BẢO QUẢN SỐ TRONG CÁC THƯ VIỆN VÀ CƠ QUAN LƯU TRỮ

ThS Lê Bá Lâm
Trung tâm Thư viện và Tri thức số, ĐHQG Hà Nội

Tóm tắt: Bài báo giới thiệu về bảo quản số- vấn đề đang được quan tâm trong các thư viện và cơ quan lưu trữ; phân tích, minh họa mô hình Hệ thống thông tin lưu trữ mở (The Open Archival Information System- OAIS) và các yếu tố Quản lý - Công nghệ - Nội dung để xây dựng thành công một dự án bảo quản số, đồng thời nêu những thách thức và chiến lược trong bảo quản số.

Từ khóa: Bảo quản số; đối tượng số; bộ sưu tập số; thư viện số; lưu trữ số; mô hình OAIS.

DIGITAL PRESERVATION IN LIBRARIES AND ARCHIVING AGENCIES

Abstract: This article introduces digital preservation is the issue which is concerned in libraries and archives; Analysis, illustration The Open Archival Information System (OAIS) models and elements about Management-Technology-Content for successfully built a digital preservation project and finish article presents challenges and strategies in digital preservation.

Keywords: Digital preservation; digital objects; digital collections; digital Library; digital archives; OAIS models.

GIỚI THIỆU

Bảo quản, lưu trữ số (sau đây gọi tắt là Bảo quản số) là một lĩnh vực, chủ đề không mới của nghiên cứu và thực tiễn trong các thư viện và cơ quan lưu trữ, nhưng gần đây, vấn đề này được các nhà nghiên cứu đặc biệt quan tâm và nhiều kết quả nghiên cứu được đưa ra nhằm giúp các dự án bảo quản số trong thực tế đạt hiệu quả cao nhất.

Bảo quản số tập trung vào chiến lược, chính sách, công nghệ và dữ liệu nhằm đảm bảo các đối tượng và bộ sưu tập số luôn sẵn sàng cho việc tìm kiếm, truy cập và sử dụng được ở hiện tại và tương lai. Bảo quản số cũng chính là đảm bảo an ninh, an toàn cho các tài liệu được sinh ra ở định dạng số cũng như các tài liệu dạng truyền thống đã được chuyển đổi số thông qua quá trình số hóa.

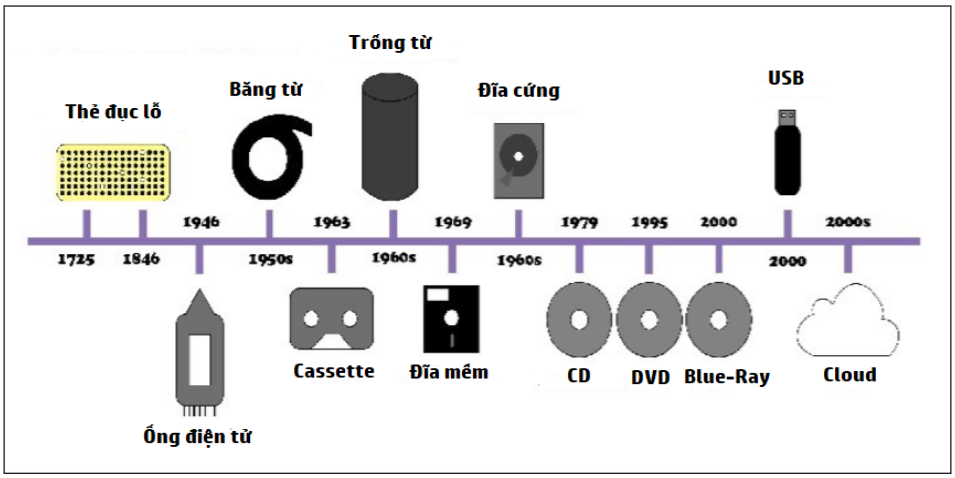
Theo Thư viện Quốc hội Mỹ, bảo quản số là "các hoạt động quản lý nội dung số giúp đảm bảo, khả năng truy cập liên tục vào các đối tượng số" [Library of Congress, 2013]. Corrado & Moulaison (2014) thì cho rằng, bảo quản số là một vấn đề phức tạp về kỹ thuật, xã hội, kinh tế và của các tổ chức. Tính phức tạp của nó trong thư viện bắt nguồn từ thực tế là nó được đan xen vào quá trình tạo, sử dụng và duy trì các bộ sưu tập và tài liệu số. Tính bền vững của tài liệu số phụ thuộc vào việc quản lý, phòng các rủi ro trong bảo quản, chính sách tổ chức, cam kết thể chế và cơ sở hạ tầng kỹ thuật.

Tại Hội thảo Annual Conference, Washington, D.C., June 24, 2007, American Library Association's (2007) đưa ra định nghĩa: Bảo quản số là sự kết hợp các chính sách, chiến lược và hành động để đảm bảo tính chân thực, chính xác của nội dung thông tin qua thời gian, bất chấp sự thay đổi, lỗi thời của công nghệ. Bảo quản số áp dụng chung cho tài liệu số nguyên gốc (born digital materials) và tài liệu số hóa (digitalized materials) là kết quả của quá trình số hóa.

Như vậy có thể thấy rằng, những phát biểu, nhận định và định nghĩa đưa ra trên đây đều khẳng định bảo quản số là một loạt những hoạt động từ quản lý đến công nghệ và triển khai xây dựng nội dung số cho các bộ sưu tập số, giúp cho việc truy cập vào các đối tượng số được thường xuyên, liên tục và lâu dài cho dù công nghệ phát triển, thay đổi hàng ngày, hàng giờ và làm cho mọi thứ đều trở nên nhanh chóng bị lỗi thời.

1. SỰ NHANH CHÓNG LỖI THỜI CỦA CÔNG NGHỆ

Tài liệu giấy, tác phẩm nghệ thuật có thể cho phép người dùng tin đọc, sử dụng thông tin được trong nhiều năm, nhiều thế kỷ hoặc thậm chí thiên niên kỷ. Với mục tiêu chuyển định dạng số để lưu giữ, bảo quản lâu dài thì cũng chưa hẳn các tác phẩm, công trình đó có thể yên tâm sử dụng mãi mãi nếu không có sự quan tâm trong công tác bảo quản số do tốc độ thay đổi công nghệ nhanh chóng, tài liệu có thể không truy cập được chỉ sau một vài năm được tạo ra. Khi thông tin được tạo ra bằng kỹ thuật số và các công nghệ mới phát triển, các định dạng cũ hơn sẽ trở nên lỗi thời, do đó nội dung tài liệu có thể không truy cập được khi sử dụng các phần mềm, ứng dụng mới hoặc do hỏng hóc của các bộ lưu trữ.



