VẤN ĐỀ LƯU TRỮ WEB TẠI MỘT SỐ THƯ VIỆN NƯỚC NGOÀI

MỘT SỐ TIÊU CHUẨN VÀ PHẦN MỀM TRONG QUY TRÌNH TẠO LƯU TRỮ WEB

Hiệp hội Lưu trữ Internet quốc tế (International Internet Preservation Consortium - IIPC) định nghĩa lưu trữ web là tập hợp các quy trình thu thập các thông tin, các trang web từ World Wide Web, lưu giữ chúng và sau đó cung cấp quyền truy cập đến cho người sử dụng. Việc quy định ở phạm vi toàn cầu đối với quá trình lưu trữ web hiện đang sử dụng tiêu chuẩn quốc tế ISO 14721:2012 (Hệ thống truyền tải thông tin và dữ liệu không gian - Hệ thống thông tin lưu trữ mở (OAIS) - Mô hình tham chiếu). Tiêu chuẩn này được xem xét và cập nhật thường xuyên. Theo báo cáo kỹ thuật ISO/TR 14873:2013, lưu trữ web đề cập đến các hoạt động lựa chọn, thu thập, lưu trữ và quản lý quyền truy cập vào các tệp dữ liệu của nguồn tài nguyên từ internet trong một khoảng thời gian nhất định. Hiện nay, ngoài Tiêu chuẩn quốc tế ISO 14721:2012, còn có một số tiêu chuẩn đang được nghiên cứu và sử dụng (Bảng 1) [4].

Bảng 1. *Một số Tiêu chuẩn quốc tế về lưu trữ web*

Tiêu chuẩn	Mô tả	Địa chỉ
ISO 14721:2012 Mô hình tham chiếu cho hệ thống thông tin lưu trữ mở (OAIS)	Tiêu chuẩn quy định toàn bộ các vấn đề cơ bản trong việc lưu trữ Dữ liệu lớn; các tiêu chuẩn siêu dữ liệu mở đang được sử dụng phổ biến nhất như: Dublin Core, EAD, METS, từ điển siêu dữ liệu để lưu trữ lâu dài các tài liệu điện tử	https://public.ccsds.org/pubs/650x0m2.pdf
ISO 16363:2012 Kiểm tra và chứng nhận các kho lưu trữ kỹ thuật số đáng tin cậy (TDS)	Thiết lập các số liệu toàn diện của một kho lưu trữ theo đúng tiêu chuẩn dựa trên OAIS	https://public.ccsds.org/pubs/652x0m1.pdf
ISO 28500:2017 Thông tin và tài liệu - định dạng file WARC	Tiêu chuẩn này xem xét các thành phần có ứng dụng định dạng WARC nhằm tránh trùng lặp giữa các file trong kho lưu trữ điện tử, xem xét khả năng lưu siêu dữ liệu, giao thức nén tệp, v.v.	https://www.iso.org/ standard/68004.html
ISO/TR 14873:2013 Thông tin và tài liệu - Số liệu thống kê và những vấn đề về chất lượng đối với việc lưu trữ web	Xác định số liệu thống kê, điều kiện và tiêu chí chất lượng để lưu trữ web. Tiêu chuẩn này đề cập đến nhu cầu và thực tiễn của các cơ quan, tổ chức như: thư viện, cơ quan lưu trữ, bảo tàng, trung tâm nghiên cứu, kho lưu trữ di sản.	https://www.iso.org/standard/55211.html
ISO 16919:2014 Hệ thống chuyển giao thông tin và dữ liệu không gian - Những yêu cầu đối với các cơ quan thực hiện việc kiểm tra và chứng nhận về chất lượng của các kho lưu trữ kỹ thuật số.	Chủ yếu dành cho những người thành lập và quản lý cơ quan/tổ chức thực hiện việc đánh giá và chứng nhận kho lưu trữ kỹ thuật số. Tiêu chuẩn này cũng giúp cho các cơ quan/tổ chức có thể tự đo lường một cách khách quan độ tin cậy đối với kho lưu trữ của mình.	https://public.ccsds.org/Pubs/652x1m2.pdf

Để tiến hành thu thập thông tin lưu trữ trên internet, thông thường, các thư viện, tổ chức sẽ sử dụng phần mềm nguồn mở. Phần mềm phổ biến nhất được sử dụng trong các thư viện quốc gia và các tổ chức lưu trữ web là phần mềm nguồn mở Heritrix. Bên cạnh đó, một số phần mềm khác cũng đang được các cơ quan/tổ chức sử dụng trong quá trình lưu trữ web (Bảng 2) [4].

