionality: User preferences in algorithmic word cloud design** [Manuscript submitted for publication].
- Selfiecity. (n.d.). **Selfiecity**. <https://selfiecity.net/>
- Shakespeare's Circuits. (n.d.). **Shakespeare's Circuits**. <http://shakespearescircuits.northwestern.edu>
- Six Degrees of Francis Bacon. (n.d.). **Six Degrees of Francis Bacon**. <http://www.sixdegreesoffrancisbacon.com/>
- Sriwrote, P., Leong, W. Q., Polpanumas, C., Thanyawong, S., Tjhi, W. C., Aroonmanakun, W., & Rutherford, A. (2025). The Thai Universal Dependency Treebank. **Transactions of the Association for Computational Linguistics**, 13, 376–391. https://doi.org/10.1162/tacl_a_00745
- Tauber, J. (n.d.). **Digital Tolkien Project**. <https://digitaltolkien.com/>
- Teerarojanarat, S., & Tingsabadh, K. (2011a). A GIS-based approach for dialect boundary studies. **Dialectologia**, 6(6), 55–75.
- Teerarojanarat, S., & Tingsabadh, K. (2011b). Using GIS for linguistic study: A case of dialect change in the Northeastern region of Thailand. **Procedia – Social and Behavioral Sciences**, 21, 362–371.
- TEI Consortium. (2024). **Guidelines for electronic text encoding and interchange** (Version 4.8.1). <https://tei-c.org/release/doc/tei-p5-doc/en/html/index.html>
- The Stanford Natural Language Processing Group. (n.d.). **Software – The Stanford Natural Language Processing Group**. <https://nlp.stanford.edu/software/>
- The World of Dante. (n.d.). **The World of Dante**. <http://www.worldofdante.org>
- Underwood, T. (2016). Distant reading and recent intellectual history. In M. K. Gold & L. F. Klein (Eds.), **Debates in the Digital Humanities 2016**. <https://dhdebates.gc.cuny.edu/read/untitled/section/3b96956c-aab2-4037-9894-dc4ff9aa1ec5>