

TÍNH MỞ, QUỐC TẾ VÀ TOÀN DIỆN CỦA CÁC CÔNG CỤ LƯU TRỮ PHÙ HỢP VỚI WEB

Gordon Mohr

Cơ quan lưu trữ Internet, California, Hoa Kỳ

Giới thiệu bộ công cụ lưu trữ trên web của Cơ quan lưu trữ Internet, bao gồm: máy dò tìm lưu trữ Heritrix, kho mẫu tệp lưu trữ WARC, dịch vụ duyệt lưu trữ Wayback và các tiện ích định chỉ mục và tìm kiếm toàn văn NutchWAX.

Năm 2003, Cơ quan lưu trữ Internet (the Internet Archive- IA) bắt đầu phát triển các công cụ mã nguồn mở (open source tools) cho việc lưu trữ trên web với sự hỗ trợ và giúp đỡ của nhiều thư viện quốc gia.

Cơ quan lưu trữ Internet là một thư viện Internet phi lợi nhuận đặt trụ sở ở San Francisco, bang California, Hoa Kỳ. Cơ quan được biết đến nhờ “Máy Wayback”. Máy này cung cấp truy cập công cộng tới một kho lưu trữ trên trang web có hơn 100 tỷ URL (chỉ số định vị nguồn đồng nhất) kể từ năm 1996. IA cũng đóng vai trò chủ đạo trong nỗ lực số hóa mảng tài nguyên sách thuộc Tổ chức liên kết nội dung mở (the Open Content Alliance); và là chủ của các bộ sưu tập nội dung nghe nhìn miễn phí phổ cập, gồm hàng ngàn buổi biểu diễn nhạc sống và các cuộc trình diễn giáo dục.

Mảng tài liệu rời trong sưu tập Web của IA đến từ các nguồn tặng có nội dung chưa xử lý của Alexa Internet- một đối tác thương mại. Tuy nhiên, vào năm 2003, Cơ quan lưu trữ Internet, cùng với các thư viện quốc gia muốn tiếp tục hình thành một Liên hiệp bảo quản quốc tế trên Internet (the International Internet Preservation Consortium- IIPC) nhận thấy có nhu cầu về những công cụ và tiêu chuẩn mới, được xây dựng theo mô hình hợp tác mở, cho việc lưu trữ trên web. Những công cụ đầu tiên còn thiếu tính linh hoạt, tâm điểm lưu trữ, và khả năng cấp giấy phép không gây phiền hà với phương pháp tiếp cận mã nguồn mở.

Việc phát triển các công cụ này bắt đầu với máy dò tìm *Heritrix* để thu thập các nội dung trên web. Chúng cũng đã phát triển để tiếp nhận một kho mẫu chuẩn *WARC* nhằm lưu trữ và trao đổi các nội dung trên web; một dịch vụ duyệt *Wayback* để xem các nội dung lưu trữ; và các tiện ích tìm kiếm, *NutchWAX* để

định chỉ mục và hỏi đáp toàn văn nội dung lưu trữ chỉ dùng phần mềm miễn phí. Tất cả các công cụ này hiện sẵn có để tải xuống và sử dụng miễn phí với đầy đủ mã nguồn. Thoạt đầu, phần mềm được ứng dụng trên nền chéo Java, và sẵn sàng kết nối và tùy biến cho các dự án khác.

Máy dò tìm lưu trữ Heritrix. Heritrix được thiết kế để lưu trữ nội dung một cách trung thành và đầy đủ, với mức độ cấu hình và tùy biến cao. Ở Cơ quan lưu trữ Internet và các nơi khác, Heritrix đã được sử dụng để dò tìm với các tần suất khác nhau: hàng ngày, hàng tuần, hàng tháng hoặc một lần- và phạm vi - từ một vài ngàn tới hàng tỷ URL nắm bắt được. Heritrix có thể được điều khiển từ xa bởi phần mềm khác hoặc được mở rộng đáng kể với những mô đun mã mới.

Khổ mẫu tập lưu trữ WARC. Trong hơn một thập kỷ, Cơ quan lưu trữ Internet đã lưu trữ các nội dung web nắm bắt được theo dạng thức các câu trả lời được móc nối với nhau đơn giản, gọi là ARC. Để xử lý tốt hơn các siêu dữ liệu về sưu tập-thời gian, sự nhân rộng- sự thu hẹp, sự tiến hóa của dạng thức và các nhu cầu lưu trữ liên quan khác, Cơ quan lưu trữ và các thư viện khác đã thiết kế một khổ mẫu kế vị gọi là WARC, hiện đang được coi như một tiêu chuẩn quốc tế ISO.

Dịch vụ duyệt lưu trữ Wayback. Phần mềm Wayback cho phép xem và duyệt các nội dung web trên cơ sở URL theo một trình duyệt, như thể xem một trang web gốc gần như vào lúc mong muốn. Wayback có thể tạo điều kiện truy nhập tới các bộ sưu tập hàng tỷ URL.

Các tiện ích định chỉ mục và tìm kiếm toàn văn NutchWAX. NutchWAX bổ sung thêm một cách tìm kiếm Google vào các kho lưu trữ tài nguyên web trên cơ sở các dự án mã nguồn mở khác. Các dự án này bao gồm cả Lucene, một máy tìm toàn văn thô sơ; Nutch, một phương án thích nghi của Lucene với nội dung Web; và Hadoop, một hệ thống song hành hóa các công việc xử lý lớn. NutchWAX hiện đang được sử dụng trong các cấu hình phân tán ở Cơ quan lưu trữ Internet để tạo khả năng tìm kiếm trong các bộ sưu tập có chứa hơn một tỷ URL nắm bắt được.

Cơ quan lưu trữ Internet chủ động duy trì từng dự án nghiên cứu phần mềm này, các cải tiến thường được các đối tác thư viện của IA và người sử dụng phần mềm khác tài trợ và đóng góp. Tuy nhiên, phần mềm lưu trữ web tiếp tục phải đối

mặt với những thách thức nghiêm trọng như web ngày một phát triển về quy mô, về tính đa dạng của các loại nội dung và công nghệ, và về mức độ thao túng xuyên tạc nội dung (tài liệu rác trên web).

Phương pháp tiếp cận mang tính hợp tác, mã nguồn mở và tính quốc tế đã được áp dụng để xây dựng một cách hiệu quả một cơ sở năng lực lưu trữ web cùng chia sẻ, và IA đang tìm những quan hệ hợp tác mới để tiếp tục bước tiến này, xây dựng những chức năng mới.

Vũ Văn Sơn dịch

Tài liệu gốc: “Asian Digital Libraries: Looking Back 10 Years and Forging New Frontiers”, pp. 7-8