

การจัดเก็บรักษาเว็บถาวร (Web archiving): แนวคิดใหม่เพื่อการพัฒนาแหล่งสารสนเทศ

ในยุคดิจิทัล

วชิราภรณ์ คลังธนบูรณ์¹

จินดารัตน์ เบอรพันธุ์²

บทคัดย่อ

บทความนี้ชี้ให้เห็นความสำคัญของการจัดเก็บรักษาเว็บถาวร (Web archiving) ในฐานะที่เป็นแนวคิดใหม่สำหรับการพัฒนาแหล่งสารสนเทศใหม่ในยุคดิจิทัล การจัดเก็บรักษาเว็บถาวรเป็นแนวทางหนึ่งที่ช่วยให้มั่นใจว่าในระยะยาวผู้ใช้สามารถเข้าถึงและใช้ข้อมูลบนเว็บที่เปลี่ยนอยู่ตลอดเวลาได้ เนื้อหาที่น่าสนใจครอบคลุมความหมาย โครงการริเริ่มการจัดเก็บรักษาเว็บถาวรที่เกิดขึ้นในหอสมุดแห่งชาติ และสถาบันสารสนเทศต่าง ๆ ทั่วโลก ประเภทและกระบวนการจัดเก็บรักษาเว็บถาวร รวมถึงเครื่องมือสำคัญและมาตรฐานที่เกี่ยวข้องกับการจัดเก็บรักษาเว็บถาวร

คำสำคัญ: การจัดเก็บรักษาเว็บถาวร คลังเก็บเว็บถาวร แหล่งสารสนเทศ

¹ ภาควิชาบรรณารักษศาสตร์ คณะอักษรศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อีเมล: Wachiraporn.K@chula.ac.th

² ภาควิชาบรรณารักษศาสตร์ คณะอักษรศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อีเมล: Chindarat.B@chula.ac.th

Received 31 March 2019; Revised 2 April 2019; Accepted 15 April 2019

Web Archiving: A New Concept of Information Source Development in the Digital Age

Wachiraporn Klungthanaboon¹

Chindarat Berpan²

Abstract

This article indicates the significance of web archiving as a new concept for developing information sources in the digital age. Web archiving is also a new approach to ensuring the long-term accessibility of dynamic web content. The paper covers the definition of web archiving, web archiving initiatives which started by national libraries and information institutions across the globe, the types and process of web archiving as well as some important tools and relevant standards.

Keywords: Web archiving, Web archives, Information sources

¹ Department of Library Science, Faculty of Arts, Chulalongkorn University Email:

Wachiraporn.K@chula.ac.th

² Department of Library Science, Faculty of Arts, Chulalongkorn University Email: Chindarat.B@chula.ac.th

บทนำ

ปัจจุบันเว็บไซต์ได้กลายเป็นแหล่งบันทึก จัดเก็บ และเผยแพร่สารสนเทศที่มีบทบาทต่อการดำเนินกิจกรรมต่าง ๆ ของผู้คนในสังคมเพิ่มขึ้น หน่วยงานทั้งภาครัฐและเอกชนจำนวนมากพัฒนาเว็บไซต์ เพื่อใช้เป็นช่องทางในการเผยแพร่สารสนเทศ ข่าวสาร องค์ความรู้ และติดต่อสื่อสาร ตลอดจนสร้างปฏิสัมพันธ์กับสาธารณชนและองค์กรอื่น ผู้ประกอบการรุ่นใหม่ต่างก็หันมาใช้เว็บไซต์เป็นช่องทางในการสร้างตลาดออนไลน์เพื่อขายผลิตภัณฑ์และบริการเพิ่มมากขึ้น ในขณะที่นักวิชาการและบุคคลทั่วไปนิยมแบ่งปันความรู้และประสบการณ์ผ่านเว็บไซต์มากขึ้นเช่นกัน ดังจะเห็นได้ว่าการผลิตและเผยแพร่หนังสือ วารสาร นิตยสาร รายงานการวิจัย และหนังสือพิมพ์รูปแบบอิเล็กทรอนิกส์ผ่านเว็บไซต์เพิ่มขึ้นอย่างต่อเนื่อง ในขณะที่การผลิตและเผยแพร่สารสนเทศในรูปแบบสื่อสิ่งพิมพ์ลดลงเรื่อย ๆ แม้แต่ในชีวิตประจำวันยังพบว่าผู้คนนิยมเขียนบล็อกเล่าเรื่องราวที่สนใจบนเว็บไซต์แทนการบันทึกประจำวันในสมุดไดอารี่ และจัดทำแกลลอรีภาพบนเว็บแทนการจัดภาพในอัลบั้มแบบดั้งเดิม ปรากฏการณ์เหล่านี้ยิ่งตอกย้ำให้เห็นความสำคัญของเว็บไซต์ในฐานะแหล่งบันทึก จัดเก็บ และเผยแพร่สารสนเทศในสังคมดิจิทัลได้อย่างชัดเจนยิ่งขึ้น (Gomes & Costa, 2014, p. 106)

ในโลกยุคดิจิทัลเว็บไซต์ถือเป็นแหล่งสารสนเทศที่รวบรวมข้อมูลวิชาการสาขาต่าง ๆ จำนวนมหาศาลที่เข้าถึงได้สะดวกทุกที่ทุกเวลา การนำเสนอสารสนเทศหลากหลายรูปแบบ ไม่ว่าจะเป็นอักษร ภาพ กราฟิก เสียง หรือภาพเคลื่อนไหว สามารถตอบสนองความต้องการของผู้ใช้ได้ทุกกลุ่ม สารสนเทศที่เผยแพร่บนเว็บไซต์สะท้อนให้เห็นปรากฏการณ์ทางสังคมและวิถีชีวิตของคนในสังคมด้านต่าง ๆ ที่เกิดขึ้นได้เป็นอย่างดี ด้วยเหตุนี้เว็บไซต์จึงกลายเป็นแหล่งสารสนเทศที่มีความสำคัญต่อหน่วยงาน องค์กร ปึงเอกบุคคล ในด้านการดำรงชีวิต และพัฒนาการความก้าวหน้าทางวิชาการสาขาต่าง ๆ อย่างต่อเนื่อง

อย่างไรก็ตามเนื่องจากเว็บไซต์มีลักษณะเป็นพลวัต สารสนเทศที่อยู่บนเว็บไซต์จึงมีระยะเวลาจำกัด ไม่คงอยู่ถาวร (Ephemeral) เหมือนกับสื่อสิ่งพิมพ์ จากผลการศึกษาจำนวนมากแสดงให้เห็นว่า หลังเวลาผ่านไปหนึ่งปี มีเว็บเพจถึงร้อยละ 80 หายไปจากรูปแบบดั้งเดิมที่ปรากฏในครั้งแรก (Ntoulas, Cho, & Olston, 2004 cited in Costa, Gomes, & Silva, 2017, p. 191) และรายการอ้างอิงงานวิชาการบนเว็บไซต์ในบทความวิชาการ ถึงร้อยละ 13 สูญหายหลังจากผ่านไปเพียง 27 เดือน (Dellavalle et al., 2003 cited in Costa et al., 2017, p. 191) ดังนั้นสารสนเทศบนเว็บไซต์ที่ยังปรากฏ

ในรูปแบบเดิมนานกว่าหนึ่งปีจึงมีจำนวนน้อยมาก (Costa et al., 2017; Gomes & Costa, 2014, p. 107) ทั้งนี้อาจเป็นเพราะผู้ผลิตและเผยแพร่สารสนเทศไม่มีงบประมาณเพียงพอที่จะดำเนินการจัดเก็บรักษาสารสนเทศทั้งหมดที่เคยเผยแพร่ผ่านเว็บไซต์ กอปรกับมีความจำเป็นที่จะต้องปรับเปลี่ยนสารสนเทศบนหน้าเว็บให้เป็นปัจจุบันอยู่เสมอ เนื่องจากเป็นคุณลักษณะเด่นข้อหนึ่งของเว็บไซต์ ยิ่งไปกว่านั้นมีสารสนเทศที่ทรงคุณค่าจำนวนไม่น้อยเผยแพร่บนเว็บไซต์เพียงแหล่งเดียว ดังนั้นการเปลี่ยนแปลงที่อยู่ การปรับปรุงเนื้อหา หรือการปิดตัวเว็บไซต์ ทำให้สารสนเทศดังกล่าวหายไปด้วยอย่างถาวร ส่งผลให้ผู้ที่มีความต้องการใช้ไม่อาจเข้าถึงสารสนเทศเหล่านั้นได้อีก สภาพเช่นนี้ก่อให้เกิดช่องว่างของความรู้ระหว่างปัจจุบันกับอนาคต และมีผลกระทบต่อพัฒนาการความก้าวหน้าทางวิชาการในสาขาต่าง ๆ ตลอดจนทำให้ความทรงจำเกี่ยวกับเรื่องราวที่ผ่านมาของบุคคลและสังคมเลือนหายไปด้วย ตัวอย่างเช่น การสูญหายของภาพถ่ายที่นำมาแบ่งปันกันดูทางเว็บไซต์ เป็นต้น

ด้วยตระหนักถึงความสูญเสียที่จะเกิดขึ้นจากการที่ความรู้และมรดกทางวัฒนธรรมอันทรงคุณค่าบนเว็บไซต์สูญหายไป หน่วยงานด้านสารสนเทศจำนวนมากจึงเริ่มเห็นความจำเป็นในการอนุรักษ์สารสนเทศออนไลน์บนเว็บไซต์ให้คงอยู่เพื่อให้สามารถเข้าถึงและใช้ประโยชน์ด้านต่าง ๆ ได้ในอนาคต ด้วยวิธีการจัดเก็บรักษาเว็บถาวร (Web archiving) คลังเก็บเว็บถาวร (Web archive) ที่เกิดขึ้นจึงเป็นแหล่งสารสนเทศที่ทวีความสำคัญยิ่งขึ้นเรื่อย ๆ โดยเฉพาะอย่างยิ่งในยุคที่สารสนเทศและเทคโนโลยีสารสนเทศได้เข้ามาเป็นส่วนหนึ่งในการดำเนินชีวิตของผู้คน และเพื่อให้ผู้มีส่วนได้ส่วนเสียตระหนักถึงคุณค่าและเข้าใจภาพรวมของการดำเนินการจัดเก็บรักษาเว็บถาวร บทความนี้จึงมีวัตถุประสงค์ให้ผู้สนใจได้รับความรู้เบื้องต้นเกี่ยวกับการจัดเก็บรักษาเว็บถาวร ทั้งในด้านการทฤษฎี ความเป็นมาและโครงการริเริ่มการจัดเก็บรักษาเว็บถาวรที่เกิดขึ้นในสังคมประเภทและกระบวนการจัดเก็บรักษาเว็บถาวร ตลอดจนเครื่องมือและมาตรฐานที่เกี่ยวข้องกับการจัดเก็บรักษาเว็บถาวร

ความหมายของการจัดเก็บรักษาเว็บถาวร

การจัดเก็บรักษาเว็บถาวร (Web archiving) หมายถึงกระบวนการอันประกอบด้วยการรวบรวมจัดเก็บเว็บไซต์ และสารสนเทศที่มีอยู่บนเว็บไซต์เหล่านั้น และสงวนรักษาไว้เพื่อให้สามารถเข้าถึงและนำมาใช้ประโยชน์ได้เมื่อมีความต้องการในอนาคต แนวคิดนี้มีลักษณะ