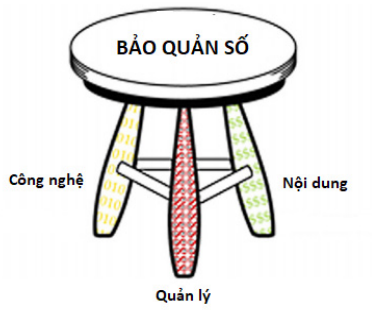
Hình 1. *Các thiết bị lưu trữ thông tin theo thời gian*
(Nguồn: https://slidetodoc.com/digital-asset-management-systems-and-digital-preservation-euan/)

Ví dụ, ở Việt Nam, một người đã viết công trình khoa học của mình vào đầu những năm 1990 bằng phần mềm Bked, VietStar, một trong những ứng dụng xử lý văn bản phổ biến nhất trong thời kỳ đó. Họ đã lưu trữ công trình của mình trên một đĩa mềm và bây giờ muốn tham khảo lại nhưng không còn ổ đĩa để đọc nó nữa. Và hiện tại, giả sử vẫn còn ổ đĩa và các phần cứng để lấy file tài liệu ra và file ở trong tình trạng có thể truy xuất được thì điều đó cũng không đảm bảo chắc chắn rằng file tài liệu đó được mở ra và đọc được bình thường vì không ai còn sử dụng những phần mềm đã tạo ra file đó. Tại thời điểm này, các định dạng phổ biến cho lưu trữ là PDF, PDF/A là tiêu chuẩn quốc tế (ISO) ISO 19005-1 được thiết kế cho lưu trữ lâu dài, các tài liệu dạng văn bản được số hóa cũng thường lưu ở định dạng này và rồi 30-40 năm nữa liệu tình trạng không đọc được các định dạng PDF có xảy ra như với các file Bked, VietStar nêu trên?

2. CÁC YẾU TỐ CHÍNH TRONG BẢO QUẢN SỐ

Để bắt đầu một dự án hoặc chương trình nào đó, bao giờ cũng có những khó khăn, bắt đầu từ đâu, các bước tiếp theo là gì và kết thúc như thế nào, đặc biệt là những vấn đề mới và phải lựa chọn công nghệ phù hợp như dự án bảo quản số của một tổ chức. Tuy nhiên, những thách thức trong bảo quản số không phải là những vấn đề không thể vượt qua nếu có quyết tâm và ủng hộ tuyệt đối từ những nhà quản lý và những người triển khai trực tiếp. Các tài liệu in ấn, bản thảo hoặc hiện vật có thể tồn tại nhiều năm mà không bị hỏng hóc đáng kể, hoặc tốn thêm nhiều chi phí để bảo quản, nhưng đối với các đối tượng số thì không hẳn như vậy. Các đối tượng số không thể cứ nằm trong các bộ sưu tập, kho lưu trữ mà không được bảo quản, chăm sóc thường xuyên do các yếu tố công nghệ như lạc hậu, hỏng hóc của các bộ phận lưu trữ, các phần cứng và phần mềm khác,…

Các nhà nghiên cứu về bảo quản số đưa ra nhiều quan điểm khác nhau nhưng tựu chung đều thống nhất 3 yếu tố chính, quyết định trong bảo quản số và coi nó như là một chiếc ghế ba chân, đó là: Quản lý, Công nghệ và Nội dung.



Hình 2. *Các yếu tố quyết định trong bảo quản tri thức số [Kenney & McGovern, 2003]*

Quan điểm thống nhất về 3 yếu tố: Quản lý, Công nghệ và Nội dung như 3 chân của một chiếc ghế cho thấy tầm quan trọng của mỗi yếu tố. Các yếu tố phụ thuộc lẫn nhau, có mối quan hệ mật thiết với nhau và không thể tách rời.

2.1. Yếu tố quản lý

Vấn đề bảo quản số đầu tiên được đề cập, đó là Quản lý. Các khía cạnh quản lý bao gồm lập kế hoạch, chuẩn bị nguồn lực tài chính, nhân sự, lựa chọn công nghệ và chuẩn bị nội dung cũng như các vấn đề về chính sách hay giám sát triển khai,…

2.1.1. Lập kế hoạch và thiết lập các chính sách

Kế hoạch và các chính sách là văn bản chính thức và có tính pháp lý được tổ chức phê duyệt trước khi đưa vào triển khai dự án bảo quản tri thức số. Văn bản này là cơ sở quan trọng để thực hiện các bước, công việc theo nội dung đã vạch ra. Những phòng ban, cá nhân được giao nhiệm vụ sẽ căn cứ vào kế hoạch đó để làm việc với các nhóm, các bên liên quan theo lộ trình trong kế hoạch để đi đến đích. Các chính sách ở đây có thể là lựa chọn nội dung, phân quyền truy cập sử dụng cho các đối tượng trong hay ngoài tổ chức, trách nhiệm của cán bộ và người dùng tin, hoặc các hướng dẫn,…

Một bản kế hoạch mẫu cho bảo quản số bao gồm 9 hạng mục đã được Christoph Becker và cộng sự đưa ra như dưới đây [Becker & cộng sự, 2009]:

1. Nhận diện kế hoạch, tạo điều kiện thuận lợi cho mọi người tiếp cận.
2. Các nguyên tắc xây dựng.
3. Bối cảnh xây dựng.
4. Mô tả về bộ sưu tập và các đối tượng số.
5. Các yêu cầu đối với việc bảo quản tri thức số.
6. Các kinh nghiệm và minh chứng.
7. Tài chính.
8. Vai trò và trách nhiệm của các cá nhân.
9. Kế hoạch triển khai.

Mặc dù không phải tất cả các chương trình, kế hoạch bảo quản số đều phải đảm bảo hay tuân thủ 9 nội dung trên, nhưng với mức độ chi tiết đó, nó cung cấp cho các nhà quản lý xây dựng kế hoạch bảo quản số được đầy đủ, chu đáo và hỗ trợ việc ra các quyết định đúng đắn.

