CAM KHÔNG PHẢI LÀ QUẢ DUY NHẤT HAY NGUYÊN NHÂN CÁC THƯ VIỆN SỐ CHƯA ĐÁP ỨNG ĐƯỢC HOÀN TOÀN CHO NGHIÊN CỨU ĐIỆN TỬ

Dana McKay
Thư viện Đại học Công nghệ Swinburne, Ôxtrâylia

1. Mở đầu

Việc nắm bắt và bảo quản dữ liệu nghiên cứu để các nhà nghiên cứu khác sử dụng đã được khuyến khích như một hoạt động hữu ích ngay từ năm 2000 [1]. Mối quan tâm về việc bảo quản dữ liệu và xúc tiến "nghiên cứu điện tử" (cũng được biết đến như là khoa học điện tử và hạ tầng cơ sở mạng) chỉ tăng lên khi những nhà tài trợ nghiên cứu và chính phủ các nước nhận ra "một trận đại hồng thủy dữ liệu" [2] và dữ liệu ngày càng được nhìn nhận như lợi ích công cộng, đặc biệt khi nó được trả bằng tiền công quỹ [3].

Trong khi việc chia sẻ dữ liệu đang được các cơ quan nhà nước thúc đẩy và lập kế hoạch [4-6] (kể cả ở Ôxtrâylia [7]), ít có nghiên cứu tập trung vào công việc và mối quan tâm liên quan đến chia sẻ và quản trị dữ liệu của các nhà nghiên cứu. Công trình hiện có trong lĩnh vực này có xu hướng tập trung vào những nhóm nghiên cứu đã tiến hành quản trị và chia sẻ dữ liệu, thí dụ [8], hoặc nhằm vào cách thức các thư viện số có thể tận dụng được từ các nghiên cứu, mà trong đó, đã đối chiếu so sánh một khối lượng lớn dữ liệu điện tử.

Mặc dù sức hấp dẫn của các thư viện số dữ liệu đã rõ ràng, nhưng nếu không hiểu công việc và mối quan tâm hiện nay của các nhà nghiên cứu thì bất kỳ dự định xây dựng thư viện số dữ liệu nào chưa chắc sẽ thành công. Một lý do thất bại như vậy của các thư viện số không định hướng theo mối quan tâm của nhà nghiên cứu là phương diện kỹ thuật. Một điều thấy rõ từ các công trình nghiên cứu hiện có là các siêu dữ liệu và các yêu cầu kỹ thuật đối với các thư viện số sẽ hoàn toàn mang tính chuyên ngành [8,13,14] và các khía cạnh kỹ thuật của vấn đề đang thực sự là điều thách đố [15]. Một lý do nữa, sự thiếu hiểu biết về công việc và mối quan tâm xoay quanh dữ liệu của các nhà nghiên cứu sẽ dẫn tới thất bại của các thư viện số, được thể hiện rõ nét bởi khó khăn trong việc sắp xếp kho tin của cơ sở. Ở Swinburne, chúng tôi tiến hành một cuộc điều tra cấp cơ sở để hiểu được đầy đủ cách tiếp cận của các nhà nghiên cứu đối với dữ liệu: họ thu thập dữ liệu nào, họ sử dụng những dữ liệu đó như thế nào, họ có chia sẻ dữ liệu hay không và họ cảm thấy như thế nào về quan điểm của cơ sở đối với việc quản trị dữ liệu đó? Trong bài này, chúng tôi trình bày những kết quả của cuộc điều tra và ý nghĩa của chúng đối với bất kỳ dự định tạo lập một thư viện số hoặc kho dữ liệu nào tại cơ sở.

2. Cơ sở lý luận

Ngay trong các ngành khoa học cũng đã dấy lên một số quan ngại về chia sẻ dữ liệu: từ năm 2003, Lyons đã nhận ra

rằng, những quan ngại về quyền sở hữu trí tuệ là một rào cản lớn đối với việc lưu trữ dữ liệu trong một thư viện số [26]; năm 2000, Corti nêu lên ý nghĩa đạo đức của chia sẻ dữ liệu (đặc biệt đối với các dữ liệu nhạy cảm) [27]; năm 2005, Humphrey bày tỏ quan ngại về tác động của loại văn hoá nghiên cứu theo kiểu cá nhân chủ nghĩa trong chia sẻ dữ liệu [28].