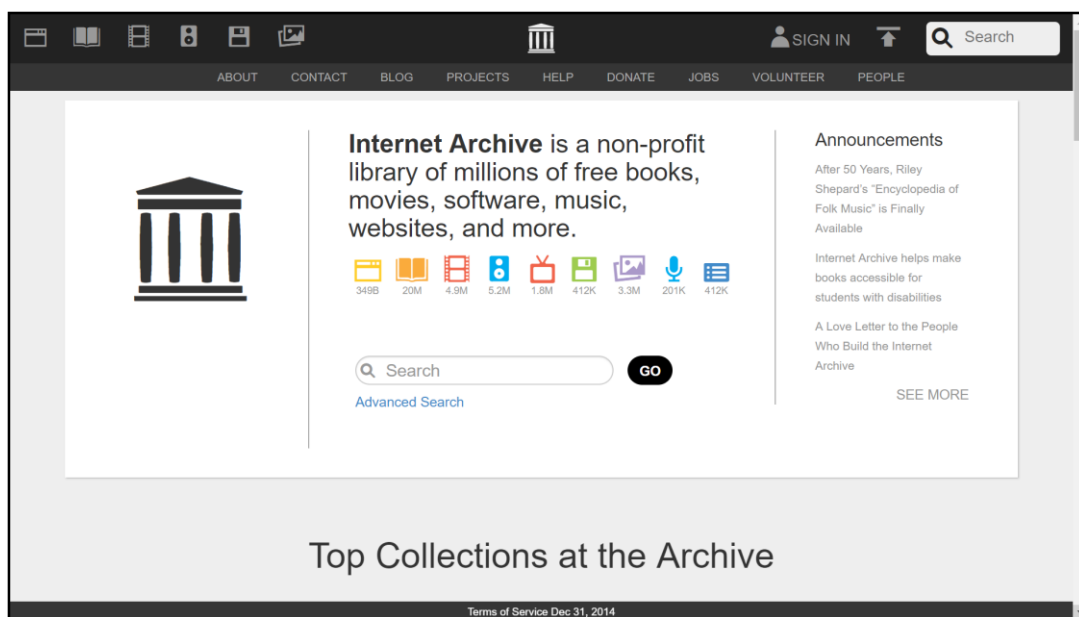
เช่นเดียวกับการอนุรักษ์เอกสารจดหมายเหตุให้นักวิจัย นักประวัติศาสตร์ และประชาชนทั่วไปใช้ประโยชน์ในด้านต่าง ๆ (The National Archives, 2011)

การจัดเก็บรักษาเว็บถาวรนอกจากเก็บรักษาสารสนเทศเว็บดิจิทัล(Born-digital web content) ให้คงอยู่ถาวรเพื่อใช้ประโยชน์ได้ในอนาคตแล้ว ยังมีส่วนช่วยอนุรักษ์เอกสารที่หน่วยงานได้เคยแปลงเป็นดิจิทัลและเผยแพร่ออนไลน์ (Digitized resources) ในภายหลังด้วย ดังนั้นคลังเก็บเว็บถาวรที่พัฒนาขึ้นจึงช่วยให้ผู้ใช้ในอนาคตสามารถสืบค้นเอกสารต่าง ๆ จำนวนมากได้ในเวลาไม่กี่วินาที ก่อให้เกิดการใช้ใหม่ อันจะส่งผลต่อการพัฒนาความก้าวหน้าทางวิชาการทุกด้าน โดยเฉพาะด้านการศึกษาค้นคว้างานวิจัยเชิงประวัติ

โครงการริเริ่มการจัดเก็บรักษาเว็บถาวร

โครงการริเริ่มจัดเก็บรักษาเว็บถาวร (Web archive initiatives) ที่เกิดขึ้นเป็นโครงการแรกคือ "CyberCemetery" เกิดขึ้นในปี 1995 โดยความร่วมมือระหว่าง The University of North Texas Libraries และสำนักพิมพ์รัฐบาลสหรัฐอเมริกา (The U.S. Government Printing Office) โครงการนี้เป็นส่วนหนึ่งของ The Federal Depository Library Program มีวัตถุประสงค์เพื่อให้บริการแก่สาธารณะในการเข้าถึงเว็บไซต์และสิ่งพิมพ์ของรัฐที่หมดอายุการใช้งาน ซึ่งเป็นการสนับสนุนพันธกิจของมหาวิทยาลัยในการจัดหา อนุรักษ์ และให้บริการการเข้าถึงความรู้ที่บันทึกบนเว็บไซต์หน่วยงานของรัฐ โครงการนี้ได้จัดเก็บรักษาสารสนเทศของรัฐบาลที่ครอบคลุมหัวข้อต่าง ๆ ที่สนับสนุนหลักสูตรและความแข็งแกร่งของโปรแกรมการศึกษาของมหาวิทยาลัย อย่างไรก็ตาม ปัจจุบันโครงการนี้ได้หยุดดำเนินการแล้ว ("List of web archiving initiatives," n.d.)

ต่อมาในปี 1996 มีโครงการการจัดเก็บรักษาเว็บถาวรเพิ่มมากขึ้นในภูมิภาคต่าง ๆ เช่น โครงการ The Internet Archive ในประเทศสหรัฐอเมริกา และ โครงการ PANDORA โดยหอสมุดแห่งชาติประเทศออสเตรเลีย เป็นต้น



ภาพ 1 Internet Archive (<https://archive.org/>)

โครงการ The Internet Archive (<https://archive.org/>) เป็นโครงการแรกเริ่มที่จัดเก็บรักษาเว็บถาวรที่ใหญ่และเก่าแก่ที่สุดในโลกที่ยังดำเนินการอยู่จนถึงปัจจุบัน ข้อมูลที่จัดเก็บย้อนหลังตั้งแต่ปี 1996 โครงการนี้ได้พัฒนา Heritrix: Internet Archive Web Crawler ซึ่งเป็นตัวเก็บรวบรวมแบบโอเพนซอร์ส (Open-source crawler) และได้ให้บริการการจัดเก็บรักษาเว็บถาวรหลากหลายบริการ อาทิ Archive-It (<https://archive-it.org/>), Save page now และ Domain level contract crawl เป็นต้น นอกจากนั้นยังให้บริการ Wayback Machine แก่สาธารณะในการเข้าถึงคลังเก็บเว็บถาวร (Web archive) (“List of web archiving initiatives,” n.d.) ปัจจุบันจัดเก็บรักษาทั้งเว็บเพจ หนังสือ คลิปบันทึกเสียง คลิปวิดีโอ ภาพ และโปรแกรมซอฟต์แวร์ การดำเนินงานของโครงการเป็นไปตามพันธกิจในการเป็นองค์กรไม่แสวงหาผลกำไรที่มีเป้าหมายจะนำเสนอการเข้าถึงความรู้ทั้งหมดแบบสากล (Universal Access to All Knowledge) (Pennock, 2013, p. 5)



ภาพ 2 โครงการ PANDORA (<https://pandora.nla.gov.au/>)

ส่วนโครงการ PANDORA (<https://pandora.nla.gov.au/>) ดำเนินการโดยหอสมุดแห่งชาติออสเตรเลีย โครงการนี้ได้พัฒนาระบบ PANDAS (PANDORA Digital Archiving System) เพื่อจัดเก็บรักษาสิ่งพิมพ์ออนไลน์ที่มีเนื้อหาเกี่ยวกับออสเตรเลียหรือผลิตในออสเตรเลีย ในการดำเนินงานมีความร่วมมือกับห้องสมุดอื่น ๆ ตลอดจนองค์กรสำคัญในประเทศออสเตรเลียที่ร่วมกันเก็บรวบรวมสิ่งพิมพ์ออนไลน์ มีบรรณารักษ์เป็นผู้คัดเลือก โดยพิจารณาและคำนึงถึงคุณค่าทางประวัติศาสตร์เป็นสำคัญ (National Library of Australia, 2018; Pennock, 2013)

หลังจากนั้นหอสมุดแห่งชาติในประเทศต่าง ๆ ได้ทยอยริเริ่มดำเนินโครงการการจัดเก็บรักษาเว็บถาวรเพิ่มขึ้น เช่น

หอสมุดรัฐสภาอเมริกัน (Library of Congress) ประเทศสหรัฐอเมริกา ได้พัฒนา The MINERVA (Mapping the Internet Electronic Resources Virtual Archive) ในปี 2000 เพื่อจัดเก็บสิ่งพิมพ์ออนไลน์ที่มีเนื้อหาเกี่ยวข้องกับเหตุการณ์สำคัญด้านต่าง ๆ เช่น การเมือง นโยบายสาธารณะ และ กีฬาโอลิมปิก เป็นต้น (Library of Congress, 2018)

หอสมุดแห่งชาติประเทศสาธารณรัฐเช็ก ได้พัฒนาโครงการ Webarchiv - Archive of the Czech Republic Web ในปี 2000 เพื่ออนุรักษ์เอกสารที่มีคุณค่าด้านวัฒนธรรม ศิลปะ และประวัติศาสตร์ที่เผยแพร่ออนไลน์ โครงการ WebArchiv นี้ได้เข้าร่วมเป็นสมาชิกของภาคีการสงวน

รักษานเทอร์เน็ตสากล (The International Internet Preservation Consortium – IIPC) ตั้งแต่ปี 2007 (National Library of the Czech Republic, 2018)

ส่วนหอสมุดแห่งชาติประเทศญี่ปุ่น (The National Diet Library) ได้พัฒนาโครงการจัดเก็บรักษาเว็บถาวร (Web Archiving Project - WARP) ในปี 2002 กฎหมายหอสมุดแห่งชาติของญี่ปุ่นที่ผ่านการปรับปรุงในปี 2009 และมีผลบังคับใช้ในปี 2010 ได้อนุญาตให้หอสมุดแห่งชาติจัดเก็บเว็บไซต์หน่วยงานราชการของประเทศ ศาล และมหาวิทยาลัย รวมทั้งเว็บไซต์ด้านวัฒนธรรมและเหตุการณ์ระหว่างประเทศที่จัดขึ้นในประเทศญี่ปุ่น และสิ่งพิมพ์ต่อเนื่องออนไลน์ (Online periodicals) ที่ได้รับอนุญาตจากเว็บมาสเตอร์ (National Diet Library, n.d.)

หอสมุดแห่งชาติอังกฤษ (The British Library) เริ่มดำเนินโครงการ UK Web Archive ในปี 2004 โดยรวบรวมเว็บไซต์ที่เผยแพร่งานวิจัยที่สะท้อนความหลากหลายในการดำเนินชีวิต ความสนใจ และกิจกรรม รวมทั้งสารสนเทศในรูปแบบต่าง ๆ เช่น วรรณกรรมที่ไม่ได้เผยแพร่กว้างขวาง (Grey literature) เรื่องย่อ รายงาน และนโยบาย เป็นต้น นอกจากนี้หอสมุดแห่งชาติอังกฤษยังเป็นผู้นำเครือข่ายการจัดเก็บรักษาเว็บถาวรโดยร่วมมือกับห้องสมุดในสหราชอาณาจักร 5 แห่ง ทำหน้าที่รวบรวมทรัพยากรสารสนเทศตามกฎหมายนำฝาก (Legal deposit libraries) ได้แก่ The National Library of Scotland, The National Library of Wales, Cambridge University Library, The Bodleian Library in Oxford, และ The Library of Trinity College Dublin อย่างไรก็ตามการเข้าถึงเว็บที่จัดเก็บจะต้องได้รับอนุญาตจากห้องสมุดแห่งใดแห่งหนึ่งในหกแห่งนี้ (The British Library, n.d.)