Becker và cộng sự xác định năm vấn đề có thể sẽ tác động đến việc lập một kế hoạch mới, đó là: (1) Nhu cầu xây dựng một bộ sưu tập mới, (2) Thay đổi một bộ sưu tập, (3) Thay đổi môi trường lưu trữ và bảo quản, (4) Thay đổi mục tiêu và (5) Đánh giá định kỳ [Becker & cộng sự, 2009]. Trong 5 vấn đề trên thì đánh giá định kỳ là vấn đề rất cần thiết và quan trọng. Nó có thể đánh giá hiệu quả sử dụng, các công nghệ đã đầu tư có còn ổn định và cho phép duy trì, mức độ phát triển các đối tượng số và các bộ sưu tập,… Các vấn đề này sẽ ảnh hưởng lớn đến việc ra quyết định tiếp theo của các nhà quản lý là có tiếp tục duy trì, cho tồn tại hay đầu tư các nguồn lực để tiếp tục phát triển. Nếu không có đánh giá định kỳ thì nhà quản lý không thể nắm được sự vận hành và hiệu quả của đầu tư cho dự án, không nắm được tình trạng hiện tại của vấn đề.

2.1.2. Quyết sách công nghệ

Những quyết sách về công nghệ rất được quan tâm để đảm bảo cho hệ thống bảo quản được lâu dài, nâng cao tính sẵn sàng phục vụ người dùng, đặc biệt trong bối cảnh các công nghệ phần cứng, phần mềm thay đổi nhanh chóng.

2.1.3. Câu hỏi về bản quyền

Khi xây dựng kế hoạch, một vấn đề đặc biệt quan trọng cần lưu ý, đó là vấn đề bản quyền tài liệu. Nó là một dạng tài sản thuộc sở hữu trí tuệ. Thường thì tài liệu văn bản hoặc hình ảnh là nội dung có bản quyền. Tài liệu được sinh

ra từ đầu đã là định dạng số (Born Digital) hay từ công tác số hóa (Digitization) đều phải được giải quyết vấn đề bản quyền. Có thể là quyền của tác giả hay tác giả đã nhượng quyền cho một nhà xuất bản. Một vấn đề ai cũng hiểu, đó là phiên bản số hóa của tài sản trí tuệ khác với các loại tài sản khác là chúng có thể được chia sẻ mà bản gốc thì vẫn còn nguyên vẹn.

2.1.4. Các nguồn lực

Đó là các vấn đề về nguồn nhân lực, nguồn tài chính,… Nguồn nhân lực cần có trình độ, các kỹ năng cần thiết để vận hành hệ thống bảo quản số và đòi hỏi nhiều cấp độ và đa dạng ở chuyên môn. Đầu tiên là đội ngũ công nghệ thông tin (IT) để vận hành hệ thống, tiếp đến là các cán bộ chuyên môn thư viện để mô tả, biên mục, tổ chức xây dựng các bộ sưu tập. Các đối tượng số được bảo quản tốt thế nào đi nữa mà không tổ chức tốt, không có các mô tả siêu dữ liệu thì cũng sẽ hạn chế trong tìm kiếm, truy xuất thông tin để sử dụng.

2.1.5. Khả năng tiếp cận và tính bền vững

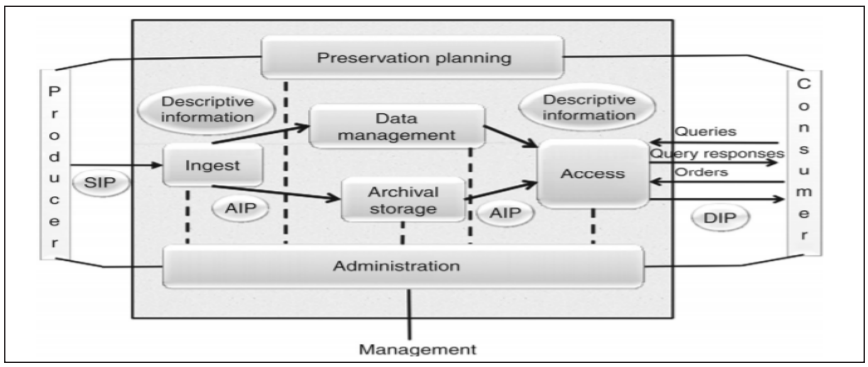
Cùng với các yếu tố trên thì khả năng tiếp cận cộng đồng và tính bền vững cần được tính tới trong quản lý. Một dự án bảo quản tri thức số có hiệu quả hay không phải được nhiều người biết đến và càng nhiều người sử dụng càng tốt (trừ những dự án liên quan đến an ninh hay quốc phòng). Làm tốt vấn đề này, ngoài việc mang lại danh tiếng cho tổ chức, thể hiện được trách nhiệm quốc gia, còn có thể mang lại nguồn lực tài chính, giúp duy trì hệ thống bền vững, mở rộng kho lưu trữ, nâng cấp hạ tầng công nghệ hoặc bổ sung đối tượng số có giá trị vào bộ sưu tập từ tài nguyên của cộng đồng đóng góp,…

2.2. Yếu tố công nghệ

Bảo quản số không phải phụ thuộc tất cả vào công nghệ, tuy nhiên không thể thực hiện nó mà không có hạ tầng công nghệ là các phần cứng, phần mềm, bộ lưu trữ, đường truyền, hệ thống mạng hay các vấn đề liên quan đến bảo mật,…

The Open Archival Information System (OAIS - Hệ thống thông tin lưu trữ mở) là mô hình hiện đại, tin cậy được xem là tiêu chuẩn cho các kho bảo quản số. Mô hình OAIS mô tả cách bảo quản các đối tượng số từ thu thập đến đăng tải, quản lý, xử lý và phục vụ người dùng. OAIS có thể áp dụng cho nhiều trường hợp bảo quản số khác nhau, nên không bắt buộc các tổ chức có dự án phải tuân thủ nghiêm ngặt các quy trình trong mô hình mà có thể mềm dẻo để áp dụng sử dụng nó. Mô hình OAIS là một tiêu chuẩn quốc tế (ISO), tiêu chuẩn ISO 14721. OAIS được phát triển bởi CCSDS (Consultative Committee for Space Data Systems) vào ngày 04/4/1994. SIP (Submission Information Package) là gói thông tin đưa vào, SIP sẽ chứa đối tượng số và siêu dữ liệu; AIP (Archival Information Package) là gói lưu trữ thông tin; PDI (Preservation Description Information) là thông tin mô tả đối tượng bảo quản.