Các bài học rút ra từ kinh nghiệm lưu trữ ở cơ sở rất đáng được nghiên cứu. Các kho lưu trữ ở cơ sở có tiềm năng mang lại lợi ích cho cả nhà nghiên cứu thuộc cơ quan lẫn nhà nghiên cứu riêng lẻ [29, 30], tuy nhiên, một khi được thực thi, chúng thường không suôn sẻ [31-33]. Các nhà nghiên cứu, không tham gia trao đổi xuất bản phẩm trước khi cơ quan của họ thành lập kho lưu trữ, đều lưỡng lự gắn bó với các kho này, vì họ coi chúng là gánh nặng, sự thách đố về mặt công nghệ [17], có nguy cơ vi phạm bản quyền và làm mất thời gian nghiên cứu của họ [34]. Ngược lại, những người đã có hệ thống chia sẻ xuất bản phẩm tại chỗ, như những nhà vật lý đã lưu trữ tại địa chỉ Arxiv.org, cũng không quan tâm đến kho lưu trữ, bởi vì họ không thấy có lợi cho mình; mối quan tâm chính của họ là chuyên ngành, hơn là cơ quan [25]. Một số biện pháp đã chứng tỏ có giá trị trong tìm kiếm tài liệu lưu trữ tại các kho lưu trữ ở cơ sở, mang lại lợi ích ngay lập tức và hữu hình cho nhà nghiên cứu dưới dạng các trang tổ chức theo diện nghiên cứu đã thành công [17], vì đã giảm gánh nặng cho nhà nghiên cứu bằng cách sử dụng tài liệu lưu trữ được phương tiện hóa hoàn toàn [20]. Dường như những điều hạn chế đối với lưu trữ xuất bản phẩm cũng sẽ hạn chế lưu trữ dữ liệu, và các cách tiếp cận để giải quyết những vấn đề này phải hướng về nhà nghiên cứu. Có quá nhiều điều chứng tỏ rằng, chỉ nguyên mệnh lệnh không thôi thì chưa có tác dụng [16, 35]. Một khuyến khích khả dĩ hướng nhà nghiên cứu vào lưu trữ dữ liệu trong thư viện số (và bằng cách mở rộng chia sẻ dữ liệu) là tiềm năng trích dẫn tập hợp dữ liệu hệt như đã làm với việc trích dẫn xuất bản phẩm nghiên cứu [26, 36], tuy nhiên, để điều này có ích, việc trích dẫn dữ liệu phải được thừa nhận theo các sơ đồ đánh giá chất lượng nghiên cứu.

Chúng tôi tin tưởng rằng, biết được cách thức mà đông đảo các nhà nghiên cứu khác nhau trong một cơ quan đang thu thập, chia sẻ và quản lý các dữ liệu nghiên cứu của họ là bước đầu tiên tiến tới việc đưa ra chính sách và các công cụ hữu ích xung quanh việc chia sẻ và quản lý dữ liệu ở cấp cơ sở.

3. Phương pháp luận

Bản điều tra của chúng tôi gồm 55 câu hỏi, và được thiết kế sao cho không người trả lời nào phải xem hết 55 câu. Các câu hỏi hiển thị được lựa chọn để tương thích với các câu trả lời đã cho trước đó. Để nắm bắt được rộng rãi hơn kinh nghiệm của các nhà nghiên cứu, ngoài các câu hỏi về lượng, chúng tôi đưa vào một số câu hỏi yêu cầu trả lời bằng văn bản tự do. Để tập trung các câu trả lời của người tham gia vào kinh nghiệm của họ một cách thực tế hơn [39], đối với một vài câu hỏi, chúng tôi đề nghị họ bình luận về một tình tiết cụ thể: những dữ liệu mà họ sử dụng khi viết xuất bản phẩm mới nhất của họ. Bản điều tra được

thực hiện trong vòng một tháng.

Chúng tôi thận trọng xem xét trên diện rộng các định nghĩa của cả dữ liệu lẫn nghiên cứu. Vì thế, chúng tôi đã quảng bá bản điều tra qua "Bản tin nghiên cứu của Swinburne" (Swinburne's Research Bulletin[1], một bản tin trực tuyến phục vụ cán bộ nghiên cứu và sinh viên đã tốt nghiệp), và cho phép những người trả lời tự mình lựa chọn. Chúng tôi gửi thư điện tử nhắc nhở tất cả cán bộ một tuần trước khi cuộc điều tra kết thúc. Để tránh sự thiên lệch trong việc tự lựa chọn người trả lời, chúng tôi đưa ra một phần thưởng cho những người trả lời đã điền đầy đủ vào bản điều tra: cơ may được nhận máy nghe nhạc nén iPod 30 Gb có khả năng xem hình.