Netarkivet โครงการริเริ่มการจัดเก็บรักษาเว็บถาวรในประเทศเดนมาร์ก เริ่มดำเนินการในปี 2005 โดยหอสมุดแห่งชาติ 2 แห่ง¹ ได้แก่ The State and University Library ที่เมืองอาร์ฮุส และ The Royal Library ที่เมืองโคเปนเฮเกน เว็บที่จัดเก็บได้ดำเนินการผ่านกฎหมายนำฝาก (Legal deposit) เรียบร้อยแล้ว (Brügger & Schroeder, 2017, p. 66; Pennock, 2013)

¹ เมื่อวันที่ 1 มกราคม ค.ศ. 2017 หอสมุดแห่งชาติเดิม 2 แห่ง ได้แก่ The State and University Library ที่เมืองอาร์ฮุส และ The Royal Library ที่เมืองโคเปนเฮเกน ได้รวมเป็นหน่วยงานเดียว โดยใช้ชื่อว่า “The Royal Danish Library”

The Nordic Web Archive (NWA) เป็นโครงการความร่วมมือระหว่างหอสมุดแห่งชาติ ฟินแลนด์ ไอซ์แลนด์ เดนมาร์ก นอร์เวย์ และสวีเดน ที่ได้พัฒนาโปรแกรมแบบโอเพนซอร์ซ (Open source) เพื่อช่วยในการจัดเก็บและเข้าถึงคอลเลกชันเว็บ (Pennock, 2013)

นอกจากการริเริ่มโครงการฯ โดยหอสมุดแห่งชาติและหน่วยงานสารสนเทศในประเทศต่าง ๆ แล้ว พัฒนาการอีกขั้นหนึ่งที่แสดงให้เห็นถึงความสำคัญของการจัดเก็บรักษาเว็บถาวรในสังคมโลก คือการจัดตั้งภาคีการสงวนรักษาอินเทอร์เน็ตสากล (IIPC) ในปี 2003 โดยเป็นความร่วมมือระหว่างสถาบันต่าง ๆ ในหลากหลายภูมิภาคของโลก ที่จะช่วยกันอนุรักษ์เนื้อหาบนอินเทอร์เน็ตเพื่อผู้ใช้ในอนาคต รวมทั้งสนับสนุนและส่งเสริมให้เกิดความก้าวหน้าในความร่วมมือระหว่างประเทศด้านการพัฒนามาตรฐาน และเครื่องมือโอเพนซอร์สสำหรับการจัดเก็บรักษาเว็บ (Niu, 2012) ต่อมาในปี 2008 ภาคีการสงวนรักษาอินเทอร์เน็ตสากล (IIPC) ได้จัดพิมพ์ผลการสำรวจข้อมูลจากประเทศสมาชิกซึ่งครอบคลุมประเด็นเกี่ยวกับการดำเนินงานคลังเก็บเว็บถาวร ได้แก่ จำนวนบุคลากร เครื่องมือที่ใช้ กฎหมาย และเกณฑ์ในการคัดเลือกเว็บสำหรับคลังเก็บเว็บถาวร (Gomes, Miranda, & Costa, 2011, p. 409)

นอกจากนี้ยังมีองค์กรไม่แสวงหาผลกำไร The Joint Information System Committee (JISC) ในประเทศสหราชอาณาจักร ที่สนับสนุนเทคโนโลยีดิจิทัลเพื่อการศึกษาและการวิจัย เห็นความสำคัญของการจัดเก็บรักษาเว็บถาวร และได้ศึกษากฎหมายที่เกี่ยวกับการจัดเก็บรักษาทรัพยากรบนอินเทอร์เน็ตในประเทศสหราชอาณาจักร สหภาพยุโรป สหรัฐอเมริกา และออสเตรเลีย เพื่อใช้เป็นแนวทางการกำหนดนโยบายของสหราชอาณาจักร นอกจากนี้ JISC ยังได้ศึกษาความเป็นไปได้ในการจัดเก็บและอนุรักษ์เว็บ การดำเนินงานของ JISC มีอิทธิพลทำให้หน่วยงานสารสนเทศต่าง ๆ ตื่นตัวและดำเนินโครงการจัดเก็บรักษาเว็บถาวรเพิ่มมากขึ้นเรื่อย ๆ (Gomes et al., 2011) ในปี 2010 JISC ได้ศึกษาพฤติกรรมการใช้และความต้องการใช้จดหมายเหตุเว็บ (Archived web resources) ของนักวิจัย เพื่อใช้ในการพิจารณาโยบายการคัดเลือกจัดเก็บรักษาเว็บถาวรของหอสมุดแห่งชาติ และกฎหมายลิขสิทธิ์ (Dougherty et al., 2010)

ในฤดูร้อนปี 2011 เป็นต้นมา The National Archives ในสหราชอาณาจักร ได้เริ่มโครงการทดลองแบบจำลองการจัดเก็บรักษาเว็บถาวรสำหรับหอจดหมายเหตุ 7 แห่ง เพื่อให้มั่นใจว่าสารสนเทศออนไลน์ที่สำคัญ ๆ ได้ถูกจัดเก็บรักษาไว้สำหรับให้ผู้คนรุ่นหลังสามารถเข้าใช้ได้ในอนาคต กระบวนการดำเนินโครงการนี้ถือได้ว่าเป็นแนวทางการปฏิบัติที่เป็นเลิศ สามารถนำมาใช้

เป็นพื้นฐานในการสร้างต้นแบบการจัดเก็บรักษาเว็บถาวรเพื่อส่งเสริมและนำมาซึ่งการให้ความช่วยเหลือโครงการบริการจัดเก็บเว็บถาวรในประเทศต่าง ๆ และควบคู่กันไปกับโครงการนี้ The National Archives ยังได้ดำเนิน โครงการ An automated web crawl of local authority and National Health Services sites (NHS) ในอีกสองปีต่อมา เพื่อจัดเก็บสารสนเทศต่าง ๆ ที่ผลิตในท้องถิ่น ซึ่งครอบคลุมชุดข้อมูลที่ยังไม่ได้จัดเก็บ โดย data.gov.uk (The National Archives, 2011)

หลังจากนั้นจำนวนโครงการริเริ่มการจัดเก็บรักษาเว็บถาวรได้เพิ่มมากขึ้นเรื่อย ๆ ตามลำดับ ประมาณว่าจนถึงกลางปี 2013 มีโครงการการจัดเก็บรักษาเว็บถาวรทั่วโลกประมาณ 64 โครงการ ข้อมูลที่จัดเก็บมีเนื้อหาครอบคลุมด้านประวัติ วัฒนธรรม วิทยาศาสตร์ และสังคม และจำนวนไฟล์ที่จัดเก็บตั้งแต่ปี 1996 จนถึงปี 2014 มีมากกว่า 181 พันล้านไฟล์ (Gomes & Costa, 2014, p. 107)

ประเภทและกระบวนการจัดเก็บรักษาเว็บถาวร

การจัดเก็บรักษาเว็บถาวรสามารถแบ่งตามลักษณะการดำเนินการได้เป็น 2 ประเภทใหญ่ ได้แก่ Micro-archiving และ Macro-archiving (Brügger, 2005, 2011; Masanès, 2006)

Micro-archiving เป็นการจัดเก็บเว็บไซต์ที่มีขอบเขตการดำเนินการขนาดเล็ก (Small scale) ทั้งในด้านพื้นที่ (จำนวนเว็บไซต์ถูกจำกัด) และเวลา (เวลาช่วงใดช่วงหนึ่ง) กล่าวคือ ดำเนินการโดยบุคคลหรือกลุ่มบุคคลที่มีความรู้ทางเทคนิคเกี่ยวกับการจัดเก็บรักษาเว็บถาวรไม่มากนัก ไม่มีคอมพิวเตอร์ที่มีสมรรถนะสูงทั้งกำลังและความจุในการจัดเก็บ มักเป็นการดำเนินการตามความจำเป็นเร่งด่วนเพื่อใช้ในการศึกษา หรือเพื่อการอื่น ๆ เช่น การดำเนินการด้านกฎหมาย และการดำเนินการในช่วงเวลาใดเวลาหนึ่งก่อนที่เว็บไซต์จะถูกเปลี่ยนแปลงหรือลบออก เป็นต้น บุคคลที่ดำเนินการอาจเป็นนักวิจัย นักศึกษา หรือบุคคลทั่วไปที่ต้องการจัดเก็บรักษาเว็บไซต์โดยใช้คอมพิวเตอร์ส่วนบุคคลที่ได้มาตรฐาน ทรัพยากรหรือเนื้อหาที่จัดเก็บจึงสัมพันธ์กับเรื่องใดเรื่องหนึ่งโดยเฉพาะเพื่อตอบสนองวัตถุประสงค์

ส่วน Macro-archiving เป็นการจัดเก็บรักษาเว็บไซต์ที่มีขอบเขตการดำเนินการขนาดใหญ่ โดยหน่วยงานที่มีผู้จัดเก็บแบบมืออาชีพ ซึ่งอาจเป็นหน่วยงานภาครัฐ ภาคเอกชน หน่วยงานระดับชาติ หรือระดับนานาชาติก็ได้ การดำเนินการใช้คอมพิวเตอร์ที่มีกำลังและความจุในการจัดเก็บจำนวนมาก เพื่อจัดเก็บมรดกทางวัฒนธรรมบางส่วนในระดับชาติหรือระดับนานาชาติ ทรัพยากรหรือเนื้อหาที่จัดเก็บมีหลากหลายครอบคลุมหัวข้อต่าง ๆ จำนวนมาก

อย่างไรก็ตาม ไม่ว่าจะเป็นการจัดเก็บรักษาเว็บถาวรประเภทใด กระบวนการดำเนินการก็มีขั้นตอนหลักคล้ายกันซึ่งแบ่งได้เป็น 4 ขั้นตอน คือเริ่มจากการประเมินและคัดเลือก (Appraisal and selection) เว็บไซต์และสารสนเทศที่สอดคล้องกับเกณฑ์ที่กำหนด แล้วจึงดำเนินการจัดหา (Acquisition) ลำดับต่อมาเป็นการจัดระบบและการจัดเก็บ (Organization and storage) การพรรณนา และการทำให้ผู้ต้องการใช้เข้าถึงได้ (Description and access) ซึ่งมักจะเป็นการเข้าถึงผ่านทางเว็บไซต์ที่จัดเก็บรักษาเว็บถาวร ทำให้หน่วยงานที่พัฒนาเว็บไซต์มีโอกาสนำเสนอการเข้าถึงสารสนเทศโดยไม่จำเป็นต้องจัดเก็บบนเว็บไซต์เข้าถึงได้ในปัจจุบัน (Live web) (Niu, 2012; The National Archives, 2011) เห็นได้ว่าขั้นตอนการดำเนินการไม่แตกต่างจากกระบวนการจัดเก็บรักษาทรัพยากรสารสนเทศที่เป็นสื่อสิ่งพิมพ์

- การประเมินและการคัดเลือก (Appraisal and selection) ครอบคลุมการประเมินคุณค่าเนื้อหาสารสนเทศหรือเอกสารต่าง ๆ บนเว็บไซต์ เพื่อพิจารณาว่าเหมาะที่จะจัดเก็บไว้สำหรับใช้ประโยชน์ในอนาคตหรือไม่ และควรเก็บนานเท่าใด เพื่อนำไปสู่การตัดสินใจคัดเลือก การประเมินคุณค่าเอกสารบนเว็บไซต์อาจพิจารณาคุณค่าในเชิงธุรกิจ เชิงหลักฐาน หรือเชิงประวัติ รวมทั้งการพิจารณาความสัมพันธ์เกี่ยวข้องกับเอกสารอื่น ๆ ที่หน่วยงานต้องการจัดเก็บ ส่วนการพิจารณาระยะเวลาที่เหมาะสมสำหรับการจัดเก็บ มักจะขึ้นอยู่กับลักษณะเนื้อหา ตัวอย่างเช่น เอกสารทางการเงินมักจะจัดเก็บอย่างน้อย 7 ปี (Niu, 2012; The National Archives, 2011) นอกจากนั้นการพิจารณาวัตถุประสงค์ของการผลิตเนื้อหาเว็บ และคุณค่าของเอกสาร ก็เป็นอีกแนวทางหนึ่งที่น่าสนใจในการคัดเลือกเว็บหรือสารสนเทศเว็บเข้าคลังเก็บเว็บถาวร (The National Archives, 2011, p. 8)

