

MỘT SỐ PHƯƠNG PHÁP LỰA CHỌN VÀ LƯU TRỮ DỮ LIỆU HIỆN NAY

MỞ ĐẦU

Kho dữ liệu nghiên cứu hiện nay là một trong những thành phần chủ chốt của hệ sinh thái khoa học mở, góp phần hình thành và phát triển các hoạt động nghiên cứu và đổi mới sáng tạo. Khi lượng dữ liệu nghiên cứu tăng lên, cùng với sự phát triển nhanh chóng của các công nghệ thì mô hình cung cấp dữ liệu cũng thay đổi. Kể từ khi các kho lưu trữ thể chế bắt đầu hình thành, việc áp dụng các sáng kiến có tính pháp lý liên quan đến khoa học mở đã được đặc biệt chú trọng. Số lượng kho lưu trữ trên thế giới đang tăng lên hàng năm. Việc tổ chức các cộng đồng khác nhau liên quan đến việc phát triển và vận hành kho lưu trữ đã góp phần vào sự phát triển bền vững của chúng và thay đổi mô hình truyền thông khoa học trong môi trường nghiên cứu.

1. LỰA CHỌN LƯU TRỮ DỮ LIỆU

Tri thức nói chung và thông tin khoa học nói riêng là một trong những nguồn tài nguyên có giá trị nhất đối với nhân loại. Những khám phá khoa học có giá trị hoặc dữ liệu thu được trong quá trình nghiên cứu đều có thể tác động và làm thay đổi nhận thức của chúng ta. Khối lượng dữ liệu không ngừng tăng lên mỗi ngày và phần lớn các thành tựu khoa học cũng được khám phá, phát triển dựa trên các nghiên cứu được công bố trong các tài liệu khoa học trước đó.

Thư viện là nơi thu thập, tổ chức sắp xếp, bảo quản và cung cấp quyền khai thác, truy cập vào nguồn kiến thức và thông tin. Bên cạnh việc cung cấp thông tin, tri thức cho người dùng, thư viện cũng thực hiện một nhiệm vụ quan trọng là lưu giữ và bảo quản nguồn tri thức và văn hóa của nhân loại cho các thế hệ sau. Các thư viện ngày nay lưu giữ và bảo quản các bộ sưu tập không chỉ bao gồm các tài liệu in như sách, tạp chí, báo, mà còn cả các bản ghi âm thanh, bản sao nghệ thuật, vi phim, video, bản đồ, ảnh, đĩa CD-ROM, phần mềm máy tính, cơ sở dữ liệu trực tuyến và các phương tiện truyền thông khác. Ngoài việc duy trì các bộ sưu tập trong các tòa nhà của mình, các thư viện

hiện đại hướng tới việc chuyển sang thư viện số và cung cấp cho người dùng quyền truy cập thông tin từ xa. Cho dù nguồn thông tin, tri thức, văn hóa được lưu giữ ở dạng giấy hay ở định dạng điện tử thì các thư viện đều phải bảo đảm để các bộ sưu tập hồ sơ được bảo quản tốt và có thể sử dụng sau này.

Ở hầu hết các thư viện, thông tin được trình bày trên giấy, gồm sách, tạp chí, báo, bản thảo và các ấn phẩm in khác vẫn chiếm số lượng khá cao. Chúng ta có sự tin tưởng với việc lưu trữ thông tin trên giấy, tuy nhiên, các nhà nghiên cứu cũng lưu ý một số khía cạnh quan trọng khi lưu trữ ở dạng này. Thứ nhất, chất lượng giấy có thể thay đổi tùy thuộc vào thời điểm ra đời, xuất bản: giấy từ thế kỷ 19 và thế kỷ 21 là hai loại giấy khác nhau. Thứ hai, ngoài sự khác nhau về chất lượng và sự thay đổi của các vật liệu để trình bày thông tin trên giấy thì những nhược điểm đối với các kho lưu trữ truyền thống như khả năng hao mòn vật lý, mối đe dọa từ độ ẩm và hỏa hoạn cũng là những vấn đề đáng lo ngại.