Chúng tôi phân tích các câu trả lời cho bản điều tra này, sử dụng những phương pháp thống kê đặc trưng cho các dữ liệu định lượng, và đặt nền phân tích lý thuyết cho các câu trả lời bằng văn bản tự do [40].

4. Kết quả

Kết quả của cuộc điều tra này quy tụ xung quanh năm chủ đề chính:

4.1. Thông tin về con người

Tổng số có 85 người trả lời bản điều tra, hai trong số họ thay mặt cho cả nhóm nghiên cứu. Mặc dù chúng tôi biết rõ có bao nhiêu cán bộ đại học làm việc ở Swinburne, thật khó có thể xác định tỷ lệ trả lời vì không có định nghĩa chắc chắn về người hoạt động nghiên cứu. Ngoài ra, chúng tôi cảm thấy điều quan trọng là phải đưa các sinh viên đã tốt nghiệp vào cuộc điều tra, vì họ thường tạo ra các dữ liệu và sau đó đi tiếp [41], và điều này phù hợp với chính sách dữ liệu của Ôxtrâylia đối với các dữ liệu đó.

4.1.1. Vai trò và kinh nghiệm nghiên cứu

Những người trả lời làm việc với những vai trò rất khác nhau trong lĩnh vực nghiên cứu, bao gồm nghiên cứu sinh sau đại học (thạc sỹ và tiến sỹ), nhà khoa học tiến hành giảng dạy và nghiên cứu (giảng viên, giảng viên cao cấp, giáo sư), thuần túy nghiên cứu (cộng tác và trợ lý nghiên cứu) và các nghề nghiệp khác gồm cả cán bộ tư vấn và quản lý.

Bốn người tham gia, đã liệt kê vai trò nghiên cứu của họ như "vai trò khác", nói rằng họ là nhà quản lý hoặc được một dự án nghiên cứu tuyển dụng.

Thời gian nghiên cứu được phân bổ rất đồng đều, với 14 người trả lời cho biết, họ đã tham gia nghiên cứu dưới hai năm ở một đầu thang bậc, và 10 người - đã làm nghiên cứu hơn 20 năm ở đầu kia. Các nhóm lớn nhất ở trong phạm vi 2-5 năm (31 người) và trong phạm vi 5-10 năm (18 người). Thật ngạc nhiên, không có mối tương liên đáng kể về mặt thống kê giữa chiều dài sự nghiệp và vị trí nghiên cứu.

4.1.2. Lĩnh vực nghiên cứu

Người trả lời điền các lĩnh vực nghiên cứu riêng của họ vào phần văn bản tự do, và các câu trả lời được chia nhỏ như *"Công nghệ sinh học môi trường bền vững [Đúng như nguyên văn]* và rộng như *"Khoa học"*. Chúng tôi sử dụng phép phân tích lý thuyết cơ sở để phân loại các lĩnh vực nghiên cứu này [40] và phát hiện ra rằng, trong số 85 người trả

[1] http://www.research.Swinburne.edu.au/researchers/bulletin/

lời, 27 người nghiên cứu khoa học xã hội, 19 - nghiên cứu khoa học, 11 - nghiên cứu luật và thương mại, 8 - nghiên cứu kỹ thuật, 8 - nghiên cứu công nghệ thông tin, 7 - nghiên cứu giáo dục và 4 - nghiên cứu thiết kế. Những lĩnh vực nghiên cứu này phản ánh khái quát sức mạnh nghiên cứu của Swinburne, vì thế chúng tôi có lý do để có thể đoan chắc rằng, mẫu điều tra này là một nghiên cứu cấp cơ sở ở Swinburne.

4.2. Thu thập dữ liệu và sử dụng

4.2.1. Các loại dữ liệu nào ?

84 người đã trả lời câu hỏi có nhiều lựa chọn về loại dữ liệu nghiên cứu mà họ sử dụng một cách điển hình, và đa số đã liệt kê từ hai trở lên. Sự phân bố các loại dữ liệu được trình bày trong Bảng 1.