Nhìn vào Hình 3 có thể nhận thấy 6 chức năng cơ bản tác động lẫn nhau trong OAIS là: (1) Đầu vào (Ingest), (2) Kho lưu trữ (Archival storage), (3) Quản lý dữ liệu (Data management), (4) Quản trị (Administration), (5) Kế hoạch bảo quản (Preservation Planning) và (6) Truy cập (Access).



Hình 3. *Mô hình Hệ thống thông tin lưu trữ mở OAIS [CCSDS, 2012]*

Bảng 1. *Các chức năng cơ bản trong mô hình OAIS [Corrado & Moulaison, 2014]*

Chức năng	Diễn giải
Đầu vào	Chức năng đầu vào cung cấp các dịch vụ và chức năng cho phép đưa các đối tượng kỹ thuật số vào hệ thống. Nó chấp nhận các gói thông tin SIP. Một gói thông tin SIP thường bao gồm thông tin nội dung và thông tin mô tả (PDI).
Kho lưu trữ	Chức năng kho lưu trữ cung cấp dịch vụ và các chức năng liên quan đến lưu trữ, bảo trì và truy xuất các gói thông tin lưu trữ (AIPs). Kho lưu trữ giúp đặt AIPs ở trạng thái lưu trữ vĩnh viễn, khôi phục thảm họa, kiểm tra lỗi và cung cấp AIPs cho thực thể truy cập.
Quản lý dữ liệu	Chức năng quản lý dữ liệu cung cấp dịch vụ liên quan đến duy trì, truy cập và quản trị siêu dữ liệu. Các chức năng bao gồm duy trì sơ đồ và chế độ xem, thực hiện cập nhật cơ sở dữ liệu và thực hiện các truy vấn và tạo báo cáo dựa trên các truy vấn quản lý dữ liệu.
Quản trị	Chức năng quản trị cung cấp dịch vụ và các chức năng hỗ trợ hoạt động tổng thể của hệ thống. Các chức năng quản trị bao gồm việc xem xét, kiểm tra đầu vào để đảm bảo chúng sẽ đáp ứng yêu cầu kho lưu trữ, các tiêu chuẩn và duy trì quản lý cấu hình của phần cứng và phần mềm hệ thống.
Kế hoạch bảo quản	Chức năng lập kế hoạch bảo quản cung cấp các dịch vụ và chức năng giám sát môi trường hoạt động của hệ thống OAIS, cung cấp các khuyến cáo để đảm bảo thông tin được lưu trữ trong OAIS vẫn có thể được truy cập trong dài hạn, ngay cả khi hệ thống công nghệ ban đầu trở nên lỗi thời. Các chức năng bao gồm đề xuất thông tin lưu trữ, cập nhật, di chuyển, báo cáo phân tích rủi ro và giám sát những thay đổi công nghệ và những thay đổi trong yêu cầu dịch vụ.
Truy cập	Chức năng truy cập cung cấp các dịch vụ và chức năng hỗ trợ người dùng cuối. Người sử dụng thông tin, bao gồm cả khả năng xác định sự tồn tại, mô tả, vị trí và tính khả dụng của thông tin được lưu trữ trong OAIS, cho phép người sử dụng yêu cầu và nhận sản phẩm thông tin cũng như đưa ra các phản hồi cho người dùng.

Khi một dự án bảo quản số đã được xác định và các chuyên gia đã thẩm định về mô hình công nghệ, khả năng vận hành cũng như đảm bảo tính duy trì và sự ổn định, nghĩa là thiết kế cho phép nội dung không thay đổi (sự toàn vẹn của các đối tượng số) và thuận lợi cho việc truy cập (các hệ thống truy xuất thông tin). Một hệ thống được thiết kế hợp lý để bảo quản tri thức số cần giải quyết một số vấn đề [Gorman & Dorne, 2009]: Sự toàn vẹn của các đối tượng số; Đảm bảo nội dung và truy cập phù hợp với công nghệ; Truy xuất thông tin; Siêu dữ liệu phục vụ cho truy cập và bảo quản số; Hệ thống lưu trữ; Sự chuyển đổi giữa các thế hệ phần cứng và phần mềm để đảm bảo khả năng truy cập liên tục.

Các vấn đề về công nghệ nêu trên nếu được đặt ra, xem xét cẩn trọng và giải quyết tốt thì dự án bảo quản tri thức số chắc chắn sẽ đạt kết quả tốt và mang lại hiệu quả phục vụ to lớn.

2.2.1. Lựa chọn phần mềm

Phần mềm hoặc ứng dụng được thiết kế để quản trị bảo quản số. Những đơn vị có tiềm lực công nghệ và nhân lực có thể thiết kế riêng cho mình một hệ thống bảo quản, trong khi số khác có thể lựa chọn các sản phẩm phần mềm thương mại hoặc mã nguồn mở. Hiện nay, có rất nhiều các phần mềm thương mại dành cho bảo quản số và đang được phát triển mạnh mẽ bởi các công ty hoạt động trong lĩnh vực thư viện số như: Ex Libris (Rosetta), OCLC (Content DM), Tinh Vân (Bookworm), Hiện đại (Kipos),… Một số sản phẩm bảo quản, lưu trữ số còn được phát triển bởi các công ty chuyên sản suất thiết bị số hóa như giải pháp Nainuwa của Treventus.

Các sản phẩm nguồn mở như DAITSS (Dark Archive in the Sunshine State) là một ứng dụng mã nguồn mở được phát triển bởi Trung tâm tự động hóa thư viện Florida (FCLA) với sự tài trợ của Viện Bảo tàng và Dịch vụ

thư viện (IMLS) hay Archivematica (https://ww.archivematica.org/en/) là một hệ thống mã nguồn mở được thiết kế để bảo quản số với các tiêu chuẩn cơ bản. Không giống như một số hệ thống bảo quản tri thức số khác là có cả giao diện cho người sử dụng, Rosetta (http://www.exlibrisgroup.com/category/RosettaOverview) không bao gồm giao diện tìm kiếm cho người dùng cuối, thay vào đó, nó sử dụng giao thức mở (OAI-PMH) để cho ứng dụng khám phá (Discovery) thu thập siêu dữ liệu và trình bày. Tinh Vân và Hiện đại là các công ty trong nước phát triển tính năng bảo quản và phục vụ tài liệu số tích hợp cùng với sản phẩm quản trị của thư viện truyền thống. Ứng dụng Bookworm của Tinh Vân còn mở rộng sử dụng mượn/đọc sách điện tử trên các thiết bị di động nhằm mang lại tiện ích cho người sử dụng và phần nào đảm bảo tính bảo mật cho tài liệu số. Ứng dụng Kipos của công ty Hiện đại tách dữ liệu số ra từng trang và áp dụng tiêu chuẩn truyền và mã hóa siêu dữ liệu METS.