Sự phức tạp của công tác bảo quản là khối lượng tài liệu ngày càng tăng và việc phải giám sát hằng ngày. Vấn đề này được giải quyết một phần bằng số hóa. Vì vậy, phương pháp thứ hai để lưu trữ bộ sưu tập thư viện là chuyển đổi tài liệu sang định dạng kỹ thuật số, cho phép lưu giữ và khai thác tài nguyên thông tin ở dạng nhỏ gọn hơn, thuận tiện hơn cho việc tìm kiếm tự động trong điều kiện công nghệ hiện nay. Cùng với đó, việc lựa chọn các tài liệu để ưu tiên chuyển sang định dạng kỹ thuật số cũng trở thành một nhiệm vụ quan trọng và cấp bách.

Một số nhiệm vụ và vấn đề phổ biến mà các thư viện phải đối mặt ở giai đoạn đầu khi triển khai việc tổ chức bảo quản dữ liệu chính là chiến lược thu thập, đánh giá, lựa chọn tài liệu và các tập dữ liệu để bảo quản lâu dài. Do đó, việc chia sẻ những kinh nghiệm, các phương pháp và cách thức chính để lưu trữ, lựa chọn và xử lý thông tin khoa học cũng đang là một trong những vấn đề quan trọng được thảo luận trong các nghiên cứu.

2. TIÊU CHÍ ĐÁNH GIÁ ĐỂ LỰA CHỌN BẢO QUẢN DỮ LIỆU

Bảo quản dữ liệu, bao gồm cả nội dung số, vẫn là một lĩnh vực tương đối mới. Các phương pháp đánh giá các tập dữ liệu để bảo quản lâu dài là một vấn đề quan trọng. Với số lượng thông tin lớn và nguồn kinh phí ngày càng hạn hẹp, chúng ta không thể lưu giữ tất cả. Do vậy, các nhà quản lý kho lưu trữ và dữ liệu phải xây dựng phương án và đưa ra quyết định những gì sẽ được lưu giữ để quản lý và bảo quản lâu dài.

Các giai đoạn chính trong quá trình chuẩn bị lưu trữ dữ liệu ở bất kỳ định dạng nào cũng phải thực hiện việc lọc dữ liệu (loại bỏ nhiễu, trùng lặp, lỗi thiếu sót,...); chia tỷ lệ dữ liệu để tạo điều kiện so sánh; chuyển đổi và thích ứng với một công cụ phân tích cụ thể; tổng hợp và tích hợp dữ liệu, kết hợp thông tin từ các nguồn khác nhau và chuyển đổi thành một hệ thống duy nhất.

Tiêu chuẩn quản lý hồ sơ của Úc đề xuất bảy tiêu chí chung khi lựa chọn tài liệu, bao gồm: phục vụ mục tiêu chung, giá trị khoa học hoặc lịch sử, tính duy nhất, tiềm năng sử dụng lại, không thể tái tạo, tính kinh tế và tính đầy đủ của tài liệu. Các khái niệm chính mà Dịch vụ Vương quốc Anh xem xét khi lựa chọn dữ liệu là: tuân thủ các tiêu chuẩn quốc gia, có giá trị khoa học hoặc lịch sử, nguồn hoặc loại dữ liệu mới, giá trị quốc tế, tính duy nhất/rủi ro mất mát, khả năng sử dụng/phân phối lại, lợi ích hoạt động, sao chép.

Trung tâm quản lý dữ liệu số (Digital Curation Centre- DDC) đưa ra đề xuất trong hướng dẫn lựa chọn dữ liệu, bao gồm: tiềm năng tái sử dụng, tuân thủ luật pháp hoặc chính sách, giá trị lâu dài, tính đầy đủ của dữ liệu. Các nhà khoa học từ Hoa Kỳ, Tây Ban Nha và Hà Lan ở giai đoạn đầu xác định bốn thành phần để chọn siêu dữ liệu, đó là: khả năng tìm thấy, khả năng truy cập, tính tương tác và tài sử dụng. Các tiêu chí chung mà các nhà nghiên cứu đề cập đến chính là tính toàn vẹn, khả năng tái tạo, khả năng khai thác, chia sẻ. Các vấn đề liên quan đến bảo quản dữ liệu số thường được đề cập tới là về tài chính, nhân sự, sở hữu trí tuệ, việc tuân thủ các tiêu chuẩn, chiến lược lưu trữ,

chứng nhận, đánh giá rủi ro,... Ở một số nước (như: Anh, Hoa Kỳ, Nga, Trung Quốc), các chính sách bảo quản dữ liệu số của thư viện thường được xây dựng và phát triển với sự trợ giúp của dịch vụ quản lý số hoặc họ sử dụng dịch vụ riêng của mình để thực hiện hoạt động này (như: Hệ thống thư viện của Viện Công nghệ Massachusetts, Quỹ Khoa học quốc gia tại Vương quốc Anh, Hiệp hội các thư viện nghiên cứu châu Âu LIBER tại Hà Lan).