Bảng 1. *Các loại dữ liệu được các nhà nghiên cứu ở Swinburne sử dụng*

Loại dữ liệu	Số lượng sử dụng	% tổng số sử dụng	% người trả lời liệt kê loại dữ liệu này
Khảo sát khoa học	23	12,6	27,4
Dữ liệu thực nghiệm	10	5,5	11,9
Dữ liệu sinh học	5	2,7	6
Phân tích dữ liệu	6	3,3	7,1
Đầu ra của chương trìnhtin học	9	4,9	10,7
Các câu trả lời theo Bản điều tra/câu hỏi	31	16,9	36,9
Các tài liệu phỏng vấn/ nhóm mục tiêu	40	21,9	47,6
Tài liệu nguồn cấp 1	19	10,3	22,6
Khảo sát của người tham gia	4	2,2	4,8
Tài liệu công bố	15	8,2	17,9
Bút ký	8	4,4	9,5
Các loại khác	13	7,1	15,5

Các loại dữ liệu mà người trả lời phân loại là "Các loại khác" bao gồm tài liệu video, lý thuyết, thống kê và nghiên cứu trường hợp cụ thể. Phần lớn các nhà nghiên cứu sử dụng kết hợp dữ liệu số hóa và phi số hóa, và một phần đáng kể các dữ liệu được sử dụng (38,6%) ở định dạng phi số hóa.

4.2.2. Chuẩn bị dữ liệu để sử dụng

Mức độ sử dụng dữ liệu số hóa: 50 nhà nghiên cứu (58,8%) chỉ tạo ra dữ liệu số hóa, 14 nhà nghiên cứu (16,5%) và 21 nhà nghiên cứu còn lại (24,7%) tạo ra kết hợp dữ liệu số hóa và phi số hóa.

Các loại xử lý dữ liệu mà những nhà nghiên cứu thường hay tham gia là phiên âm các file tiếng, chuyển định dạng dữ liệu, nhập dữ liệu vào các phần mềm phân tích thống kê, phân tích, chỉnh trang và kiểm tra dữ liệu. Số lượng xử lý được yêu cầu và khả năng áp dụng vào các dữ liệu số hóa và phi số hóa có thể xem trong Bảng 2.

Bảng 2. *Những yêu cầu xử lý dữ liệu*

Định dạng dữ liệu gốc	Tất cả dữ liệu cần xử lý (%)	Không có dữ liệu cần xử lý (%)	Một số dữ liệu cần xử lý (%)
Chỉ có dạng số hóa	20,0	54,0	10,0
Chỉ có dạng phi số hóa	78,6	21,4	0,0
Kết hợp giữa số hóa và phi số hóa	71,4	4,8	23,8
Tỷ lệ phần trăm tổng số người trả lời	51,8	36,5	10,6

Có một mối liên hệ rõ ràng giữa việc tạo ra các dữ liệu phi số hóa và việc phải xử lý các dữ liệu đó. Yêu cầu này không bình thường, thời gian tối thiểu dùng vào việc xử lý mà người trả lời báo cáo là 1-2 giờ, và một vài người trả lời đã phải mất hơn một tháng để xử lý dữ liệu của họ. Những yêu cầu xử lý này là gánh nặng đáng kể đối với nhà nghiên cứu, nhưng họ chấp nhận để có thể nghiên cứu. Bất kỳ một cách tiếp cận nào ở cơ sở đối với việc quản lý dữ liệu đều không thể dựa vào những nhà nghiên cứu quá hào phóng thời gian.

4.2.3. Sử dụng lại dữ liệu

65 trong số 85 nhà nghiên cứu nói rằng, ở một thời điểm nào đó trong quá khứ, họ đã dùng lại những dữ liệu nghiên cứu của riêng họ, chứng tỏ việc sử dụng sau này các dữ liệu nghiên cứu của riêng không chỉ là một tình huống giả thiết.

4.3. Lưu trữ dữ liệu

98% người trả lời giữ lại những dữ liệu mà họ đã sử dụng để viết báo cáo nghiên cứu vừa qua của họ, và chỉ có bốn người không lưu trữ dưới dạng số. Những nơi lưu trữ dữ liệu phổ biến nhất là "A" máy tính (43 người), USB, CD hoặc DVD (12 người), và kho số hóa an toàn (12 người).

4.4. Chia sẻ dữ liệu

4.4.1. Sử dụng lại dữ liệu của người khác

45,3% tổng số người trả lời nói rằng, họ đã sử dụng các tập hợp dữ liệu do người khác tạo lập. Những trường hợp sử dụng lại dữ liệu bao gồm xác nhận kết quả riêng của nhà nghiên cứu, những vấn đề và phạm vi nghiên cứu mới, tạo ra những câu hỏi mới và sử dụng những phương pháp phân tích mới. Ba người trả lời đã cho biết, họ có sử dụng dữ liệu từ một thư viện số dữ liệu. Những trường hợp sử dụng này phản ánh rộng rãi việc dùng lại các dữ liệu của riêng người trả lời như họ đã báo cáo.