Các phần mềm, ứng dụng mã nguồn mở khác cho bảo quản tri thức số có thể kể đến như Greenstone (https://www.greenstone.org), CDS-Invenio (https://inveniosoftware.org), Dspace (http://www.dspace.org/), Eprints (http:// www.eprints.org/), Fedora (http://fedorarepository.org/) và MyCore (https://www.mycore.com).

Việc xây dựng và quản lý một kho lưu trữ tại tổ chức đòi hỏi sự đầu tư đáng kể về nguồn lực tài chính cho hạ tầng công nghệ, nhân sự và chuyên môn, do đó, một số tổ chức quyết định giảm chi phí bằng cách tham gia vào các chương trình hợp tác hoặc thuê ngoài (sử dụng dịch vụ phần mềm) cho dự án của họ.

HathiTrust (http://www.hathitrust.org/) được công bố vào năm 2008, là một sáng kiến hợp tác của các thư viện nghiên cứu để bảo quản các tài liệu số về văn hóa. Mục tiêu ban đầu là cung cấp một nền tảng cho bảo quản, lưu trữ một khối lượng lớn các tài liệu số hóa của dự án Google Book và Open Content Alliance (OCA). Christenson (2011) nhận định "trái tim của HathiTrust là kho lưu trữ số dùng chung và vận hành bởi sự hợp tác của các thư viện nghiên cứu". Hiện tại có hơn 60 thành viên trong HathiTrust thuộc các tổ chức trên toàn thế giới.

MetaArchive (http://www.metaarchive.org/) được công bố vào năm 2003 cũng là một kho lưu trữ số cộng đồng. MetaArchive là "mạng lưu trữ kỹ thuật số phân tán do cộng đồng sở hữu và được điều hành bởi cộng đồng" [Walters & Skinner, 2010]. Các thành viên của MetaArchive đóng phí thành viên, có đơn vị cử nhân viên hoặc góp trang thiết bị. MetaArchive được phát triển bởi Đại học Stanford, có 50 thành viên đến từ 13 bang và 3 quốc gia.

Dịch vụ lưu trữ trực tuyến (hosting) hiện nay rất phát triển, các dự án tham gia sẽ phải đóng phí duy trì dịch vụ. Chi phí được tính thường dựa vào số lượng đối tượng số và/hoặc dung lượng tính bằng terabyte của các bộ sưu tập. Sử dụng dịch vụ này, tổ chức không phải lo về vấn đề hạ tầng công nghệ cũng như quản trị, sao lưu kho lưu trữ. Các tổ chức chỉ chuẩn bị đối tượng số, chăm sóc người dùng và phát triển bộ sưu tập. OCLC DigitalArchive, DuraCloud là những dịch vụ được đánh giá cao và tin cậy.

OCLC DigitalArchive (http://www.oclc.org/digital-archive.en.html) là giải pháp lưu trữ dành cho các dự án muốn sử dụng dịch vụ trực tuyến. Phần mềm CONTENTdm là phần mềm mà OCLC sử dụng cho giải pháp này.

DuraCloud (http://www.duracloud.org/) là một dịch vụ lưu trữ trực tuyến, được cung cấp bởi DuraSpace, một tổ chức phi lợi nhuận được thành lập vào năm 2009 bởi DSpace Foundation và Fedora Commons. DuraCloud sử dụng phần mềm mã nguồn mở Dspace để triển khai dịch vụ. Ngoài dịch vụ lưu trữ đối tượng số trên, DuraCloud còn cung cấp các dịch vụ khác như truy cập, chuyển đổi và chia sẻ dữ liệu.

2.2.2. Lựa chọn phần cứng

Cấu hình, số lượng, chủng loại máy chủ, bộ lưu trữ và các thành phần khác của hệ thống phụ thuộc vào kích thước các bộ sưu tập hiện tại và sự tính toán phát triển nó trong tương lai. Nhiều hệ thống bảo quản tri thức số có hệ điều hành dựa trên GNU/Linux- hoặc UNIX hoặc Windows Server và phần cứng sẽ cần phải tương thích với yêu cầu của hệ điều hành. Một yếu tố quan trọng là ngoài hệ thống hạ tầng công nghệ vận hành chính thì phải quan tâm đến hệ thống sao lưu. Những dự án lớn có thể có hệ thống sao lưu, phục hồi đặt ở một địa

điểm khác, khoảng cách đủ để bảo đảm rằng thiên tai, hỏa hoạn xảy ra ở địa điểm chính không thể tác động đến.

Việc tính toán dung lượng của hệ thống lưu trữ cũng phụ thuộc vào quyết định sẽ để bao nhiêu bản sao của đối tượng số hay định dạng của đối tượng số cũng quyết định đến dung lượng của các kho chứa. Ví dụ, tài liệu được số hóa bước 1 sẽ ở định dạng ảnh, chúng ta hoàn toàn có thể sử dụng, bảo quản ngay dữ liệu này hoặc ở bước 2 nhận dạng ký tự và chuyển đến định dạng PDF/A. Vậy, quyết định lưu giữ cả 2 hay chỉ sử dụng tài liệu đã nhận dạng ký tự cũng là một vấn đề cần tính toán và tất nhiên lưu giữ cả 2 sẽ phải tốn thêm bộ nhớ, đòi hỏi phần cứng lưu trữ có dung lượng lớn hơn.

2.2.3. Siêu dữ liệu

Siêu dữ liệu là một yếu tố quan trọng giúp cho lưu trữ và truy xuất thông tin đến đối tượng số được thuận lợi. Siêu dữ liệu cũng có thể gọi là chìa khóa để khai thác hiệu quả hệ thống bảo quản tri thức số. Mô tả cơ bản hay chi tiết phụ thuộc vào quy định và chính sách của tổ chức.