Đánh giá và lựa chọn là một công việc quan trọng và có lẽ cũng là công việc khó khăn nhất, không thể mang tính chủ quan. Công việc này cần phải được hướng dẫn thông qua các văn bản, chính sách của các đơn vị, cơ quan quản lý có liên quan trong lĩnh vực này. Các cộng đồng và tổ chức nghiên cứu cần phát triển và thống nhất về một bộ tiêu chí chung, mang tính khách quan để đánh giá tầm quan trọng lâu dài của các tập dữ liệu nghiên cứu. Những tiêu chí này phải rõ ràng và được phổ biến rộng rãi để các nhà nghiên cứu, các nhà quản lý thông tin của các cơ quan/tổ chức có thể đưa ra quyết định hợp lý và nghiêm ngặt.

3. MỘT SỐ PHƯƠNG PHÁP LƯU TRỮ DỮ LIỆU HIỆN NAY

Khi chuyển đổi các kho lưu trữ sang định dạng kỹ thuật số, các nhiệm vụ được đặt ra để bảo đảm cho việc lưu trữ tài liệu ở dạng kỹ thuật số được tốt nhất, đó là:

- Bảo đảm tính toàn vẹn của dữ liệu trong toàn bộ thời gian lưu trữ;
- Duy trì khả năng tái tạo định dạng thông tin kỹ thuật số trong suốt thời gian lưu trữ;
- Bảo đảm độ tin cậy của phương tiện lưu trữ hoặc hệ thống lưu trữ thông tin kỹ thuật số.

Hiện tại, thời gian mà phần lớn các thư viện trên thế giới chuyển đổi sang lưu trữ dữ liệu số chưa đủ lâu để có thể khẳng định rõ ràng về các vấn đề và rủi ro có thể xảy ra đối với việc bảo quản dữ liệu số lâu dài. Đó là lý do vì sao, một trong những phương pháp lưu trữ là bảo quản tài liệu cả ở giấy, định dạng gốc (bao gồm cả vi phim, đĩa quang) và cả ở dạng kỹ thuật số.

Một vấn đề chưa được giải quyết mà hầu hết

các thư viện phải đối mặt trong quá trình số hóa là sự đa dạng của nhiều định dạng dữ liệu khác nhau (văn bản, hình ảnh, video, bảng thống kê, tài nguyên có giá trị), do đó cần những cách tiếp cận khác nhau để xử lý và lưu trữ chúng. Một số dữ liệu đã được ghi lại cách đây vài thập kỷ trên phương tiện kỹ thuật số. Song, ngay cả khi chúng được lưu trữ đúng cách, thì khả năng không thể truy cập được những dữ liệu này cũng có thể xảy ra. Điều này không chỉ do phương tiện đã lỗi thời mà còn do phương pháp lưu trữ cũng như việc ngừng hỗ trợ từ các nhà sản xuất phần mềm và hệ điều hành. Một phương pháp lưu trữ cho phép giải quyết vấn đề này là mô phỏng - tạo ra một bộ phần mềm và phương tiện kỹ thuật cho phép tái tạo chức năng của môi trường phần cứng và phần mềm ban đầu. Theo kinh nghiệm, chuyên gia lưu ý đến khả năng có thể mất thông tin trong quá trình lưu trữ và khi chuyển từ phương tiện này sang phương tiện khác. Liên quan đến vấn đề này, các phương pháp quan trọng nhất để bảo quản dữ liệu kỹ thuật số là sao lưu (lưu thông tin trên một máy chủ riêng) và nên sao chép, lưu trữ dữ liệu ở nhiều nơi.