4.4.2. Chia sẻ dữ liệu cá nhân

Trong số 85 nhà nghiên cứu, 32 người chưa bao giờ chia sẻ bất kỳ dữ liệu nào, 30 người xác định có dữ liệu được người khác sử dụng lại, và 23 người không chắc (vì dữ liệu của họ nằm trong một tài liệu lưu trữ khuyết danh, và do đó không thể nói rằng nó đã được sử dụng lại). Những người rõ ràng đã chia sẻ dữ liệu, chắc chắn cũng đã sử dụng lại dữ liệu của người khác hơn là những người không chia sẻ (tl=0,007).

Những người trả lời đã chia sẻ các loại dữ liệu ít hơn là họ đã sử dụng; con số trung bình 1,5 loại dữ liệu trên một

nhà nghiên cứu được chia sẻ (so với 2,3 loại dữ liệu được sử dụng). Dữ liệu số hóa được chia sẻ nhiều hơn – chỉ 13,6% dữ liệu chia sẻ là phi số hóa, trong khi 38,6% dữ liệu được tạo lập không ở dạng số hóa (mục 4.2.1).

4.5. Sự tham gia của cơ sở vào dữ liệu nghiên cứu

4.5.1. Nhà nghiên cứu muốn gì từ Swinburne?

Khi được hỏi, Swinburne có thể giúp đỡ gì trong việc quản lý dữ liệu, đa số nhà nghiên cứu (55,64%) nói: Swinburne không giúp gì cho họ được.

Trong số 30 nhà nghiên cứu còn lại thì hình thức giúp đỡ phổ biến nhất là không gian lưu trữ, hoặc là không gian số hóa (10 - 30,3%) hoặc là không gian vật lý (3 - 9,1%). Những câu trả lời phổ biến khác đã liệt kê thêm dịch vụ sao lưu, dịch vụ chuyển đổi dữ liệu và đào tạo quản lý dữ liệu.

4.5.2. Nhà nghiên cứu cảm thấy như thế nào về chính sách dữ liệu của Ôxtrâylia?

Chúng tôi đã đề xuất cho người trả lời về một chính sách, đồng bộ với tài liệu hướng dẫn của Hệ thống dữ liệu quốc gia Ôxtrâylia (ANDS): quy định rằng, tất cả dữ liệu nghiên cứu cần phải được lưu trữ ở một kho lưu trữ dữ liệu số hóa. Khi chúng tôi hỏi những người trả lời suy nghĩ của họ về một chính sách như vậy, kết quả khả quan đến ngạc nhiên; chỉ 13 nhà nghiên cứu tuyên bố họ 'không thích' hoặc 'rất không thích' tuân theo một chính sách như thế. Tuy nhiên, lý do của điều này cũng sớm rõ - đa số các nhà nghiên cứu, mà đã cho biết ý kiến về lý do họ tuân thủ chính sách, đều coi bản thân chính sách này là đúng về lý.

Bảng 3. *Những ưu điểm nhìn thấy về chính sách lưu trữ dữ liệu bắt buộc*

Ưu điểm	số lượng	% tổng số câu trả lời	% người trả lời liệt kê ưu điểm này
Bảo quản dữ liệu	6	7,3	9,3
Sao lưu dữ liệu	9	11,0	13,9
An toàn dữ liệu	9	11,0	13,9
Truy cập dữ liệu của người khác	7	8,5	10,8
Chia sẻ dữ liệu của mình với người khác	20	24,4	30,9
Chuyển dịch văn hoá hướng về chia sẻ và cộng đồng	7	8,5	10,8
Tính công khai rõ ràng	3	3,7	4,6
Sử dụng tốt các nguồn dữ liệu	4	4,9	6,2
Bằng chứng/đối chiếu kết quả	9	11,0	13,9
"Không có lợi gì cho tôi"	5	6,1	7,7
Ưu điểm khác	3	3,7	4,6