สำหรับการคัดเลือกเว็บไซต์ยังอาจใช้เกณฑ์ที่กำหนดด้านใดด้านหนึ่งหรือหลายด้านประกอบกัน เช่น การกำหนดชื่อโดเมนเฉพาะที่เป็น .gov หรือ .edu หรือ เฉพาะชื่อโดเมนในประเทศใดประเทศหนึ่งเท่านั้น การกำหนดหัวข้อเนื้อหาและเหตุการณ์ที่สัมพันธ์กับเป้าหมายและวัตถุประสงค์ของโครงการการจัดเก็บรักษาเว็บถาวร การกำหนดเฉพาะประเภทข้อมูลที่จัดเก็บ (Genres) อาทิ จัดเก็บเฉพาะส่วนที่เป็นข้อความ บล็อก หนังสือพิมพ์ หรือ โลกเสมือน (Virtual worlds) หรือ วิดีทัศน์บนยูทูป และการกำหนดให้จัดเก็บเฉพาะไฟล์รูปแบบใดรูปแบบหนึ่ง อาทิ จัดเก็บเฉพาะไฟล์ประเภทจาวาสคริปต์ ฟลิตเอฟ หรือเอชทีเอ็มแอล เป็นต้น (Niu, 2012) อย่างไรก็ตาม เมื่อตัดสินใจคัดเลือกแล้ว เอกสารเหล่านี้จะถูกจัดการแตกต่างกันไปเพราะมีรูปแบบไฟล์

แตกต่างกัน เช่น ไฟล์สเปรดชีต, ไฟล์พีดีเอฟ และ ไฟล์ภาพ เป็นต้น (The National Archives, 2011, p. 8) ดังนั้นโดยหลักการแล้วการคัดเลือกจะขึ้นอยู่กับเกณฑ์ที่กำหนดตามวัตถุประสงค์ในการจัดเก็บรักษาเว็บถาวร หลังจากนั้นหากจะจำแนกตามประเภทข้อมูลที่จัดเก็บ (Genres) หรือรูปแบบไฟล์บางรูปแบบ (File formats) ก็ทำได้ไม่ยากโดยใช้วิธีการคัดเลือกแบบอัตโนมัติ กล่าวคือใช้ซอฟต์แวร์จำแนกประเภทของข้อมูล เช่น เสียง วิดีทัศน์ หรือ ข้อความที่เป็นวารสารออนไลน์ หรือ เป็นบล็อก หรืออาจแยกความแตกต่างระหว่างโพสต์ในบล็อก (Blog posts) กับ คอมเมนต์ (Comments) เป็นต้น นอกจากนี้อาจใช้การจำแนกโดเมนเว็บไซต์ เช่น .gov, .edu, .th, และ .au เป็นต้น สำหรับการพิจารณาคุณภาพและเนื้อหาเว็บที่ได้รับความนิยมนั้นอาจระบุจากจำนวนลิงก์ขาเข้าและผู้เข้ามาเยี่ยมชม (Incoming links and visitors) เช่น จำนวนผู้เข้าชมวิกิทัศน์ออนไลน์ หรืออาจใช้ข้อมูลการจัดอันดับจากผู้ใช้ เป็นต้น

กรณีการคัดเลือกตามหัวข้อเนื้อหาหรือเหตุการณ์นั้น จำเป็นต้องใช้วิจารณ์ญาณของคน ซึ่งอาจเป็นนักวิชาชีพสารสนเทศ หรือให้ผู้เชี่ยวชาญเฉพาะสาขาช่วยคัดเลือก อาทิ The Digital Archive for Chinese Studies (DACHS) ได้เชิญนักวิชาการผู้เชี่ยวชาญด้านจีนศึกษาให้คำแนะนำเว็บไซต์ที่เกี่ยวข้อง เป็นต้น (Lecher, 2006) วิธีการที่ใช้คนช่วยคัดเลือกเป็นวิธีที่ต้องใช้เวลาและมีค่าใช้จ่ายสูง ดังนั้นจึงนิยมใช้กับการจัดเก็บรักษาเว็บถาวรที่มีขนาดเล็ก และเพื่อลดค่าใช้จ่ายในการคัดเลือกที่กระทำโดยคน โครงการจัดเก็บรักษาเว็บถาวรบางโครงการได้รับเอาที่อยู่เว็บที่ผู้ใช้แนะนำ โดยใช้การลงทะเบียนยูอาร์แอล (Registries of URLs) แล้วนำมาคัดเลือก เช่น The Preserving and Accessing Networked Documentary Resources of Australia (PANDORA), The National Taiwan University Library Web Archives (Chen, Chen, & Ting, 2008; National Library of Australia, 2018) และ The UK government web archiving project เป็นต้น สำหรับโครงการ The UK government web archiving นี้ได้คัดเลือกเว็บไซต์ โดยใช้การลงทะเบียนเว็บไซต์ทั้งหมดของรัฐบาลกลางแห่งสหราชอาณาจักร ยูอาร์แอลที่อยู่ในทะเบียนจะถูกเสนอและบำรุงรักษาโดยผู้จัดการเว็บไซต์ เป็นต้น (Spencer, O'Reilly, & Vasile, 2009 cited in Niu, 2012)

การใช้คนคัดเลือกเว็บไซต์อาจทำได้เร็วขึ้นด้วยการใช้ทฤษฎีการประเมินระดับแมโคร (The macro-appraisal theory) ในสาขาการจัดการจดหมายเหตุ ซึ่งเป็นการประเมินและการคัดเลือกทรัพยากรเว็บโดยการรวบรวมกลุ่มเว็บเพจแทนที่จะรวบรวมเว็บเพจแต่ละเพจ (Pearce-Moses, Kaczmarek, & Pearce-Moses, R. Kaczmarek, 2005) เพื่อลดปัญหาและทำให้กระบวนการ

ประเมินมีประสิทธิภาพยิ่งขึ้น ตัวอย่างเช่น บรรณารักษ์ หรือนักจดหมายเหตุประเมินคุณค่าเว็บไซต์ ทั้งเว็บแทนเว็บเพจแต่ละเพจเพื่อตัดสินใจว่าจะจัดเก็บรักษาเว็บไซต์นั้น ๆ หรือไม่ เป็นต้น

- การจัดหา (Acquisition) วิธีการจัดหาเพื่อให้ได้เว็บไซต์มาจัดเก็บรักษาขึ้นอยู่กับปัจจัยหลายประการ เช่น ขนาดของคลังจัดเก็บตามข้อตกลงรักษาเว็บ ความสัมพันธ์ระหว่างแหล่งจัดเก็บรักษาเว็บกับเจ้าของเว็บไซต์ และลักษณะของเนื้อหาเว็บที่จัดเก็บ เป็นต้น อย่างไรก็ตาม วิธีการจัดหาที่โครงการริเริ่มจัดเก็บรักษาเว็บถาวรใช้มีหลากหลายแตกต่างกันไป เช่น การถ่ายโอนข้อมูลโดยตรง (Direct transfer) จากหน่วยงานของรัฐ หรือสำนักพิมพ์ตามข้อตกลงความร่วมมือ การแปลงข้อมูล (Convert data) จากเจ้าของฐานข้อมูลตามแบบแผนมาตรฐานเปิด อาทิ มาตรฐาน XML โดยใช้เครื่องมือต่าง ๆ เช่น DeepArc เว็บครอว์เลอร์ (Web crawlers) รวบรวมเนื้อหาจากเว็บเซิร์ฟเวอร์ เป็นต้น

อย่างไรก็ตามที่ผ่านมาหอสมุดและหอจดหมายเหตุแห่งชาติต่าง ๆ ได้รับทรัพยากรสารสนเทศผ่านการส่งต่อจากหน่วยงานรัฐ จากการบริจาค และจากสำนักพิมพ์ตามที่กฎหมายกำหนด วิธีการเช่นนี้ยังคงประยุกต์ใช้กับการจัดเก็บรักษาเว็บถาวรได้ เช่น The U.S. National Archives and Records Administration (NARA) ได้ร้องขอให้หน่วยงานของรัฐบาลกลางทุกแห่งส่งเนื้อหาพร้อมในเว็บไซต์หน่วยงานให้แก่ NARA เมื่อประธานาธิบดีคลินตันหมดวาระการดำรงตำแหน่ง (Bellardo, 2001)

สำหรับวิธีการจัดหาโดยใช้เว็บครอว์เลอร์ เป็นวิธีที่นักจดหมายเหตุเว็บ (Web archivist) นิยมใช้ กระบวนการนี้ประกอบด้วย การเก็บเกี่ยวเว็บไซต์ที่เข้าถึงได้ในปัจจุบัน (Live web) จากที่อยู่เว็บ เว็บครอว์เลอร์จะวิ่งไปยังเว็บไซต์แต่ละเว็บและได้เรียงดูเนื้อหาภายในเว็บไซต์นั้น เพื่อทำสำเนาและจัดเก็บสารสนเทศที่พบในขณะเคลื่อนที่ไป เว็บไซต์และสารสนเทศที่ถูกจัดเก็บเป็นจดหมายเหตุออนไลน์และเป็นส่วนหนึ่งของคอลเลกชันในคลังเก็บรักษาเว็บถาวร ผู้ใช้สามารถดู อ่าน และ สืบหาเนื้อหาได้เหมือนเดียวกับเว็บที่เข้าถึงได้ในปัจจุบัน ต่างกันเพียงแต่เว็บไซต์เหล่านี้ถูกจัดเก็บรักษาเป็นภาพถ่ายสารสนเทศ (Snapshots) ณ ขณะนั้น (The National Archives, 2011)

ลักษณะการทำงานของเว็บครอว์เลอร์สอดคล้องไปตามพารามิเตอร์หรือเกณฑ์การคัดเลือกเพื่อจัดเก็บเว็บตามที่ได้กำหนดไว้ เสมือนเป็นตัวกรองเฉพาะเอกสารเว็บที่ต้องการจัดเก็บรักษา เมื่อได้รายชื่อเว็บที่ตรงกับเกณฑ์แล้ว จึงดาวน์โหลดเอกสารเว็บนั้นมาจัดเก็บไว้และคอย

ติดตามเว็บนั้นเพื่อค้นหาและดาวน์โหลดเอกสารเว็บหรือเนื้อหาเว็บที่มีการเปลี่ยนแปลงเพื่อจัดเก็บเข้าคลังเก็บรักษาเว็บถาวรเพิ่มเติมตามความถี่ที่กำหนด เช่น หอสมุดแห่งชาติฝรั่งเศสมีนโยบายคัดเลือกเว็บไซต์ทุกเว็บไซต์ที่มีโดเมนระดับบนสุดตามรหัสประเทศเป็น .fr และ .re เป็นหลัก รวมทั้งเว็บไซต์ที่มีโดเมนระดับบนสุดอื่นที่ผู้พัฒนาอยู่ในประเทศฝรั่งเศส และเว็บไซต์ที่เผยแพร่เนื้อหาซึ่งแสดงให้เห็นว่าผลิตในประเทศฝรั่งเศส (Lasfargues, Oury, & Wendland, 2008) จึงกำหนดให้เว็บครอว์เลอร์จัดเก็บเว็บตามพารามิเตอร์เหล่านี้ เป็นต้น

การรวบรวมเว็บไซต์เพื่อจัดเก็บโดยใช้เว็บครอว์เลอร์กำลังเข้ามาแทนที่การรับฝาก (Depositing) อาทิ The Arizona State Library ได้เปลี่ยนวิธีการรวบรวมจากการรอส่งมาจากหน่วยงานรัฐเป็นการใช้เว็บครอว์เลอร์ (Pearce-Moses et al., 2005) Masanès (2005, pp. 72–73) ว่าการใช้เว็บครอว์เลอร์เพื่อเก็บรวบรวมเนื้อหานั้นช่วยให้สามารถจัดหาเนื้อหาได้จำนวนมากโดยเสียค่าใช้จ่ายน้อย