Bảng 2. *Một số phần mềm được sử dụng trong hoạt động lưu trữ web*

Tên phần mềm	Nhóm phát triển	Mô tả	Định dạng	Nguồn
Heritrix	Cơ quan Lưu trữ Internet (Internet Archive), Hoa Kỳ	Heritrix - phần mềm cung cấp miễn phí, là một trình thu thập dữ liệu web được thiết kế dành để lưu trữ web.	WARC	https://github.com/internetarchive/heritrix3
Open Wayback	Lauren Ko, Trường Đại học Bắc Texas, Hoa Kỳ	Open Wayback là phần mềm được các cơ quan lưu trữ web trên toàn thế giới sử dụng để "tái tạo" các trang web được lưu trữ ở trình duyệt của người dùng	ARC hoặc WARC	https://netpreserve.org
Apache Solr	Apache Software Foundation	Apache Soir là một nền tảng tìm kiếm toàn văn với mã nguồn mở, dựa trên dự án Apache Lucene. Các tính năng chính của phần mềm là: tìm kiếm toàn văn, đánh dấu kết quả, tìm kiếm theo nhiều khía cạnh, tích hợp CSDL, xử lý tài liệu có định dạng phức tạp.	-	https://lucene.apache.org/solr/
Web Curator Tool (WCT)	Thư viện Quốc gia New Zealand, Thư viện Anh	WCT là một ứng dụng dùng để quản lý quy trình làm việc với nguồn mở để lưu trữ web có chọn lọc. Ứng dụng tích hợp với trình thu thập thông tin Heritrix và hỗ trợ các quy trình chính như: thu thập, kiểm soát chất lượng và thu thập siêu dữ liệu mô tả.	WARC	http://webcuratortoo..org/
HTTrack	Xavier Roche và cộng sự	HTTrack là trình duyệt ngoại tuyến đa nền tảng trong truy cập mở, cho phép tải các trang web từ internet đến máy tính nội bộ.	HTML	http://www.httrack.com/

MỘT SỐ HOẠT ĐỘNG LƯU TRỮ WEB TRÊN THẾ GIỚI HIỆN NAY

Ở nhiều quốc gia trên thế giới, việc lưu trữ web với tên miền quốc gia được thực hiện theo các tiêu chuẩn quốc tế thống nhất với phần mềm Heritrix mở ở định dạng WARC. Tùy vào mục tiêu, nhiệm vụ của mỗi cơ quan, tổ chức, quy trình lưu trữ sẽ được thực hiện ở các mức độ khác nhau.

Hằng năm, Hiệp hội Lưu trữ Internet quốc tế tổ chức Hội nghị Lưu trữ web (Web Archiving Conference - WAC), trong đó có các bài thuyết trình về sự phát triển của công nghệ lưu trữ web nói chung và lưu trữ web có chọn lọc, cũng như các vấn đề về luật bản quyền quốc gia trong môi trường kỹ thuật số. Một trong số các vấn đề chính mà các thư viện quốc gia ở nhiều nước phải đối mặt khi hình thành kho lưu trữ web, đó là khó khăn trong việc phối hợp sao chép và bảo quản dữ liệu với chủ sở hữu bản quyền, vì các tổ chức tư nhân không muốn đáp ứng yêu cầu sao chép tài nguyên internet từ các thư viện [2].

Internet Archive - một tổ chức phi lợi nhuận được thành lập vào năm 1996, là một trong những tổ chức nổi tiếng nhất trong việc lưu trữ nội dung web trên internet. Internet Archive được xem như là một thư viện kỹ thuật số trực tuyến lớn, cho phép tất cả mọi người truy cập miễn phí nội dung tài liệu số, bao gồm các bản lưu của trang web, sách, tài liệu đồ họa, video, bản ghi âm và các phần mềm ứng

dụng. Các tổ chức lưu trữ web coi hoạt động này là sự mở rộng sứ mệnh bảo tồn di sản văn hóa của dân tộc.

Tại các thư viện quốc gia, việc phát triển trong lĩnh vực lưu trữ web đã được tiến hành từ năm 1994, bắt đầu với Dự án thí điểm xuất bản điện tử (Electronic Publication Pilot Project - EPPP) của Thư viện Quốc gia Canada. Từ năm 1996, tài nguyên web đã được công nhận là lưu trữ hợp pháp ở nhiều quốc gia. Hoạt động này bắt đầu phát triển đặc biệt tích cực trong những thập kỷ gần đây (từ năm 2000 đến 2020) và ngày càng mở rộng. Việc lưu trữ hợp pháp có thể bao gồm: các trang web, các trang riêng lẻ và các đối tượng thông tin trên các trang web (hình ảnh, văn bản điện tử, tài liệu đa phương tiện và các loại nội dung khác).