2.2.4. Định dạng tài liệu

Cơ quan đăng ký định dạng số toàn cầu The Global Digital Format Registry (GDFR) xác định hai loại định dạng riêng biệt là: định dạng nội dung và định dạng vật lý. Ví dụ, về các định dạng nội dung ảnh là JPEG (Joint Photographic Experts Group) và TIFF (Tagged Image File Format) và định dạng vật lý là ISO 966: 1988 hay còn được biết là Compact Disc File System (CDFS) được sử dụng trên đĩa CD-ROM.

Lựa chọn các định dạng file trong các dự án bảo quản tri thức số tùy theo nhu cầu và đặc tính của lưu trữ và bảo quản cũng như chức năng nhiệm vụ của các đơn vị là khác nhau nhưng về cơ bản các nhà quản lý và chuyên gia sẽ lựa chọn các định dạng file có tính mở và tính phổ biến cao. Tính mở có nghĩa là định dạng không phụ thuộc bản quyền, pháp lý khi sử dụng và tính phổ biến là mức độ định dạng được sử dụng rộng rãi, phổ thông. Các công cụ, phần mềm/ứng dụng quản trị đối tượng số cũng thường căn cứ vào tính mở, tính phổ biến để xây dựng và phát triển.

Khi đánh giá các định dạng file để đưa vào bảo quản tri thức số phải xem xét các yếu tố này. Nếu một file PDF là định dạng của một đối tượng số khác được nhúng vào thì cũng có thể chúng ta không còn được lưu giữ đầy đủ định dạng của bản gốc đó. Ưu điểm của một file PDF là hiển thị giống nhau trên những môi trường làm việc khác nhau, vì vậy nó làm cho định dạng này ngày càng trở nên phổ biến và cũng là lý do tại sao mọi người thích PDF/A, một phiên bản PDF chuyên dụng được thiết kế để bảo quản tri thức số lâu dài. PDF là định dạng của Adobe, là một tiêu chuẩn quốc tế (International Organization for Standardization-ISO). Một số ưu điểm khác của định dạng PDF là: Nội dung trình bày đa dạng cùng với khả năng bảo mật tốt; Có thể in ra trên bất cứ thiết bị nào mà vẫn giữ nguyên được định dạng; Hỗ trợ trên hầu hết các loại thiết bị di động; PDF thường có kích thước nhỏ khiến cho việc di chuyển, chia sẻ dễ dàng.

Các định dạng văn bản khác thường được sử dụng là RTF (Rich Text Format), Ngôn ngữ đánh dấu eXtensible Markup Language (XML) và Ngôn ngữ đánh dấu siêu văn bản Hypertext Markup Language (HTML). Đối với các loại bảng tính, định dạng CommaSeparated Values (CSV) hoặc OpenDocument Spreadsheets (ODS) được ưa thích sử dụng nhiều hơn vì mang tính mở thay vì sử dụng định dạng XLS, XLSX của Microsoft.

Đối với tài liệu ảnh, các định dạng thường sử dụng là TIFF và JPEG. TIFF ở dạng chưa nén nên kích thước thường lớn hơn JPEG, nhưng số lượng ứng dụng mã nguồn mở để xem ở định dạng JPEG thì chưa phát triển nhiều. Một số định dạng khác của ảnh số như Portable Network Graphics (PNG) và Scalable Vector Graphic (SVG) cũng được quan tâm và đưa vào tiêu chuẩn bảo quản.

Tài liệu dạng âm thanh và video cũng là một dạng đối tượng số cần bảo quản. Thuộc tính của loại hình tài liệu này mang đến nhiều thách thức cho các dự án bảo quản tri thức số. Ví dụ, các file video có phần ghi âm thanh riêng, hoặc có những video xuất hiện thêm các phụ đề được chèn vào sau. Vì không có khuyến cáo cho một chuẩn cụ thể nào về tài liệu có định dạng này nên các dự án sẽ phải tự quyết định xem định dạng nào tối ưu nhất cho tổ chức của

họ. Định dạng Audio Layer III thường được gọi là MP3 được nhiều người biết đến và sử dụng, nhưng đối với các chuyên gia, họ lại không ưu tiên đưa vào bảo quản vì nó sử dụng công nghệ nén dữ liệu, làm mất đi nhiều chất lượng của bản gốc. Định dạng Broadcast Wave Format (BWF) và Waveform Audio Format (WAV) là hai định dạng thường được sử dụng để bảo quản. Một số dự án lựa chọn định dạng Free Lossless Audio Codec (FLAC) cho kho lưu trữ của họ. Các định dạng video là AVI/MP4 là định dạng được nhắc đến nhiều và đưa vào lưu trữ, bảo quản tri thức số.

Cơ quan phụ trách về Thư viện và Lưu trữ Canada đã đưa ra 5 tiêu chí đánh giá các định dạng file để đưa vào lưu trữ, bảo quản tri thức số (Library and Archives Canada), phần nào đó giúp cho các nhà quản lý và công nghệ lựa chọn các định dạng tài liệu cho dự án của mình như: Tính công khai, minh bạch; Tính phổ biến; Tính ổn định và tương thích; Sự phụ thuộc và khả năng tương tác với các phần cứng, phần mềm; Tính chuẩn hóa.

2.3 Yếu tố nội dung

Nội dung là yếu tố thứ 3 trong chiếc ghế ba chân của bảo quản số. Đây có thể gọi là yếu tố trọng tâm vì chính sách, kế hoạch quản lý và công nghệ có tốt đến đâu mà không có nội dung thì sẽ thiếu đi yếu tố quyết định. Thu thập, tổ chức nội dung để lưu giữ thường liên quan đến các lĩnh vực của tổ chức. Nội dung đối tượng số để bảo quản trước mắt là tài liệu mà tổ chức sở hữu, chẳng hạn như các bộ sưu tập tài liệu nội sinh trong thư viện, cơ quan lưu trữ hay tài liệu có được từ các quan hệ và hợp tác cũng như sưu tầm của tổ chức.

2.3.1. Nội dung để người dùng sử dụng hợp pháp

Cung cấp nội dung có thể sử dụng là một trong những mục tiêu của việc duy trì hệ thống bảo quản tri thức số. Bất kể nội dung đối tượng số nào được bảo quản thì các vấn đề bản quyền tài liệu cần được đưa lên hàng đầu. Các nhà quản lý phải giải quyết để đảm bảo rằng các quyền sở hữu trí tuệ tác giả, nhà xuất bản đã được cấp phép, đảm bảo yêu cầu về mặt pháp lý để thực hiện các bước cần thiết để triển khai dự án.