Quá trình số hóa các bộ sưu tập truyền thống và việc bảo quản chúng theo thời gian làm nảy sinh các vấn đề bảo mật cơ bản liên quan đến quản lý quyền. Ví dụ, cơ chế cung cấp quyền truy cập cho các nhóm người dùng khác nhau khi tương tác với siêu dữ liệu lưu trữ là gì, quyền truy cập sẽ được cập nhật như thế nào khi trạng thái bản quyền hoặc mức độ bảo mật của tài liệu thay đổi. Các mô hình lưu trữ tham chiếu làm tiêu chuẩn cho kho lưu trữ kỹ thuật số có bảo vệ thông tin, thiết bị và tài nguyên kỹ thuật số đã được phát triển vào đầu thế kỷ 21, như CCSDS và OAIS (Hệ thống thông tin lưu trữ mở),... và hiện đang được sử dụng tại hầu hết các thư viện lớn ở Úc, Anh, Hà Lan và Hoa Kỳ.

Việc bảo quản dữ liệu nghiên cứu phức tạp hơn nhiều so với việc chỉ lưu trữ trên ổ cứng hoặc máy chủ. Trên thực tế, việc bảo quản dữ liệu kỹ thuật số phức tạp hơn việc lưu trữ giấy vì dữ liệu kỹ

thuật số có thể bị suy giảm nhanh hơn và trong một số trường hợp không thể lấy dữ liệu đó ra và đọc được. Phương án lựa chọn tốt hơn cả là nên chuyển dữ liệu đó cho một bên giám hộ đủ tiêu chuẩn, chẳng hạn như kho lưu trữ dữ liệu - nơi thực hiện các bước cần thiết để bảo quản dữ liệu này.

Hiện nay, một số lựa chọn lưu trữ kỹ thuật số dành cho các bộ sưu tập chủ đề lớn phổ biến và được chấp nhận rộng rãi nhất, có thể kể đến một số kho lưu trữ được biết đến trên toàn cầu như: OpenDOAR-danh mục toàn cầu về các kho lưu trữ truy cập mở (<https://v2.sherpa.ac.uk/opensoar/>). OpenDOAR chứa các kho lưu trữ cung cấp quyền truy cập mở miễn phí đến các kết quả, sản phẩm đầu ra và tài nguyên học thuật. OpenDOAR hiện bao gồm hơn 6.000 kho lưu trữ từ khắp nơi trên thế giới. Các loại nội dung của kho lưu trữ lớn này bao gồm: các bài báo trên tạp chí, luận văn, luận án, sách và chương sách, tài liệu hội nghị, bằng sáng chế, tài liệu đồ họa, bộ dữ liệu,... của nhiều chủ đề khác nhau; BioRxiv.org (kho lưu trữ điện tử miễn phí các bài báo khoa học và các bản in trước trong lĩnh vực sinh học-<https://www.biorxiv.org/>); Kho lưu trữ tri thức mở của Ngân hàng Thế giới (kho lưu trữ chính thức các kết quả nghiên cứu của Ngân hàng Thế giới - <https://openknowledge.worldbank.org/home>); và rất nhiều kho lưu trữ khác. Thông tin đầy đủ nhất và dữ liệu duy nhất được thu thập tại một nơi có liên quan đến người dùng làm việc trong các lĩnh vực hẹp hoặc có nhu cầu cụ thể. Khách truy cập tiết kiệm thời gian và thư viện kiểm soát được dữ liệu, theo dõi được việc sử dụng tài liệu, xác định nhu cầu của người đọc, tối ưu hóa quy trình tiếp nhận ấn phẩm mới. Ngoài ra, việc hỗ trợ kho lưu trữ cho phép phân bổ lại quỹ thư viện và chuyển trọng tâm từ lưu trữ bộ sưu tập và mua tài liệu sang quản lý nội dung, bao gồm cả quyền truy cập mở.

Trong tương lai, các kho lưu trữ thư viện với dữ liệu được chọn lọc và có chất lượng cao có thể là cơ sở cho việc phân tích máy-mô phỏng là quá trình nghiên cứu và diễn giải dữ liệu bằng phần mềm và thuật toán chuyên dụng được sử dụng để phân

tích văn bản trong công cụ tìm kiếm. Hiện nay, trong bối cảnh thư viện trong quá trình chuyển đổi sang tài nguyên kỹ thuật số, lưu trữ kỹ thuật số, dịch vụ trực tuyến hướng tới khả năng truy cập mở, có nhiều nghiên cứu đề cập đến những cơ hội mới để cải thiện hoạt động của thư viện thông qua việc sử dụng trí tuệ nhân tạo để thu thập và xử lý dữ liệu.