Bảng 4. *Những quan ngại về chính sách lưu trữ dữ liệu bắt buộc*

Quan ngại	số lượng	% tổng số	% người trả lời liệt kê câu trả lời quan ngại này
Tính bảo mật/đạo đức	24	34,3	32,4
Lạm dụng hoặc lấy cắp dữ liệu	22	31,4	29,7
Quan ngại công trình của họ có được công nhận	2	2,9	2,7
Mất thời gian tải lên mạng	9	12,9	12,2
Chi phí lưu trữ	5	7,1	6,8
Tính khả dụng của kho lưu trữ	3	4,3	4,1
Độ tin cậy của kho lưu trữ	5	7,1	6,8
Sự quan liêu	5	7,1	6,8
Sự trùng lặp của các dịch vụ ngành hiện có	2	2,9	2,7
"Không có quan ngại gì"	2	2,9	2,7
Quan ngại khác	12	17,1	16,2

Bảng 5. *Những hạn chế về dữ liệu được lưu trữ trong một thư viện số dữ liệu ở cơ sở*

Hạn chế	số lượng	% tổng số	% người trả lời câu trả lời liệt kê hạn chế này
Kiểm soát ai đã sử dụng dữ liệu	27	33,8	36,0
Giao ước đạo đức được giữ vững	15	18,8	20,0
Sự công nhận	12	15,0	16,0
Bảo vệ quyền sở hữu trí tuệ	10	12,5	13,3
Khai báo khi dữ liệu được sử dụng	6	7,5	8,0
Biết được dữ liệu sẽ được sử dụng như thế nào	5	6,3	6,7
An toàn dữ liệu	5	6,3	6,7
Truy cập tới chính dữ liệu	5	6,3	6,7
Phần phi thương mại như trong thông lệ sáng tác	3	3,8	4,0
Không có hạn chế	3	3,8	4,0
Những hạn chế khác	9	11,3	12,0

Khi được hỏi về những ưu điểm khả dĩ của một chính sách như thế, những nhà nghiên cứu đã tìm được một số ưu điểm (Bảng 3), nhưng họ cũng có một số quan ngại (Bảng 4). Hơn nữa, họ muốn đặt một số giới hạn về dữ liệu trong một thư viện như thế (Bảng 5). Điều này phản ánh những phát hiện về những kho lưu trữ ở cơ sở [42], nhấn mạnh sự song hành giữa xuất bản phẩm và dữ liệu.

Nhiều nhà nghiên cứu đã liệt kê nhiều quan ngại và hạn chế hơn là ưu điểm. Họ thấy được viễn cảnh của một chính sách hoàn toàn có nguy cơ bắt buộc họ phải chia sẻ dữ liệu của mình, và những đe dọa này nặng ký hơn cả những ưu điểm mà họ có thể tìm thấy trong một đề xuất như vậy.

5. Thảo luận

Các thư viện số có nhiều cái để cống hiến trong một tương lai chia sẻ dữ liệu. Nhiều ngành khoa học và nhà nghiên cứu đã chia sẻ dữ liệu, và nhiều người nữa đã bày tỏ mong muốn làm như vậy khi có sự ủng hộ [10, 21, 22, 43]. Các thư viện số tạo cơ hội tiết kiệm rất nhiều chi phí thu thập dữ liệu và giúp vượt qua khoảng cách và cả thời gian [7, 9, 44-46]. Các thư viện số đặc biệt rất thích hợp với các dữ liệu không gắn với bất kỳ sự nhậy cảm nào về đạo đức và dễ dàng được mô tả theo cách mà cộng đồng sử dụng dữ liệu có thể hiểu được. Các dữ liệu có xu hướng đáp ứng những tiêu chuẩn này (và thường được sử dụng trong các thí dụ về thư viện số dữ liệu) là những dữ liệu khoa học.