แม้ว่าเว็บครอว์เลอร์จะเป็นเครื่องมือสำรวจและค้นพบเว็บไซต์ใหม่ผ่านลิงก์ที่เป็นระบบและมีสมรรถนะสูง โดยเฉพาะเมื่อโครงการเริ่มจากกลุ่มเว็บไซต์ที่คัดเลือกไว้จำนวนไม่มากนัก แต่ก็มีข้อจำกัดที่พึงตระหนัก เช่น บางเว็บไซต์ป้องกันการเข้าถึงโดยให้แสดงตัวตนด้วยรหัสผ่านหรืออินเทอร์เน็ตโปรโตคอล (IP) หรือ ทรัพยากรเว็บบางส่วนต้องใช้คนในการรวบรวม เพราะเว็บครอว์เลอร์บางตัวไม่อาจเก็บเกี่ยวข้อมูลบางประเภท เช่น ข้อมูลในไฟล์ GIS ซึ่งมีเนื้อหาเว็บเป็นพลวัต และ สื่อสตรีมมิ่ง เป็นต้น

ดังนั้นการแสวงหาความร่วมมือจากเจ้าของเว็บอาจเป็นอีกวิธีหนึ่งที่จะช่วยให้การจัดหาทรัพยากรเว็บเพื่อเก็บในคลังเก็บเว็บถาวรเกิดประสิทธิผลมากขึ้น อย่างไรก็ตามในการติดต่อขออนุญาตเพื่อรวบรวมเว็บที่จะนำมาจัดเก็บจากเจ้าของลิขสิทธิ์นั้นควรคำนึงถึงสภาพแวดล้อมทางกฎหมายที่เกี่ยวข้องกับการจัดเก็บรักษาเว็บถาวรด้วย สำหรับประเทศที่มีกฎหมายการนำฝากทรัพยากรสารสนเทศที่ครอบคลุมทรัพยากรเว็บ (Electronic legal deposit) หอสมุดและหอจดหมายเหตุมีอำนาจในการรวบรวมจัดเก็บดูแลเอกสารสาธารณะ จึงไม่จำเป็นต้องขออนุญาตจากผู้ผลิตเอกสาร เช่น หอสมุดในประเทศนิวซีแลนด์ The U.S. National Archives and Records Administration (NARA) และหอจดหมายเหตุของรัฐในสหราชอาณาจักร เป็นต้น แต่บางครั้งในสภาพแวดล้อมทางกฎหมายเดียวกันนี้ โครงการจัดเก็บรักษาเว็บถาวรขนาดเล็กอาจประสบความยุ่งยากมากกว่า เพราะ

โครงการขนาดใหญ่สามารถบริหารจัดการการขออนุญาตจากเจ้าของสิทธิ์ได้มีประสิทธิภาพมากกว่า ทั้งนี้เนื่องจากความพร้อมด้านทรัพยากร เวลา และเครือข่าย

- การจัดระบบและจัดเก็บ (Organization and storage) การดำเนินงานในขั้นตอนการจัดระบบและจัดเก็บ ครอบคลุมการวางแผนเพื่อจัดเก็บชั่วคราวหรือจัดเก็บระยะยาว การกำหนดแผนการสำหรับข้อมูลที่เก็บรักษา ขั้นตอนการเก็บรักษา และการจัดระบบ สำหรับหน่วยงานบางแห่งอาจจะประกอบด้วยกิจกรรมการสงวนรักษาด้วย ในขั้นตอนนี้ ผู้ดำเนินการต้องดูแลรักษาเนื้อหาบนเว็บที่จัดเก็บให้คงความถูกต้อง และครบถ้วนตามวัตถุประสงค์ของการเก็บรวบรวมที่ได้กำหนดไว้ ซึ่งอาจแตกต่างกันไป เช่น การดูแลรักษาเฉพาะเนื้อหาวิชาการ และการดูแลรักษาเนื้อหาในฐานะเป็นหลักฐานทางกฎหมาย เป็นต้น นอกจากนั้น โครงสร้างและบริบทของทรัพยากรสารสนเทศเว็บก็ต้องดูแลรักษาเช่นกัน โดยอาจเทียบเคียงกับการจัดการจดหมายเหตุแบบดั้งเดิม ซึ่งครอบคลุมบริบทเอกสารจดหมายเหตุ แหล่งกำเนิด¹ และลำดับที่ต้นฉบับเอกสาร²

แม้ว่าการจัดเก็บรักษาเว็บไซต์จำนวนมากมักจะดูแลรักษาเนื้อหาเว็บในฐานะทรัพยากรสารสนเทศมากกว่าในฐานะหลักฐานทางกฎหมาย แนวคิดเกี่ยวกับแหล่งกำเนิดก็ยังนำมาประยุกต์ใช้กับทรัพยากรเว็บที่นำมาจัดเก็บได้

- การพรรณนาและการเข้าถึง (Description and access) หมายถึงการพรรณนาทรัพยากรสารสนเทศที่ดำเนินการจัดเก็บรักษา โดยใช้เมทาเดตาเพื่อเอื้อประโยชน์ในการสืบค้นและเข้าถึง และช่วยให้ทราบบริบทอื่น ๆ ที่สำคัญของทรัพยากรสารสนเทศเว็บที่จัดเก็บ ทำให้สามารถจัดการทรัพยากรสารสนเทศที่มีหลายเวอร์ชันได้ชัดเจน ในการจัดเก็บรักษาเว็บถาวร การสร้างเมทาเดตาและความสมบูรณ์ของเมทาเดตาที่สร้างขึ้นมีความแตกต่างกัน ซึ่งขึ้นอยู่กับขนาดของคลังเก็บรักษาเว็บและทรัพยากรที่มีอยู่ในหน่วยงานที่ดำเนินการจัดเก็บรักษาเว็บ

สำหรับคลังเก็บรักษาเว็บขนาดใหญ่มักจะใช้วิธีการสร้างเมทาเดตาอัตโนมัติ ข้อมูลเมทาเดตาบางส่วน เช่น การประทับเวลา (The timestamp) จะถูกสร้างเมื่อทรัพยากรเว็บถูกเก็บเกี่ยว

¹ แหล่งกำเนิด หมายถึงสารสนเทศเกี่ยวกับแหล่งที่มาของเอกสาร เช่น ที่อยู่ของเว็บไซต์ (URL) เนื้อหาหรือผู้ผลิตเอกสาร การดำเนินการ หรือเหตุการณ์ที่ทำให้เกิดการผลิตเอกสารดังกล่าว และห่วงโซ่การดูแลรักษา

² ลำดับที่ต้นฉบับเอกสาร หมายถึงการจัดเรียงในลักษณะที่แสดงให้เห็นความสัมพันธ์ในหมู่เอกสารด้วยกัน

ปริมาณข้อมูล (Bytes) ยูอาร์แอล ประเภทสื่อ (Internet Media Type)¹ สามารถสร้างหรือดึงมาโดยใช้เว็บเบราว์เซอร์ หากเป็นคลังเก็บเว็บขนาดเล็ก บุคคลสามารถสร้างเมตาดาตาเอง โดยใช้มาตรฐานดับลินคอร์เมตาดาตา คู่มือการให้หัวเรื่องหอสมุดรัฐสภาอเมริกัน หรือรายการควบคุมที่พัฒนาขึ้นเอง

การสร้างเมตาดาตาที่ใช้ในการจัดเก็บรักษาทรัพยากรเว็บ แบ่งได้เป็น 3 ระดับ ได้แก่ ระดับคอลเลกชัน (Collection level) ระดับ Seed (Seed level) และระดับเอกสาร (Document level) (“Archive-It user guide,” 2018)

เมตาดาตาระดับคอลเลกชัน หมายถึงการอธิบายขอบเขตของคอลเลกชันอย่างกว้าง ๆ ว่าครอบคลุมเนื้อหาเกี่ยวกับอะไร ให้ข้อมูลแก่ผู้ใช้ในการค้น เช่น ชื่อคอลเลกชัน (Collection Title) หัวข้อเรื่อง (Collection topics) ชื่อผู้สร้าง/ผู้รับผิดชอบ (Creator) หัวเรื่อง (Subject) และคำอธิบาย (Description) เป็นต้น

เมตาดาตาระดับ Seed หมายถึงการระบุยูอาร์แอลที่หน่วยงานกำหนดให้เว็บเบราว์เซอร์ไปรวบรวมสิ่งต่าง ๆ ที่อยู่ภายใต้ยูอาร์แอลนั้น อาจจะเป็นได้ทั้ง 1) เว็บไซต์ทั้งเว็บ เช่น <http://www.arts.chula.ac.th/> เป็นต้น 2) ส่วนใดส่วนหนึ่งของเว็บไซต์ เช่น <http://www.arts.chula.ac.th/~research-arts/> เป็นต้น และ 3) เอกสารรายการใดรายการหนึ่ง เช่น <http://www.arts.chula.ac.th/~research-arts/images/year59/quality/CDS2559.xls> เป็นต้น หรือหน่วยข้อมูลย่อยอื่น ๆ ใน Dublin Core Metadata Element Set เช่น ชื่อเรื่อง (Title) ชื่อผู้สร้าง/ผู้รับผิดชอบ (Creator) หัวเรื่อง (Subject) และคำอธิบาย (Description) เป็นต้น

ส่วน เมตาดาตาระดับเอกสาร หมายถึงการพรรณนาเอกสารแต่ละรายการ โดยระบุยูอาร์แอลเฉพาะให้แก่เนื้อหาเว็บที่จัดเก็บรวบรวมแต่ละรายการ และหน่วยข้อมูลย่อยอื่น ๆ ใน Dublin Core Metadata Element Set เช่น ชื่อเรื่อง (Title) ชื่อผู้สร้าง/ผู้รับผิดชอบ (Creator) หัวเรื่อง (Subject) และคำอธิบาย (Description) เป็นต้น

สำหรับความสามารถในการเข้าถึงคลังจัดเก็บเว็บ การกำหนดนโยบายการเข้าถึงการใช้และการนำมาใช้ใหม่มีความสำคัญ เพราะมีผลต่อความสำเร็จของโครงการจัดเก็บรักษาเว็บ

¹ ประเภทสื่อ (Internet Media Type) ประกอบด้วย) ส่วน คือ ชนิดของไฟล์ 2MIME Type และประเภท (ของเนื้อหา) Content type ซึ่งเป็นมาตรฐานสำหรับการแลกเปลี่ยนข้อมูลบนอินเทอร์เน็ต (เช่น text/html application/pdf เป็นต้น

ถาวร หน่วยงานที่ดำเนินการต้องพิจารณาวิธีการที่จะเอื้อให้ผู้เข้าใช้เข้าถึงแบบเปิด รวมทั้งวิธีการส่งเสริมการใช้คอลเลกชัน และวิธีการควบคุมการใช้สาธารณะ การจัดการกระบวนการเหล่านี้ถือเป็นเป้าหมายเบื้องต้นในขั้นตอนการเข้าถึง การใช้ และการนำมาใช้ใหม่ นโยบายด้านการเข้าถึงที่กำหนดควรครอบคลุมการเลือกเทคโนโลยีหรือเครื่องมือเฉพาะสำหรับผู้เข้าใช้เข้าถึงเว็บเพจที่จัดเก็บ เช่น การใช้ซอฟต์แวร์ Wayback เป็นเครื่องมือแบบเปิดให้ผู้เข้าใช้สามารถเรียกดูเว็บเพจที่จัดเก็บได้เสมือนกับได้ดูเว็บเพจที่เข้าถึงได้ในปัจจุบัน การขอใช้เนื้อหาที่จำกัดอินเทอร์เน็ตโปรโตคอลเพื่อให้เกิดความคล่องตัวในการเข้าถึงจากสถานที่ที่กำหนดเฉพาะในหน่วยงานที่ให้บริการ และการให้ผู้เข้าใช้เชื่อมโยงคอลเลกชันจากไซต์ของตนเองผ่านกล่องค้นหา (Search box) หรือการให้เชื่อมโยงกับซอฟต์แวร์ Wayback เป็นต้น (Bragg & Hanna, 2013)