Tùy thuộc vào quy mô và mục đích thu thập, người ta phân biệt hai phương pháp lưu trữ web chính, đó là: thu thập và lưu trữ tự động liên tục có hệ thống và lưu trữ có chọn lọc. Việc khai thác tên miền quốc gia nhằm mục đích lấy bản sao cấu trúc nội dung của toàn bộ hoặc một phần tên miền. Cách tiếp cận này là điển hình cho các thư viện quốc gia của các nước: Ireland Croatia, Phần Lan, Úc, Canada, Bỉ, Đức, Israel, Trung Quốc và một số quốc gia khác. Việc thu thập có chọn lọc được thực hiện ở quy mô nhỏ và có mục tiêu rõ ràng hơn. Việc lựa chọn được thực hiện trên cơ sở các tiêu chí nhất định và có trọng tâm tùy thuộc vào tính chất và chức năng, nhiệm vụ của các thư viện. Việc thu thập có chọn lọc đòi hỏi cần xem xét và tuân thủ các tiêu chuẩn đã định trước [1].

New Zealand là một trong những quốc gia đi đầu trong việc mở rộng hoạt động lưu trữ cho tất cả các tài liệu số và coi đó là hoạt động hoàn toàn hợp pháp, bao gồm các tài nguyên internet truy cập mở. Đạo luật được thông qua vào năm 2003 đã trao quyền thu thập và bảo quản tất cả các ấn phẩm điện tử và tài nguyên internet trong nước cho Thư viện Quốc gia New Zealand. Điều này tạo điều kiện thuận lợi cho việc thành lập Kho Lưu trữ Di sản kỹ thuật số quốc gia (National Digital Heritage Archive - NDHA) nhằm cung cấp việc lưu giữ vĩnh viễn thông tin kỹ thuật số của New Zealand. Chiến lược quốc gia làm nền tảng cho NDHA không phân biệt giữa nội dung do các tổ chức có ủy quyền tạo ra hay nội dung do công dân tạo ra. Các tài liệu số lưu trữ trong NDHA được chuyển đến Thư viện Quốc gia từ bốn nguồn: các nguồn lưu trữ hợp pháp, các tìm kiếm trên internet, tài liệu được quyên góp và các nguồn tài liệu số hóa. Các sáng kiến lưu trữ mang tính pháp lý tương tự cũng được triển khai ở một số nước như: Anh, Đức, Na Uy, Úc, Litva, Estonia, Nhật Bản và nhiều quốc gia khác.

Những quy định mở về pháp lý trong việc lưu trữ nguồn tài liệu điện tử đã cho phép Thư viện Quốc gia Anh có thể thực hiện việc thu thập di sản điện tử của toàn xã hội. Phối hợp với năm thư viện lớn nhất của Vương quốc Anh, Thư viện Quốc gia Anh thu thập các trang web tên miền quốc gia và là thành viên chính thức của Hiệp hội Lưu trữ internet thế giới và tham gia các dự án lưu trữ web quốc tế.

Một kinh nghiệm đáng lưu ý khác là của Thư viện Quốc gia Úc. Hiện nay, Thư viện đang thu thập và bảo tồn vào kho lưu trữ web của Úc để truy cập lâu dài thông qua cổng thông tin quốc gia Trove. Bộ sưu tập của kho lưu trữ được xây dựng dựa trên nền tảng thu thập có chọn lọc. Thu thập các trang web của Chính phủ Úc và thu thập toàn bộ tên miền "au" (tên miền chính thức mới của Úc).

Theo một nghiên cứu của tổ chức Internet Archive, ngày càng có nhiều thư viện quốc gia muốn hướng đến việc tự xử lý việc lưu trữ web. Việc xây dựng các kho lưu trữ web đang được đưa vào hoạt động tạo lập bộ sưu tập của nhiều thư viện nhằm hỗ trợ cho mục đích học tập, nghiên cứu. Một số thư viện quốc gia đang hợp tác với Internet Archive để phát triển các bộ sưu tập tài liệu web và đăng ký sử dụng phần mềm của tổ chức này. Trong số đó có Thư viện Quốc gia Ireland - thư viện đã tạo lập được các bộ sưu tập web chọn lọc, có thể tìm thấy trên "Archive-it.org".