Một phương pháp lưu trữ khác giúp giải quyết vấn đề hao mòn và lỗi thời là cập nhật thường xuyên-chuyển thông tin từ một phương tiện vật lý này sang một phương tiện vật lý khác hoặc hệ điều hành khác. Di chuyển dữ liệu trong thư viện là quy trình lưu trữ phức tạp, cần phải tính đến không chỉ sự khác biệt về công nghệ, mà còn cả các chi tiết cụ thể khác trong quá trình chuyển đổi.

Các phương pháp lưu trữ thông tin trong thư viện và trung tâm thông tin cần được sử dụng kết hợp các công cụ phần mềm và công nghệ là: bảo quản, chuyển tài liệu sang định dạng kỹ thuật số, mô phỏng, sao lưu, sao chép, cập nhật thường xuyên, di chuyển dữ liệu, lưu trữ đám mây. Bắt đầu từ giai đoạn lập kế hoạch và kết thúc bằng việc kiểm tra tính toàn vẹn của thông tin đã lưu. Công việc này cần có sự tham gia của các chuyên gia về công nghệ và các lập trình viên.

KẾT LUẬN

Các thư viện hiện đang phải đối mặt với bốn thách thức toàn cầu trong việc bảo tồn các bộ sưu tập, đó là:

- Tạo điều kiện để bảo quản bộ sưu tập truyền thống;
- Chuyển đổi tài liệu sang định dạng kỹ thuật số;
- Duy trì tính an toàn của tài liệu ở bất kỳ định dạng nào;
- Bảo đảm an ninh dữ liệu ở mọi giai đoạn lưu trữ.

Một trong những vấn đề chính liên quan đến việc truyền và lưu trữ dữ liệu là lựa chọn phương tiện lưu trữ cho dữ liệu số. Tính ổn định, độ tin cậy và khả năng tương thích của phương tiện lưu trữ với các hệ thống lưu trữ hiện có cũng cần được xem xét.

Phát triển quy trình di chuyển dữ liệu là một

nhiệm vụ quan trọng bao gồm việc tìm ra các thuật toán đặc biệt không chỉ có thể chuyển dữ liệu hiệu quả sang phương tiện kỹ thuật số mà còn duy trì tính toàn vẹn của chúng sau khi chuyển.

Khi sử dụng các công nghệ tiên tiến để giải quyết các vấn đề của thư viện và đáp ứng nhu cầu thay đổi của người dùng, các thư viện cần xử lý các vấn đề lưu trữ cả các bộ sưu tập truyền thống và tài liệu kỹ thuật số. Do vậy, việc bảo đảm lưu trữ thông tin lâu dài là nhiệm vụ quan trọng mà một thư viện hiện đại cần phải giải quyết. Vấn đề đặt ra ở đây là, các công nghệ (bao gồm cả công nghệ thông tin) và phần mềm không ngừng phát triển, và thay đổi. Chính vì vậy, các thư viện và trung tâm thông tin cũng cần phải lựa chọn các giải pháp phù hợp để việc lưu trữ hiệu quả nhất và tối ưu hóa các dịch vụ của mình.

Nguyễn Thị Tú Quyên

TÀI LIỆU THAM KHẢO

1. Anuradha Maurya#, and Subaveerapandiyan A (2022). Research Data Preservation Practices of Library and Information Science Faculties. Journal of Library & Information Technology 42, 4, 259-264.
2. Data management costing tool and checklist, version 3, UK Data Service. UK Data Service (UKDS). 2015. URL: ukdataservice.ac.uk/manage-data/plan/costing.
3. Data Management Plan (DMP) Guide. Truy cập tại: <https://instr.iastate.libguides.com/dmp/step5> (Tháng 6/2025).
4. Directory of Open Access Repositories (OpenDOAR). Truy cập tại: <https://library.iimcal.ac.in/opendoar/> (Tháng 6/2025).
5. Whyte A., Wilson A. How to Appraise and Select Research Data for Curation. DCC How-to Guides. Edinburgh: Digital Curation Centre. 2010. URL: dcc.ac.uk/resources/how-guides/appraise-select-data.
6. E. B. Бескаравайная1, И. А. Митрошин (2024). Современные методы хранения данных в библиотеках: ключевые аспекты и решения. Научные и технические библиотеки, 8, 78-97.