ในการเข้าถึงเนื้อหาของโครงการจัดเก็บรักษาเว็บถาวรบางโครงการ หน่วยงานอนุญาตให้สาธารณชนเข้าถึงเนื้อหาที่จัดเก็บได้ แม้ว่าเนื้อหาบางส่วนกำหนดไว้ว่าต้องเก็บไว้ก่อนระยะเวลาหนึ่ง บางแห่งให้เข้าถึงเฉพาะยูอาร์แอล หรือเฉพาะโดเมนที่เป็นคอลเลกชันของบุคคลแต่ละบุคคล (Bragg & Hanna, 2013) กล่าวได้ว่าลักษณะและสภาพการเข้าถึงมักจะขึ้นอยู่กับสภาพแวดล้อมทางกฎหมายของประเทศเจ้าของโครงการที่ดำเนินการจัดเก็บรักษาเว็บถาวร (Niu, 2012) ตัวอย่างเช่น

กฎหมายการนำฝากทรัพย์สินทางปัญญาของประเทศนิวซีแลนด์ (The legal deposit legislation of New Zealand) อนุญาตให้หอสมุดแห่งชาตินิวซีแลนด์ จัดเก็บรักษาเอกสารต่าง ๆ ที่เผยแพร่บนเว็บไซต์นิวซีแลนด์และอนุญาตให้ผู้เข้าใช้เข้าถึงเอกสารที่จัดเก็บรักษาบนเว็บไซต์ดังกล่าวได้ (National Library of New Zealand, 2018)

หอสมุดรัฐสภาอเมริกัน จัดทำรายการบรรณานุกรมเว็บไซต์ทั้งหมดที่จัดเก็บเพื่อการเข้าถึงสาธารณะโดยให้เข้าถึงได้เฉพาะเว็บเพจที่ผู้ผลิตอนุญาตเท่านั้น (Grotke & Jones, 2010)

นอกจากนี้ความสามารถในการค้นหาสารสนเทศจากคลังเก็บเว็บถาวรยังแตกต่างกันไปตามความสมบูรณ์ของเมทาตาและเครื่องมือที่ใช้ในการสืบค้นและการจัดทำดัชนี เช่น การจัดเก็บรักษาเว็บถาวรของหอสมุดรัฐสภาอเมริกันและนิวซีแลนด์สนับสนุนการสืบค้นผ่านรายการควบคุม ดังนั้นจึงใช้หัวเรื่องในรายการบันทึกเมทาตาที่เว็บไซต์ซึ่งถูกจัดเก็บ ตรงกันข้ามกับเว็บไซต์ที่ใช้ Wayback Machine จัดเก็บ สามารถสืบค้นได้เฉพาะทางยูอาร์แอลเท่านั้น ในขณะที่

เว็บซึ่งจัดเก็บโดยใช้โปรแกรมสืบค้น NutchWax จะสนับสนุนการสืบค้นเนื้อหาเต็ม (Full-text search) โดยมีส่วนต่อประสานการเข้าถึงขั้นสูงบางส่วนถูกสร้างขึ้น (Niu, 2012)

สำหรับขอบเขตในการให้เข้าถึงสารสนเทศที่รวบรวมนั้น The National Archives (2011, pp. 10–11) ได้แนะนำว่า ควรอยู่บนพื้นฐานการพิจารณาเหตุผลในการจัดเก็บรักษาเว็บถาวร และลักษณะการได้รับอนุญาตให้เข้าถึงว่าเป็นแบบเปิด แบบจำกัด หรือแบบปิด กล่าวคือ หากทรัพยากรเว็บถูกจัดเก็บเนื่องจากเหตุผลที่เกี่ยวข้องกับการปฏิบัติตามกฎหมาย การให้เข้าถึงและให้ใช้มักจะจำกัดเฉพาะภายในหน่วยงาน และองค์กรที่รับผิดชอบ หากเป็นองค์กรด้านมรดกวัฒนธรรมมักจะอนุญาตให้นักวิจัยเข้าถึงได้ในสถานที่และเข้าถึงทางออนไลน์ นอกเสียจากว่าเป็นเนื้อหาที่ถูกกำหนดขอบเขตการใช้ นอกจากนั้นประเด็นด้านลิขสิทธิ์และการออกใบอนุญาตอาจเป็นอุปสรรคต่อการเข้าถึงเป็นจำนวนมาก ๆ ได้ ในหลายประเทศได้ขยายการใช้กฎหมายการนำฝากทรัพยากรสารสนเทศเพื่ออนุญาตให้สามารถจัดเก็บรักษาเว็บถาวรทั้งประเทศ และอนุญาตให้ใช้ทรัพยากรสารสนเทศที่จัดเก็บเฉพาะในสถานที่ที่เป็นห้องอ่านของหน่วยงานรวบรวมระดับชาติที่เกี่ยวข้อง ในด้านการส่งเสริมการใช้ หน่วยงานที่ดำเนินการอาจสร้างเว็บท่า (Robust portal) เพื่อบริการข้อมูลเกี่ยวกับการจัดเก็บรักษาเว็บให้แก่สาธารณะ นักวิชาชีพสารสนเทศและผู้สอนการใช้เว็บดังกล่าว

นอกจากการดำเนินการทั้งสี่ขั้นตอนที่ได้กล่าวมาแล้ว กระบวนการพัฒนาการจัดเก็บรักษาเว็บถาวรยังควรคำนึงถึงการรักษาคุณภาพคลังเก็บรักษาเว็บอย่างยั่งยืน โดยจัดการประกันคุณภาพและการวิเคราะห์ (Quality assurance and analysis) (Bragg & Hanna, 2013)

การประกันคุณภาพและการวิเคราะห์ (Quality Assurance and Analysis) หมายถึงการดำเนินการทบทวนคอลเลกชันที่จัดเก็บรักษาว่าส่งผลให้เกิดความสำเร็จตามเป้าหมายที่กำหนด ณ จุดแรกเริ่มของการพัฒนาโครงการจัดเก็บรักษาเว็บถาวรหรือไม่ กล่าวคือ หลังจากจัดเก็บข้อมูลจากเว็บไซต์ที่ต้องการแล้ว หน่วยงานควรทบทวนเนื้อหาที่ได้จัดเก็บ พิจารณาคุณภาพ และความครบถ้วนสมบูรณ์ของเนื้อหา การดำเนินการอาจทำผ่านรายงานที่ผลิตจากเว็บครอว์เลอร์หรือโดยการ Clicking ผ่านเว็บไซต์ที่จัดเก็บเองด้วยเครื่องมือการเข้าถึงคล้ายโปรแกรม Wayback

เครื่องมือสำคัญและมาตรฐานที่เกี่ยวข้องกับการจัดเก็บรักษาเว็บถาวร

กระบวนการพัฒนากลังเก็บเว็บถาวรจำเป็นต้องใช้เครื่องมือหรือโปรแกรมประเภทต่าง ๆ (Web archiving tools/software) ช่วยดำเนินการเพื่อให้เกิดประสิทธิภาพและประสิทธิผล เครื่องมือ

ที่ใช้สามารถแบ่งได้เป็น 4 ประเภทตามกระบวนการจัดเก็บ ได้แก่ 1) เครื่องมือสำหรับการจัดหาเว็บ (Acquisition tools) 2) เครื่องมือสำหรับการดูแลรักษาข้อมูล (Curator tools) 3) เครื่องมือสำหรับการเข้าถึงและการค้นหา (Access and finding aids) และ 4) เครื่องมือสำหรับการจัดการรูปแบบเว็บถาวร (Web archiving formats) (Ball, 2010)

1) เครื่องมือสำหรับการจัดหาเว็บ (Acquisition tools) โปรแกรมที่นิยมเลือกใช้กัน เช่น

Heritrix (<http://crawler.archive.org>) พัฒนาโดย Internet Archive ร่วมกับ Nordic National Libraries ในปี 2002 Heritrix เป็นโปรแกรมรวบรวมเว็บที่มีคุณภาพ เป็นโอเพนซอร์ซ จึงสามารถแก้ไขและดัดแปลง Source code ได้ ต่อมาได้มีการพัฒนามาอย่างต่อเนื่องและเปลี่ยนเป็น Archive-It (www.archive-it.org) ในปี 2006

HTTrack (<http://www.httrack.com>) พัฒนาโดย Xavier Roche and others เป็นโปรแกรมอรรถประโยชน์ ช่วยค้นหาเว็บแบบออฟไลน์ที่ใช้ง่ายและไม่เสียค่าใช้จ่าย

2) เครื่องมือสำหรับการดูแลรักษาข้อมูล (Curator tools) ตัวอย่างเครื่องมือที่ใช้ดูแลรักษาข้อมูล เช่น

Web Curator Tool (WCT) (<http://webcurator.sourceforge.net>) พัฒนาโดยหอสมุดแห่งชาตินิวซีแลนด์ หอสมุดแห่งชาติอังกฤษ และภาควิชาการสงวนรักษาดิจิทัล (IIIPC) เพื่อใช้จัดการกระบวนการเก็บเกี่ยวเว็บที่ได้คัดเลือกแล้ว เหมาะสำหรับผู้ที่ไม่มีความรู้ด้านเทคนิค (Ball, 2010) Web Curator Tool (WCT) สนับสนุนการจัดเก็บรักษาเว็บถาวรทุกขั้นตอน ทั้งการคัดเลือก การอนุญาต การพรรณนา การเก็บเกี่ยว และการประกันคุณภาพ ประกอบด้วยโมดูลหลัก 3 โมดูล ได้แก่ 1) WCT Core Component เป็น Web server ที่จัดลำดับงาน (Scheduler) และขั้นตอนหรือองค์ประกอบหลักอื่น ๆ ในการจัดเก็บรักษาเว็บถาวร เช่น การอนุญาต (Authorization) การแจกแจง (Identification) การจัดลำดับตารางเวลา (Scheduling) ส่วนต่อประสานกับผู้ใช้ (User interface) และ หน่วยที่ดูแลการเก็บเกี่ยว (Harvest agent) เป็นต้น 2) Harvester – Heritrix ทำงานร่วมกับ Core ในฐานะหน่วยดูแลการเก็บเกี่ยว (Harvest agent) นอกจากนั้นยังเป็นแอปพลิเคชันอิสระสำหรับเก็บรวบรวมข้อมูล (Crawl) และเก็บเกี่ยว (Harvest) เว็บไซต์ตามคำสั่งของนักจดหมายเหตุ และ 3) Wayback เป็นแอปพลิเคชันอิสระที่พัฒนาโดยหน่วยงาน Internet Archive ทำงานในส่วนที่ติดต่อผู้ใช้เพื่อเข้าถึงเนื้อหาที่จัดเก็บรักษาไว้ (Pennock, 2013)