Tại Liên bang Nga, năm 2017, Thư viện Tổng thống đã đặc biệt chú ý đến việc lưu trữ các tài nguyên web liên quan đến lịch sử và sự phát triển của nước Nga. Ngày nay, theo định kỳ, Thư viện Tổng thống lưu trữ các tài nguyên như: trang web chính thức của Tổng thống Nga "kremlin.ru" và trang web chính thức của Chính phủ Nga "Government.ru". Dự án thu thập dữ liệu tên miền quốc gia vào một kho lưu trữ điện tử duy nhất đang được thực hiện bởi một tổ chức phi lợi nhuận "Văn hóa thông tin" (Информационная культура) dựa trên những tiêu chuẩn quốc tế và sử dụng phần mềm Heritrix mở ở định dạng WARC. Kho lưu trữ kỹ thuật số quốc gia của Nga được lưu trữ trên miền ".org", trong đó có cơ sở dữ liệu dành riêng cho các tài khoản trên mạng xã hội liên quan đến hoạt động của các cơ quan thuộc chính phủ và các thành viên đại diện của các cơ quan đó [4].

MỘT SỐ VẤN ĐỀ KHI THU THẬP TÀI LIỆU TRÊN WEB

Khi thực hiện các dự án lưu trữ web, các thư viện quốc gia phải đối mặt với thực tế là việc thu thập liên tục các tên miền quốc gia đã tạo ra một khối lượng thông tin khổng lồ, do đó, rất khó trong việc xử lý, kiểm soát và quản lý. Để bảo đảm an toàn cho một khối lượng thông tin lớn như vậy, đòi hỏi cần phải có sự đầu tư tài chính khá lớn.

Vấn đề cấp bách nhất mà các thư viện nước ngoài gặp phải khi hình thành kho lưu trữ web đó là việc tích hợp các quy trình lưu trữ web vào một môi trường chung để tạo lập các bộ sưu tập. Theo Hiệp hội Lưu trữ Internet, một trong những khó khăn chính của quá trình lưu trữ web là vấn đề lưu trữ một lượng lớn thông tin và thông tin này cũng có thể bị biến mất theo thời gian khỏi địa chỉ mạng hiện tại hoặc phải chuyển sang một mức độ web cao hơn. Do đó, việc phát triển các phần mềm mới ưu việt hơn để tìm kiếm, xử lý và hình thành các kho lưu trữ web vẫn đang được tiếp tục phát triển.

Bên cạnh đó là các vấn đề về bản quyền khi sao chép tài liệu, cũng như các vấn đề liên quan đến kỹ thuật khi chỉnh sửa các đối tượng khó sao chép. Một giải pháp toàn diện được đưa ra là các thư viện chuyển sang các dịch vụ lưu trữ web toàn cầu (Internet Archive) hoặc thu hẹp phạm vi hình thành các kho lưu trữ web thành các bộ sưu tập chuyên đề nhỏ.

Việc tạo ra các bộ sưu tập theo chủ đề và các kho lưu trữ vi mô cũng cần phải tuân thủ các chuẩn mực và tiêu chuẩn quốc tế về sao chép tài liệu kỹ thuật số và cung cấp quyền truy cập dựa theo các điều khoản đã thỏa thuận với người giữ bản quyền. An toàn nhất trên quan điểm pháp lý, đó là tải xuống các tài liệu đã được cung cấp theo giấy phép Creative Commons CC BY, CC BY-SA và CC0.

Tùy thuộc vào tầm quan trọng của chiến lược, nhiệm vụ, cũng như khả năng về cơ sở vật chất, kỹ thuật, pháp lý, các thư viện có thể thực hiện nhiều cách tiếp cận khác nhau để lưu trữ nguồn tài nguyên trên internet: từ việc ghi lại các trang web riêng lẻ đến hoạt động sao chép toàn bộ tên miền cấp cao nhất. Mục đích chính của việc lưu trữ web là lưu giữ vĩnh viễn các bản ghi từ internet để phục vụ cho các mục đích khoa học, nghiên cứu, giáo dục, nghề nghiệp và các mục đích khác.

Nguyễn Thị Tú Quyên

TÀI LIỆU THAM KHẢO

1. About IIPC. Truy cập tại: https://netpreserve.org/about-us/ (Tháng 9/2023).

2. International Internet Preservation Consortium. URL: https://netpreserve.org/

3. ISO/TR 14873:2013. Information and documentation - Statistics and quality issues for web archiving. URL: https://www.iso.org/standard/55211.html

4. Научные и технические библиотеки. Проблемы отечественного и зарубежного веб-архивирования в библиотеках. Веб-архивирование как область деятельности. 2022, 12, 104.