Mặc dù có sự hứa hẹn cố hữu trong các thư viện số, chúng không giải quyết được mọi vấn đề thông tin [47], và trên thực tế theo nghĩa tổng quát, bị nhiều người trong giới học thuật coi là còn đang bộn bề và có vấn đề [48]; dường như chúng không thể đáp ứng hoàn toàn cho các dữ liệu. Mặc dù những người trả lời cuộc điều tra của chúng tôi tuyên bố rằng, họ muốn gửi dữ liệu trong một thư viện số dữ liệu ở cơ sở, thì kinh nghiệm của kho lưu trữ dữ liệu ở cơ sở cho thấy, các chỉ thị "đó là nguyên tắc", đơn giản không phải là yếu tố động viên đối với các nhà nghiên cứu [16, 35]. Tương tự như vậy, cuộc điều tra cũng cho thấy có sự quan ngại liên quan đến việc ký gửi trong thư viện số dữ liệu; khi các nhà nghiên cứu quan ngại về việc ký gửi tài liệu trong một kho lưu trữ cơ sở thì họ sẽ từ chối làm như thế [17, 35]. Tính khả dụng của bất kỳ thư viện số dữ liệu nào được xem như là mối quan ngại của nhiều người trả lời chúng tôi; qua sách báo, chúng tôi biết rằng, các nhà nghiên cứu đang mong muốn "giảm bớt hỗn loạn" [49], và rằng, các thư viện số đã bị nhiều người trong giới học thuật coi là không khả dụng [48]. Rõ ràng là, để mô tả đầy đủ, các siêu dữ liệu trong thư viện số phải hoàn toàn chính xác; những người trả lời cuộc điều tra của chúng tôi quan tâm đến thời gian tải dữ liệu lên mạng cho thư viện số, và kinh nghiệm đã qua cho thấy, các nhà nghiên cứu bị bối rối khi được yêu cầu tạo lập siêu dữ liệu [50] và phải phấn đấu để làm như vậy [21]. Tiêu chuẩn về tính mở vốn có trong kiến trúc của nhiều hệ thống thư viện số có lẽ là mối quan ngại đối với những người trả lời của chúng tôi, họ bị phiền hà về những câu hỏi liên quan đến quyền sở hữu trí tuệ và đạo đức. Nhiều dữ liệu do những người trả lời tạo lập (38,6%) không phải lúc sinh ra đã ở dạng số hóa, và những người trả lời mô

tả quá trình số hóa là vất vả và chậm chạp (một phát hiện được phản ánh trong một tài liệu khác) [12]; chúng tôi có thể cho rằng, chưa chắc họ đã muốn số hóa dữ liệu chỉ để lưu trữ trong một thư viện số dữ liệu.

Để cách tiếp cận quản lý dữ liệu ở cơ sở được thành công, trước hết nó phải hỗ trợ quản lý dữ liệu và cung cấp dịch vụ mà nhà nghiên cứu thấy là hữu ích: hiện nay ở Swinburne, các dịch vụ được nhà nghiên cứu thấy hữu ích nhất là các dịch vụ số và không gian lưu trữ vật lý dành cho dữ liệu của họ, các dịch vụ sao lưu số và sự giúp đỡ số hóa. Đây là toàn bộ các dịch vụ có thể cung cấp rộng rãi mà cách tiếp cận thư viện số dữ liệu đưa ra cho nhà nghiên cứu (và như đã nói trong mục 4.5, các nhà nghiên cứu thấy thư viện số dữ liệu còn nhiều vấn đề hơn là các ưu điểm).

6. Kết luận và công việc trong tương lai

Các thư viện số dữ liệu có nhiều hứa hẹn, đặc biệt cho những nhà nghiên cứu làm việc với các dữ liệu không bị trở ngại vì các ràng buộc đạo đức và với những từ mô tả siêu dữ liệu (metadata descriptors) trong sáng, rõ ràng (thường là các dữ liệu khoa học). Chúng tôi thấy rằng, các chính sách quản lý dữ liệu không được sự ủng hộ và tán đồng chung của các nhà nghiên cứu. Sự thiếu nhiệt tình không phải là do lưỡng lự chia sẻ dữ liệu (đa số các nhà nghiên cứu đều muốn chia sẻ), mà do mong muốn về việc kiểm soát các dữ liệu nghiên cứu tế nhị hơn là thấy một chính sách hoàn chỉnh với thư viện số ở cơ sở. Các nhà nghiên cứu không những muốn kiểm soát tế nhị các dữ liệu nghiên cứu của họ, mà họ còn tạo ra một số lượng đáng kể các dữ liệu nghiên cứu không ở dạng số. Từ kinh nghiệm kho lưu trữ ở cơ sở, rõ ràng là các nhà nghiên cứu chỉ ủng hộ các dịch vụ có lợi cho họ. Với những hạn chế này, các thư viện số dữ liệu rõ ràng không thể đáp ứng đầy đủ nghĩa vụ của các cơ sở về dữ liệu.

Mặc dù có thái độ dè dặt đối với các thư viện số dữ liệu, các nhà nghiên cứu thấy được giá trị nào đó ở sự giúp đỡ của cơ sở trong việc quản lý dữ liệu. Toàn bộ các dịch vụ tàng trữ, sao lưu và chuyển đổi dữ liệu được một số nhà nghiên cứu coi là hữu ích trong cuộc điều tra của chúng tôi và có thể là một cách giúp cho các cơ sở tạo dựng năng lực quản lý dữ liệu nghiên cứu. Tương tự như vậy, một số nhà nghiên cứu đã quen với việc lưu trữ các tập hợp dữ liệu, và có lẽ họ sẽ không chỉ tuân thủ mà còn có khả năng quản lý dữ liệu của họ trong một thư viện số. Như vậy, các cơ sở có lẽ nên xem xét tạo ra các thư viện số dữ liệu cho dù chúng không phải là giải pháp toàn bộ.