NetarchiveSuite (<http://netarchive.dk/suite>) พัฒนาโดย Royal Library and The State and University Library ของ Netarchive.dk ประเทศเดนมาร์ก ใน ค.ศ. 2004 หลังจากนั้นตั้งแต่ ค.ศ. 2008 หอสมุดแห่งชาติประเทศฝรั่งเศส (Bibliothèque nationale de France) และ The Österreichische Nationalbibliothek ได้พัฒนาเพิ่มเติม NetarchiveSuite เป็นแอปพลิเคชันการจัดเก็บรักษาเว็บถาวรที่ช่วยบรรณารักษ์ในการให้ความหมายและควบคุมการเก็บเกี่ยววัตถุเว็บ (Web material) เหมาะสำหรับการเก็บเกี่ยวขนาดเล็กไปจนถึงการเก็บเกี่ยวขนาดใหญ่ทั้งโดเมนประเทศ (National domains) ระบบนี้สามารถใช้กับเครื่องระบบปฏิบัติการต่าง ๆ ได้หลายระบบ และมีโมดูลการจัดเก็บที่ปลอดภัย สามารถจัดการสำเนาหลายอย่าง รวมทั้งการประกันคุณภาพ (Ball, 2010; Pennock, 2013) NetarchiveSuite ประกอบด้วย 4 โมดูลหลัก ได้แก่ 1) Core Netarchive suite เป็นโมดูลที่เสนอกรอบการทำงานภาพรวมและส่วนต่อประสานกับผู้ใช้ที่สนับสนุนการจัดเก็บรักษาเว็บถาวร 2) Harvester module ใช้ Heritrix เป็นโปรแกรมหลักในการเก็บเกี่ยวรวบรวมเนื้อหา 3) โมดูลการเข้าถึงสำหรับจัดการเอกสารจดหมายเหตุของระบบการจัดเก็บรวบรวม (Access module for archival management of the crawl) และ 4) โมดูลการเข้าถึงเพื่อการเข้าถึง Wayback (Access module for access)

PANDAS (PANDORA Digital Archiving System) (<http://pandora.nla.gov.au/pandas.html>) พัฒนาโดยหอสมุดแห่งชาติออสเตรเลีย และเริ่มใช้งานใน ค.ศ. 2001 PANDAS เป็นระบบจัดการบนเว็บ (Web-based management system) ที่ช่วยดำเนินงานเกี่ยวกับการจัดเก็บ (Archiving) และการสงวนรักษาเอกสารออนไลน์ เหมาะสำหรับห้องสมุดต่าง ๆ ที่จะนำไปใช้บนพื้นฐานการคั่นทุน (Ball, 2010; Pennock, 2013) เนื่องจากระบบ PANDAS เขียนด้วยภาษาจาวา (Java) และเพิร์ล (Perl) จึงทำให้มีส่วนต่อประสานที่เป็นมิตรกับผู้ใช้งานที่จัดการขั้นตอนการจัดเก็บรักษาเว็บถาวร ทั้งการคัดเลือก การอนุญาต การจัดเวลา การเก็บเกี่ยว การประกันคุณภาพ การสงวนรักษา และการเข้าถึง ระบบนี้ประกอบด้วย 3 โมดูล ได้แก่ 1) PANDAS เป็นส่วนต่อประสานกับผู้ใช้เพื่อจัดการกิจกรรมต่าง ๆ ที่เกิดขึ้นระหว่างการจัดเก็บรักษาเว็บถาวร 2) เว็บครอว์เลอร์ใช้ HTTrack ติดต่อเข้าโดยตรงกับ PANDAS และ 3) PANDORA เป็นองค์ประกอบที่สร้างส่วนต่อประสานเพื่อการคั่นเว็บที่จัดเก็บ โดยการสร้างรายการชื่อเรื่องและหัวเรื่อง

3) เครื่องมือสำหรับการเข้าถึงและการค้นหา (Access and finding aids) ตัวอย่างเครื่องมือที่ใช้ในการเข้าถึงและค้นหา เช่น

Wayback (<http://archive-access.sourceforge.net/projects/wayback>) พัฒนา โดย Internet Archive ซึ่งอนุญาตให้ผู้ใช้งานเห็นเวอร์ชันของเว็บเพจที่เก็บเกี่ยวมาในช่วงเวลาต่าง ๆ ได้

NutchWAX (Nutch with Web Archive eXtensions) (<http://archive-access.sourceforge.net/projects/nutchwax>) พัฒนา โดย Internet Archive และ Nordic National Libraries ใช้สำหรับการทำดัชนีและการค้นหาล้างเก็บเว็บถาวร (Web archives) โดยใช้โปรแกรมค้นหา NutchWAX และส่วนขยายเพื่อค้นหาเว็บที่ได้จัดเก็บไว้

SOLR พัฒนาโดย The Apache Lucene project เป็นเครื่องมือค้นหาที่มีแพลตฟอร์มการค้นหาแบบโอเพนซอร์ซ เขียนด้วยภาษาจาวา ใช้ค้นหาข้อความเต็ม (Full text search) และ ค้นหาแบบแฟเซต (Faceted search) จัดทำดัชนีแบบเรียลไทม์ (Real-time indexing) และจัดกลุ่มได้แบบพลวัต (Dynamic clustering) นิยมใช้ในการค้นหาสารสนเทศจากองค์กรต่าง ๆ (Enterprise search) และ กรณีที่ใช้การวิเคราะห์ (Analytics use cases) (Apache Solr https://en.wikipedia.org/wiki/Apache_Solr)

4) เครื่องมือสำหรับการจัดการรูปแบบเว็บถาวร (Web archiving format) ตัวอย่างเครื่องมือประเภทนี้ เช่น

ARC (<http://www.archive.org/web/researcher/ArcFileFormat.php>) พัฒนา โดย Mike Burner และ Brewster Kahle จาก Internet Archive

DAT (http://www.archive.org/web/researcher/dat_file_format.php) บรรจุเมตาดาตาเกี่ยวกับเอกสารที่จัดเก็บในไฟล์ ARC

WARC (<http://bibnum.bnf.fr/WARC/>) พัฒนาโดย ภาควิชาการสงวนรักษาดิจิทัลแห่งชาติ (IIPC) และ ISO TC46/SC4/WG12 จาก ARC ซึ่งเป็นมาตรฐานสากลสำหรับการจัดการรูปแบบไฟล์เว็บถาวร ISO28500:2009, Information and documentation – WARC file format

นอกจากเครื่องมือแล้ว การพิจารณากำหนดมาตรฐานที่ใช้ในการจัดเก็บรักษาเว็บถาวรและการสงวนรักษาก็มีความสำคัญที่ต้องคำนึงถึงเช่นกัน โดยทั่วไปมาตรฐานที่เกี่ยวข้องมีดังนี้ (Pennock, 2013)

เทคโนโลยี HTML และ HTML/XML มาตรฐานภาษาที่ใช้ผลิตเอกสารเว็บเพื่อให้เกิดการทำงานร่วมกันได้ ได้แก่ HTML (Hypertext Markup Language) ซึ่งเป็นภาษาแรก ๆ ของอินเทอร์เน็ต หลังจากนั้นก็มีการพัฒนาภาษาอื่น ๆ อีกมากมาย เช่น XML (Extensible Markup Language) XSLT (eXtensible Language Stylesheet Transformations) เป็นต้น ทั้งนี้มีหน่วยงาน W3C เป็นผู้ดูแลและให้ข้อมูลเกี่ยวกับมาตรฐานภาษาเว็บ

สำหรับการจัดเก็บรักษาเว็บถาวร ภาษา HTML5 ช่วยให้การเก็บเกี่ยววัตถุดิบง่ายขึ้น นอกจากนั้นการใช้เว็บเบราว์เซอร์จำเป็นต้องรู้จักมาตรฐาน CSS (Cascading Style Sheets) JavaScript และ HTTP มาตรฐานเว็บอีกมาตรฐานหนึ่งที่เว็บมาสเตอร์ส่วนใหญ่จะใช้เพื่อแจ้งว่าอนุญาตให้เก็บรวบรวมข้อมูลด้วยเว็บเบราว์เซอร์หรือไม่ คือ โพรโตคอล robots.txt (Pennock, 2013)

ISO 28500:2017 WARC file format (<https://www.iso.org/standard/68004.html>) หรือที่รู้จักกันในชื่อมาตรฐาน WARC ออกแบบโดยสมาชิกรัฐสภาการสงวนรักษาอินเทอร์เน็ตสากล (IIPC) เพื่ออธิบายรายละเอียดสำหรับการบรรจุไฟล์หลายไฟล์ที่รวบรวมจากเว็บไซต์เว็บหนึ่งหรือหลายเว็บไซต์ไว้ด้วยกัน รวมทั้งการดูแลและพรรณนาความสัมพันธ์ระหว่างไฟล์เหล่านั้น ในขณะเดียวกันก็อนุญาตให้เพิ่มเติมเมตาดาตาอื่น ๆ เพื่อวัตถุประสงค์ในการเก็บด้วย มาตรฐาน WARC ได้รับการอนุมัติให้เป็นมาตรฐาน ISO ในเดือนพฤษภาคม ค.ศ. 2009 และปรับปรุงล่าสุด ปี 2017

ISO Technical Report 14873:2013 (<https://www.iso.org/standard/55211.html>) เป็นเอกสารมาตรฐานที่อธิบายเกี่ยวกับสถิติ คำศัพท์ และเกณฑ์คุณภาพสำหรับการจัดเก็บรักษาเว็บถาวร ซึ่งเผยแพร่ใน ค.ศ. 2013 มาตรฐานนี้เหมาะสำหรับองค์กรทางวัฒนธรรม เช่น ห้องสมุด หอจดหมายเหตุ พิพิธภัณฑ์ และ สถาบันวิจัย ใช้ประกอบการพิจารณาจัดทำคลังเก็บเว็บถาวร (Web archives) โดยเฉพาะสถิติการพัฒนาคอลเลกชันและตัวชี้วัดคุณภาพที่เกี่ยวข้องกับความสามารถในการเข้าถึงเนื้อหาอย่างต่อเนื่อง (Pennock, 2013)

ISO 14721:2012 Space data and information transfer systems หรือที่รู้จักกันในชื่อว่า Open Archive Information System (OAIS) Reference Model อธิบายกรอบการทำงานสำหรับการสงวนรักษาเนื้อหาดิจิทัลที่อยู่ในคลัง ถึงแม้ว่าจะไม่ได้เป็นโมเดลเฉพาะเจาะจงสำหรับคลังเก็บรักษาเว็บถาวรก็ตาม แต่ก็ยังให้ภาพขั้นตอนการทำงาน (Workflow) ในการจัดเก็บและการจัดการเนื้อหา

ดิจิทัลที่สลับซับซ้อนได้อย่างดี โดยไม่ได้เน้นเรื่องการแก้ไขปัญหาด้านเทคนิคหรือเครื่องมือ (<https://www.iso.org/standard/57284.html>)

มาตรฐานเมทาตาสำหรับการพรรณนา ใช้พรรณนาเว็บที่เก็บเกี่ยวมาให้ความหมาย และค้นคืนได้ โดยทั่วไปแบ่งเป็น 3 ประเภทตามวัตถุประสงค์ ได้แก่ 1) เมทาตาเชิงพรรณนา (Descriptive metadata) 2) เมทาตาเพื่อการบริหาร (Administrative metadata) และ 3) เมทาตาเพื่อการอนุรักษ์ (Preservation metadata) หรือเมทาตาเกี่ยวกับแหล่งที่มา (Provenance metadata) ในการพรรณนาเอกสารเว็บที่จัดเก็บรักษาดาว ต้องอาศัยการเชื่อมโยงวิธีการลงรายการ บรรณานุกรมแบบที่ใช้ในห้องสมุดและวิธีการลงรายการสำหรับเอกสารจดหมายเหตุ (Dooley & Bowers, 2018) มาตรฐานเมทาตาที่นิยมใช้พรรณนาเว็บที่เก็บเกี่ยว ได้แก่ มาตรฐาน MACHINEReadable Cataloging (MARC21), ISAD(G): General International Standard Archival Description และ Dublin Core Metadata Schema