2.3.2. Phát triển nội dung

Phát triển nội dung số ở đây cũng tương tự như sự phát triển nội dung, các bộ sưu tập tài liệu in trong các thư viện, cơ quan lưu trữ hay các bảo tàng, nghĩa là các hoạt động trong đó có thể làm gia tăng và cũng có cả thanh lọc. Để có nội dung tốt, các cơ quan, tổ chức, đơn vị đều có bộ phận thẩm định, giám tuyển chất lượng tài liệu để bổ sung vào bộ sưu tập.

Website của IBM về Big data có đăng tải thông tin: "90% dữ liệu trên thế giới ngày nay được tạo ra chỉ trong hai năm qua", vì vậy các thư viện, cơ quan lưu trữ không thể sưu tầm tất cả mà phải có chọn lọc.

- Các bộ sưu tập ban đầu

Đa số các tổ chức khi bắt đầu vào một chương trình bảo quản tri thức số sẽ có sẵn các đối tượng số để từ đó căn cứ vào nội dung, chủ đề, thuộc tính, định dạng,… để xây dựng các bộ sưu tập ban đầu. Các đối tượng số này cũng có thể được tạo ra từ công tác số hóa hay chuyển đổi định dạng. Đối với thư viện đại học, các đối tượng số ban đầu có thể là khóa luận, luận văn, luận án hay các bài trong kỷ yếu hội nghị hội thảo. Một số đơn vị có xuất bản tạp chí, đây cũng là nguồn tài liệu số có thể đưa vào lưu trữ, bảo quản ban đầu để phục vụ lâu dài. Kiểm kê, phân loại, chuyển định dạng tài liệu (ví dụ, từ bản word sang pdf) là những công việc phải triển khai để xây dựng các bộ sưu tập ban đầu.

- Phát triển bộ sưu tập mới

Sự phối hợp với các thành viên của tổ chức, mở rộng quan hệ hợp tác, tăng cường sưu tầm hay tiếp nhận trao đổi, tặng biếu hoặc tăng cường đội ngũ cộng tác viên là những biện pháp cơ bản gia tăng nguồn nội dung để mở rộng, có thêm các chủ đề để xây dựng các bộ sưu tập mới. Việc phối hợp thường xuyên với các nhà xuất bản để nhận thông tin, mua bản quyền sử dụng các đối tượng số cũng là một phương án mà các dự án bảo quản tri thức số thường áp dụng. Một nguồn tài liệu có giá trị khác là từ các cá nhân và các địa phương, họ có trong tay các tài liệu quý và cũng có nhu cầu bảo tồn nhưng không có kinh phí và công nghệ, khi đó thỏa thuận giữa tổ chức và các đối tượng trên để đạt mục đích thỏa mãn cả 2 phía là lựa chọn không thể tốt hơn. Có thể đặt tên giải pháp này là "Đôi bên cùng có lợi".

Sử dụng nội lực để số hóa các nguồn nội dung của tổ chức là một phương án gia tăng các đối tượng số và bộ sưu tập hữu hiệu. Khó khăn nhất của công tác này là các thỏa thuận để đạt được sự đồng ý của các cá nhân và tổ chức.

3. NHỮNG THÁCH THỨC VÀ CHIẾN LƯỢC TRONG BẢO QUẢN TRI THỨC SỐ

3.1. Thách thức

Không giống như tài liệu truyền thống, khi mà nội dung và vật mang tin không thể tách rời, các đối tượng số lại không được gắn với bất kỳ phương tiện lưu trữ cố định nào. Nội dung được mã hóa bởi các byte, bit dạng 0 1 và sao chép từ bộ lưu trữ này sang bộ lưu trữ khác hoặc truyền tải qua mạng. Việc không gắn liền đối tượng số với vật mang tin cố định dễ dẫn đến bị thay đổi, hư hỏng thậm chí bị phá hủy hoàn toàn và các mô tả siêu dữ liệu tách biệt hẳn với nội dung các đối tượng số cũng gây khó khăn cho việc xác định nguồn gốc hay các quyền đối với đối tượng số. Do các siêu dữ liệu tách biệt với đối tượng số nên một đối tượng số (có thể có nhiều bản sao) đồng thời cũng có nhiều siêu dữ liệu khác nhau nên việc xác định chính xác ở các kho lưu trữ khác nhau hoặc ngay trên cùng một kho lưu trữ cũng là một thách thức.

Brown (2013) chỉ ra 2 mối đe dọa đối với các đối tượng số:

- Sự mất mát đối tượng dữ liệu bởi yếu tố vật lý khi mã hóa thông tin.

- Sự mất mát đối tượng thông tin bởi yếu tố xác thực thông tin.

Một thách thức khác là xác định bản sao nào của đối tượng số là đối tượng được dùng để đưa vào bảo quản. Thông tin số rất linh hoạt và dễ thay đổi. Thibodeau (2012) mô tả nó là "đa hình thái", sự đa hình thái này là kết quả của các tác động: Thay đổi thiết bị lưu trữ; Xác định ranh giới giữa các đối tượng số; Mối quan hệ phức tạp giữa các đối tượng dữ liệu được lưu trữ trong hệ thống và các đối tượng được trình bày cho người dùng thông qua hệ thống trực tuyến; Xử lý dữ liệu của máy tính và kết xuất, truyền tải thông tin.

Như vậy, có rất nhiều thách thức đặt ra đối với các dự án bảo quản tri thức số, trong đó có cả yếu tố khách quan và chủ quan; cả những rủi ro về công nghệ và con người. Để giảm thiểu các rủi ro đó, các kế hoạch phải được lập chi tiết, cẩn thận và thường xuyên kiểm tra, đặc biệt hệ thống sao lưu, phục hồi dự phòng phải vận hành tốt và định kỳ theo lịch định.