Làm thế nào để thực thi một chính sách dữ liệu đầy đủ có lợi cho mọi nhà nghiên cứu vẫn còn là một vấn đề để ngỏ cho nghiên cứu, cũng như vậy đối với một giải pháp kỹ thuật tốt nhất đáp ứng

các giao ước của cơ sở về dữ liệu. Các thư viện số dữ liệu rõ ràng là một phần của giải pháp này, mặc dù một số dữ liệu sẽ cần phải số hóa, một số sẽ yêu cầu những chuẩn siêu dữ liệu mới và một số chỉ yêu cầu sự truy cập tinh tế hơn so với khả năng cung cấp hiện nay của các thư viện số. Một điều còn quan trọng hơn cả cách thức hoạt động của các thư viện số, đó là hiểu sâu hơn về công tác dữ liệu thực tế và những mối quan tâm của tất cả các nhà nghiên cứu, bởi vì không có sự quan tâm này, cả chính sách lẫn giải pháp kỹ thuật đều không thể đáp ứng nhu cầu của nhà nghiên cứu, và các hệ thống định hướng nghiên cứu mà không đáp ứng nhu cầu của nhà nghiên cứu đều chắc chắn thất bại.

Vũ Văn Sơn *(Dịch)*

Nguồn: *"the Role of Digital Libraries in a Time of Global Change", pp. 236-249*

Tài liệu tham khảo

1. Corti, L. Progress and Problems of Preserving and Finding Access to Qualitative Data for Social Research – the International Picture of Emerging Culture. Forum: Qualitative Sozialforschung/ Forum: Qualitative Social Research [Online journal] 1 (2000)

2. Hey, T., Trefethen, A.: The Data Deluge: An E-Science Perspective. In: Berman, F., Fox, G.C., Hey A.J.G.(eds.) Grid Computing: Making the Global Infrastructure a Reality, pp. 809-824. John Wiley and Sons, Ltd., Chichester (2003)

3. Arzberger, P., Schroeder, P., Beaulieu, A., Casey, K., Laaksonen, L., Moorman, D., Uhlir, P., Wouters, P., : Promoting Access to Public Research Data for Scientific, Economic and Social Development. Data Science Journal 3, 135-152 (2004)

4. Heery, R.,Duke, M., Day, M., Lyon, L., Hursthouse, M.B., Frey, J.G., Cole, S.,J., Gutteridge, C., Carr, L.: Integrating Research Data into Public Workflow. The Ebank UK Perspective. PV 2004 Ensuring the Long-term Preservation and Adding Value to the Scientific and Technical Data. European Space Agency, Frascati, Italy (2004)

5. Zhuge, H.: China,s E-Science Knowledge Grid Environment. Intelligent Systems. IEEE 19, 13-17 (2004)

6. National Science Foundation Cyberinfrastructure Council : Cyberinfrastructure vision for 21st Century Discovery: National Science Foundation, Arlington, VA (2007)

7. The ANDS Technical Working Group: Towards an Australian Data Commons, Australian National Data Service, Canberra, Australia (2007)

8. Shadbolt, A., van der Knijff, D., Young, E., Winton, L.: Sustainable Paths for Data Intensive Research Communities at the University of Melbourne: A Report for Australian Partnership for Sustainable Repositories. University of Melbourne, Melbourne (2006)

9. Borgman, C., Wallis, J., Enyedi, N.: Building Digital Libraries for Scientific Data: An Exploratory Study of Data Practices in Habitat Ecology. Research and Advanced Technology for Digital Libraries, 170-183 (2006)

10. Karasti, H., Baker, K.S.: Digital Data Practices and the Long-Term Ecological Research Program Growing Global, International Journal of Digital Curation 3, 42-58 (2008)

11. ANDS: Research Data Policy and the "Australian Code for the Responsible Conduct of Research". Australian National Data Service, Canberra, Australia (2009)

12. Henry, M.: Dreaming of Data: The Library's Role in Supporting E-Research and Data Management. In: Australian Library and Information Association Biennial Conference. Australian Library and Information Association, Alice Springs (2008)