มาตรฐานการให้หัวเรื่อง มาตรฐานที่ใช้ในการกำหนดหัวเรื่อง เช่น Library of Congress Subject Heading (LCSH) และ Dewey Decimal Classification (DDC) เป็นต้น

นอกจากนี้ยังมีการใช้มาตรฐาน Metadata Encoding and Transmission Standard (METS) และ PREservation Metadata: Implementation Strategies (PREMIS) ร่วมกับ WARC containers เพื่อบันทึกข้อมูลเพิ่มเติมเกี่ยวกับคลังเก็บเว็บถาวร (Pennock, 2013)

บทสรุป

แม้ว่าคลังเก็บเว็บถาวรจะเป็นแหล่งสารสนเทศที่มีความสำคัญต่อการพัฒนาการความก้าวหน้าทั้งทางด้านสังคม เศรษฐกิจ และวิชาการมากขึ้นเรื่อย ๆ การดำเนินโครงการการจัดเก็บรักษาเว็บถาวรก็ยังมีไม่มากนัก โดยเฉพาะในประเทศกำลังพัฒนา ทั้งนี้อาจเป็นเพราะกระบวนการดำเนินงานมีความยุ่งยากซับซ้อน จำเป็นต้องใช้งบประมาณจำนวนมากอย่างต่อเนื่อง และต้องได้รับความร่วมมือจากบุคลากรและองค์กรในภาคส่วนต่าง ๆ หลายฝ่ายที่เกี่ยวข้อง กอปรกับมีโอกาสสูงที่จะต้องเผชิญกับความท้าทายทั้งทางเทคนิค กฎหมาย เศรษฐศาสตร์ และการจัดการ

ความท้าทายทางเทคนิค อาจมีความยุ่งยากในการให้เว็บครอว์เลอร์จัดเก็บรักษาเว็บ 2.0¹ ถาวร ให้ครบถ้วนสมบูรณ์ เนื่องจากเว็บ 2.0 มีแอปพลิเคชัน JavaScript จำนวนมาก เนื้อหา

¹ เว็บ 2.0 เป็นลักษณะของเว็บไซต์ที่ใช้กันในปัจจุบัน

หลากหลายประเภท และมีการเปลี่ยนแปลงตลอดเวลาอย่างรวดเร็ว นอกจากนั้นยังอาจประสบปัญหาความสามารถในการทำงานได้ของเทคโนโลยีอนาคตที่ใช้ในการเข้าถึงและเปิดอ่านเนื้อหาที่จัดเก็บไว้ ซึ่งอาจเป็นได้ทั้งฮาร์ดแวร์และซอฟต์แวร์

ความท้าทายทางกฎหมาย โดยเฉพาะเรื่องลิขสิทธิ์ อาทิ ประเด็นเกี่ยวกับผู้มีสิทธิ์ในสื่อดิจิทัล การทำสำเนาได้ ตลอดจนการเข้าถึงและการเผยแพร่ได้ นอกจากนั้นเนื้อหาที่เผยแพร่ส่วนใหญ่เป็นเนื้อหาที่ผู้ใช้แต่ละคนสร้างขึ้น (User generated content) จึงยิ่งอาจเพิ่มปัญหาเรื่องลิขสิทธิ์และการใช้เวลาในการขออนุญาตจากเจ้าของเนื้อหา (Pennock, 2013)

ความท้าทายทางเศรษฐศาสตร์ เนื่องจากการพัฒนาคลังจัดเก็บรักษาเว็บถาวรเป็นโครงการต่อเนื่องที่ต้องใช้เวลาและทรัพยากรจำนวนมาก ทั้งบุคลากร เทคโนโลยี และงบประมาณ ดังนั้นการดำเนินงานจึงจำเป็นต้องได้รับการสนับสนุนและการจัดสรรงบประมาณอย่างต่อเนื่องเพื่อให้เกิดความยั่งยืน สภาพการณ์เช่นนี้นับเป็นความท้าทายสำคัญที่หน่วยงานรับผิดชอบจะต้องหาแนวทางรับมือและจัดการ (Grotke & Neubert, 2011)

ความท้าทายทางการจัดการ อาทิ การพิจารณาตัดสินใจกำหนดนโยบายคัดเลือกที่เหมาะสมสอดคล้องกับวัตถุประสงค์ ศักยภาพขององค์กร และข้อจำกัดด้านเทคโนโลยีการจัดเก็บรักษาเว็บถาวร และ การแสวงหา ดูแลความร่วมมือระหว่างผู้มีส่วนร่วมทั้งหมดให้ดำเนินไปได้อย่างมั่นคง และจริงจังเพื่อให้การพัฒนาเนื้อหาในคลังเก็บเว็บถาวรดำรงอยู่ได้อย่างยั่งยืน เป็นต้น

ด้วยเหตุนี้ก่อนการดำเนินการ หน่วยงานจึงควรมีการวางแผนและศึกษาประเด็นท้าทายด้านต่าง ๆ ดังกล่าวอย่างรอบด้านรัดกุมเพื่อวางแผนทางการหลีกเลี่ยงความล้มเหลวที่จะเกิดขึ้น

References

- Archive-It user guide. (2018). Retrieved June 21, 2018, from <https://support.archive-it.org/hc/en-us/categories/201179946-Archive-It-User-Guide>
- Ball, A. (2010). **Web archiving (version 1.1)**. Edinburgh, UK. Retrieved from <http://www.dcc.ac.uk/sites/default/files/documents/reports/sarwa-v1.1.pdf>
- Bellardo, L. J. (2001). **Memorandum to chief information officers: Snapshot of public web sites**. Retrieved from <https://www.archives.gov/records-mgmt/basics/snapshot-public-web-sites.html>
- Bragg, M., & Hanna, K. (2013). **The web archiving life cycle model**. Retrieved from <https://archive-it.org/learn-more/>
- Brügger, N. (2005). **Archiving websites: General considerations and strategies. Analysis** (Vol. 1st). Århus: The Centre for Internet Research. <https://doi.org/10.1080/03615260801971129>
- Brügger, N. (2011). **Web archiving – Between past , present , and future**. In M. Consalvo & C. Ess (Eds.), *The Handbook of Internet Studies* (pp. 24–42). Oxford, UK: Wiley-Blackwell. Retrieved from <https://doi.org/10.1002/9781444314861.ch2>
- Brügger, N., & Schroeder, R. (Eds.). (2017). **The web as history**. London, UK: UCL Press. <https://doi.org/10.14324/111.9781911307563>
- Chen, K. H., Chen, Y. L., & Ting, P. F. (2008). **Developing National Taiwan University web archiving system**. In *Proceedings from IWWA '08: 8th International Workshop for Web Archiving* (1–8). Denmark.
- Costa, M., Gomes, D., & Silva, M. J. (2017). **The evolution of web archiving**. *International Journal on Digital Libraries*, 18(3), 191–205. <https://doi.org/10.1007/s00799-016-0171-9>
- Dellavalle, R. P., Hester, E. J., Heilig, L. F., Drake, A. L., Kuntzman, J. W., Graber, M., & Schilling, L. M. (2003). **Going, going, gone: Lost Internet references**. *Science*, 301(July), 174–175.

- Dooley, J. M., & Bowers, K. (2018). **Descriptive metadata for web archiving: Recommendations of the OCLC research library partnership web archiving metadata working group**. Dublin, OH: OCLC Research. <https://doi.org/10.25333/C3005C>
- Dougherty, M., Meyer, E. T., Madsen, C. M., van den Heuvel, C., Thomas, A., & Wyatt, S. (2010). **Researcher engagement with web archives: State of the art**. London, UK. <https://doi.org/10.2139/ssrn.1714997>
- Gomes, D., & Costa, M. (2014). **The importance of web archives for humanities**. *International Journal of Humanities and Arts Computing*, 8(1), 106–123. <https://doi.org/10.3366/ijhac.2014.0122>
- Gomes, D., Miranda, J., & Costa, M. (2011). **A survey on web archiving initiatives**. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 6966 LNCS, 408–420. https://doi.org/10.1007/978-3-642-24469-8_41
- Grotke, A., & Jones, G. (2010). **DigiBoard: A tool to streamline complex web archiving activities at the Library of Congress**. In *Proceedings from IAWW '10: 10th International Web Archiving Workshop*, Vienna. Vienna.
- Grotke, A., & Neubert, M. (2011). **Building the “ North Africa & the Middle East 2011 ” web archive an international collaboration**. Retrieved 2018 July 17. URL: <https://www.mela.us/wp-content/uploads/2016/03/Building-the-North-AfricaGrotke-Neubert.pdf>
- Lasfargues, F., Oury, C., & Wendland, B. (2008). **Legal deposit of the French web: Harvesting strategies for a national domain**. In *Proceedings from IAWW '08: 8th International Workshop for Web Archiving*. Aarhus, Denmark. Retrieved URL: <https://hal-bnf.archives-ouvertes.fr/hal-01098538/document>
- Lecher, H. E. (2006). **Small scale academic web archiving: DACHS**. In *Web Archiving* (pp. 213–226). Berlin, Germany: Springer-Verlag.

Library of Congress. (2018). **Collections with archived web sites**. Retrieved 2018 April 1. URL:
<https://www.loc.gov/websites/collections>

List of web archiving initiatives. (n.d.). Retrieved 2018 April 1. URL:
https://en.wikipedia.org/wiki/List_of_Web_archiving_initiatives

Masanès, J. (2005). **Web archiving methods and approaches: A comparative study**. Library Trends, 54(1), 72–90. <https://doi.org/10.1353/lib.2006.0005>

Masanès, J. (2006). **Web archiving**. (J. Masanes, Ed.). Paris, France: Springer. Retrieved from
<https://link.springer.com/content/pdf/10.1007%2F978-3-540-46332-0.pdf>

National Diet Library. (n.d.). **Let's WARP to the past of the Web!** Retrieved 2018 December 15. URL: http://warp.da.ndl.go.jp/info/WARP_en.html

National Library of Australia. (2018). **About PANDORA**. Retrieved 2018 April 1. URL:
<https://pandora.nla.gov.au/about.html>

National Library of New Zealand. (2018). **New Zealand Web Archive**. Retrieved 2018 December 14. URL: <https://natlib.govt.nz/collections/a-z/new-zealand-web-archive>

National Library of the Czech Republic. (2018). **Webarchiv preserves Czech web**. Retrieved 2018 December 15. URL: <https://www.webarchiv.cz/en/about>

Niu, J. (2012). **An overview of web archiving**. D-Lib Magazine, 18(3/4).
<https://doi.org/10.1045/march2012-niu1>

Ntoulas, A., Cho, J., & Olston, C. (2004). **What's new on the web?: The evolution of the web from a search engine perspective**. Proceedings of the 13th Conference on World Wide Web, 1–12. <https://doi.org/10.1145/988672.988674>

Pearce-Moses, R., Kaczmarek, J., & Pearce-Moses, R. Kaczmarek, J. (2005). **An Arizona model for preservation and access of web documents**. DttP: Documents to the People, 33(1), 17–24. Retrieved URL: <http://www.lib.az.us/diggovt/documents/pdf/AzModel.pdf>

Pennock, M. (2013). **Web-archiving**. In DPC Technology Watch Reports. London, UK.
<http://dx.doi.org/10.7207/twr13-01>

Spencer, A., O'Reilly, B., & Vasile, G. (2009). **Past and present: Using the UK Government web archive to bridge the continuity gap**. In Proceedings from IAWAW '09: 9th International Web Archiving Workshop, Corfu. (38–43).

The British Library. (n.d.). **What is the UK Web Archive?** Retrieved 2018 April 1. URL:
<https://www.webarchive.org.uk/ukwa/info/about>

The National Archives. (2011). **Web archiving guidance**. London, UK. Retrieved URL:
<http://www.nationalarchives.gov.uk/documents/information-management/web-archiving-guidance.pdf>