3.2. Chiến lược

Không có một quy chuẩn nào về chiến lược bảo quản tri thức số cho chúng ta học tập. Cách tiếp cận tốt nhất có lẽ là sự kết hợp và lựa chọn phù hợp với tổ chức tùy thuộc vào sự thay đổi của môi trường công nghệ và các loại hình đối tượng số cần bảo quản. Các chiến lược bảo quản nhằm giải quyết các rủi ro bao gồm:

- Sao lưu cả cơ sở dữ liệu, đơn giản gọi là "tạo một bản sao cơ sở dữ liệu", đề cập đến việc tạo nhiều bản sao của các đối tượng. Biện pháp này không phải là một chiến lược bảo quản lâu dài mà đúng hơn nó như là một biện pháp phòng ngừa, bảo vệ dữ liệu khỏi các lỗi do yếu tố vật lý [DPM Tutorial, 2003-15].

- Làm tươi dữ liệu (Refresh) để giảm thiểu sự lỗi thời của thiết bị. Có thể hiểu là thay bộ lưu trữ, thay phần cứng mới.

- Chuẩn hóa định dạng đối tượng số là một hình thức thay đổi định dạng được thực hiện khi thu thập hoặc nhập để đưa vào kho lưu trữ. Mục tiêu của chuẩn hóa là chuyển đổi dữ liệu thành các định dạng mở và nhất quán hoặc để giảm thiểu số lượng các định dạng được quản lý trong một kho lưu trữ.

- Mô phỏng là một chiến lược để chống lại sự lỗi thời của công nghệ. Thay vì chuyển đổi đối tượng số sang các định dạng mới, mô phỏng vẫn giữ các đối tượng số ở dạng ban đầu, nhưng tái tạo lại chức năng của một nền tảng lỗi thời, phần lớn thông qua việc sử dụng phần mềm mô phỏng. Mô phỏng thường được sử dụng trong việc bảo quản trò chơi nhưng cũng có thể áp dụng để bảo quản các đối tượng đa phương tiện trong bảo quản tri thức số.

KẾT LUẬN

Với những nội dung trên đây, có thể thấy tầm quan trọng của bảo quản số nhằm duy trì tài nguyên thông tin tri thức số lâu dài và bền vững. Việc bảo quản số không đơn thuần là thường xuyên sao lưu và phục hồi dữ liệu khi

các đối tượng số bị hỏng hóc do trang thiết bị, hạ tầng công nghệ và có thể là chủ quan của con người mà bảo quản số là một chuỗi công việc chuyên nghiệp từ quản lý, lập kế hoạch, tài chính, các chính sách, lựa chọn công nghệ, xây dựng và phát triển các đối tượng số để đưa vào bộ sưu tập cho người sử dụng,… Một yếu tố khác không thể thiếu đó là thường xuyên xem xét, đánh giá hiệu quả của kho bảo quản số, qua đó các nhà lãnh đạo, quản lý có những quyết sách phù hợp để duy trì, nâng cao chất lượng cũng như đảm bảo an toàn, an ninh hệ thống nhằm mục tiêu bảo quản số tốt nhất và lâu dài nhất.

TÀI LIỆU THAM KHẢO

1. American Library Association's (ALA), 2007. Annual Conference, Washington, D.C., June 24, 2007. Available from: https://www.ala.org/alcts/resources/preserv/defdigpres0408.
2. Becker et al., 2009. Systematic planning for digital preservation: evaluating potential strategies and building preservation plans. International Journal on Digital Libraries volume 10, pages133-157, 2009.
3. Brown, A., 2013. Practical Digital Preservation: A How-To Guide for Organizations of Any Size. Neal-Schuman, Chicago.
4. Candela, L., Castelli, D., Pagano, P., Thano, C., Ioannidis, Y., Koutrika, G., and Schuldt, H., 2007. Setting the foundations of digital libraries: the DELOS manifesto. D-Lib Mag., 13 (3), 4. Available from: http://www.dlib.org/dlib/march07/castelli/03castelli.html.
5. Christenson, H., 2011. HathiTrust: a research library at web scale. Lib.Res. Tech. Serv. 55 (2), 93-102.
6. CCSDS: Consultative Committee for Space Data Systems, 2012. Reference Model for an Open Archival Information System (OAIS). Washington, DC: CCSDS. Available from: http://public.ccsds.org/publications/archive/650x0m2.pdf.
7. Corrado, E.M., Moulaison, H.L., 2014. Digital Preservation for Libraries, Archives, and Museums. Rowman & Littlefield, Lanham, MA.
8. DPM Tutorial, 2003-15. Digital Preservation Management. Cornell University Library. Available from: http:// www.dpworkshop.org/.
9. Gorman, G.E. và Dorne D.G., 2009. Bảo quản tài liệu số và đào tạo quản trị thông tin trong bối cảnh châu Á. Đại hội cán bộ thư viện các nước Đông Nam Á lần thứ XIV (CONSAL XIV), Hà Nội, 21-23 tháng 4 2009. (Lê Thùy Dương dịch).
10. Kenney, A.R., McGovern, N.Y., 2003. The five organizational stages of digital preservation. In: Hodges, P., Bonn, M., Sandler, M., Wilkin, J.P. (Eds.), Digital Libraries: A Vision for the Twenty-First Century, A Festschrift to Honor Wendy Lougee. The University of Michigan Scholarly Monograph Series. Available from: http://quod. lib.umich.edu/s/spobooks/bbv9812.0001.001/--digital-libraries-a-vision-for-the-21st-century.
11. Library and Archives Canada, "Library and Archives Canada, Local Digital Format Registry (LDFR) File Format Guidelines for Preservation and Long-term Access Version 1.0," accessed April 23, 2013, http://www.collectionscanada.gc.ca/obj/012018/f2/012018-2200-e .pdf.
12. Library of Congress, 2013. "Formats, Evaluation Factors, and Relationships," last modified March 20, 2013, http://www.digitalpreservation.gov/formats/intro/format_eval_rel .shtml.
13. Thibodeau, K., 2012. Wrestling with shaper-shifters: perspectives on preserving memory in the digital age. In: Proceedings of the Memory of the World in the Digital Age: Digitization and Preservation, pp. 15-23. Available from: http://www.ciscra.org/docs/UNESCO_MOW2012_Proceedings_FINAL_ENG_Compressed.pdf.
14. Walters, T.O., Skinner, K., 2010. Economics, sustainability, and the cooperative model in digital preservation. Lib. Hi Tech. 28 (2), 259-272.

(Ngày Tòa soạn nhận được bài: 12-11-2021; Ngày phản biện đánh giá: 06-01-2022; Ngày chấp nhận đăng: 15-3-2022).