Tạo bộ dữ liệu nghiên cứu sẵn sàng từ đăng ký hành chính quốc gia

Päivi Kankaanranta and Aino Melakari***

Tóm tắt:

Để cải thiện khả năng truy cập vào đăng ký quốc gia và dữ liệu hành chính cho mục đích nghiên cứu, Cơ quan Thống kê Phần Lan tạo ra các mô-đun dữ liệu nghiên cứu tạo sẵn mà ít tốn công để phổ biến hơn so với các tập dữ liệu được điều chỉnh theo yêu cầu của các nhà nghiên cứu. Trái ngược với các cuộc khảo sát mẫu do nhiều cơ quan thống kê quốc gia cũng như Eurostat cung cấp sẵn cho các nhà nghiên cứu, các mô-đun tạo sẵn này chứa một tập hợp các biến tiêu chuẩn với thông tin ở cấp độ kinh doanh cá thể, hộ gia đình hoặc cá nhân được trích xuất từ sổ đăng ký hành chính cơ bản. Việc sử dụng và sản xuất các bộ dữ liệu tạo sẵn được quy định bởi Đạo luật Thống kê Phần Lan và các đạo luật thống kê của Liên minh Châu Âu. Do đó, dữ liệu được đặt biệt hiệu để tất cả các số nhận dạng trực tiếp đều bị xóa khỏi dữ liệu. Để có thể cung cấp dữ liệu vi mô với thông tin chi tiết cao, các mô-đun tạo sẵn bao gồm một số biến giới hạn liên quan đến các chủ đề cụ thể. Các nhà nghiên cứu chỉ sử dụng các mô-đun có liên quan đến lĩnh vực nghiên cứu của họ, điều này cho phép Cơ quan Thống kê Phần Lan đáp ứng các yêu cầu của GDPR đối với việc giảm thiểu dữ liệu. Tất cả các mô-đun dữ liệu tạo sẵn có thể được liên kết với nhau và với dữ liệu vi mô từ các nguồn khác với các mã nhận dạng giả riêng lẻ để xây dựng tập dữ liệu phong phú. Tập hợp các biến bao gồm trong mỗi mô-đun được mô tả trong danh mục siêu dữ liệu trên trang web Cơ quan Thống kê Phần Lan theo cách cho phép các nhà nghiên cứu xác định các tệp dữ liệu phù hợp với mục đích của họ. Các mô-đun dữ liệu tạo sẵn chỉ có thể được sử dụng thông qua hệ thống truy cập từ xa an toàn và mỗi tệp đầu ra đều được kiểm tra tính bảo mật. Các mô-đun dữ liệu tạo sẵn tạo thành một trong những chức năng cơ bản của Dịch vụ Nghiên cứu tại Cơ quan Thống kê Phần Lan. Gần 70% tất cả các dự án nghiên cứu đang sử dụng bộ dữ liệu tạo sẵn. Có một nhóm riêng biệt, những người duy trì các bộ dữ liệu này trên cơ sở toàn thời gian. Cùng với các mô-đun dữ liệu được tạo sẵn và hệ thống truy cập từ xa đảm bảo rằng việc sử dụng dữ liệu dựa trên đăng ký cấp đơn vị cho mục đích nghiên cứu tuân thủ các chính sách bảo mật và bảo vệ dữ liệu thống kê.

* *Cơ quan Thống kê Phần Lan, paivi.kankaanranta@stat.fi*

** *Cơ quan Thống kê Phần Lan, aino.melakari@stat.fi*

1. Giới thiệu

Một trong những mục tiêu chiến lược của Cơ quan Thống kê Phần Lan là cải thiện khả năng truy cập vào đăng ký quốc gia và các nguồn dữ liệu hành chính khác cho các mục đích nghiên cứu khoa học và thống kê. Hồ sơ hành chính được sử dụng rộng rãi để đưa ra số liệu thống kê chính thức ở Phần Lan (xem Cơ quan Thống kê Phần Lan, 2004). Do đó, phần lớn dữ liệu lưu trữ tại Cơ quan Thống kê Phần Lan bao gồm dữ liệu dựa trên đăng ký cũng được các nhà nghiên cứu quan tâm. Vì việc chỉnh sửa tập dữ liệu nghiên cứu từ nguồn lưu trữ trong cơ quan thống kê và các nguồn bên ngoài theo yêu cầu của nhà nghiên cứu là rất khó, nên Dịch vụ nghiên cứu tại Thống kê Phần Lan phát hành các mô-đun dữ liệu tạo sẵn. Các mô-đun dữ liệu tạo sẵn này được cung cấp cho các nhà nghiên cứu trên cơ sở "nguyên trạng": chúng chứa một tập hợp các biến tiêu chuẩn với thông tin cấp đơn vị về tất cả các doanh nghiệp, hộ gia đình và cá nhân được trích xuất từ hồ sơ hành chính mà nhiều cơ quan chức năng duy trì. Thông thường, các cơ quan thống kê quốc gia và các nhà cung cấp dữ liệu khác, chẳng hạn như Eurostat, cung cấp quyền truy cập vào dữ liệu vi mô bao gồm thông tin về một mẫu người trả lời khảo sát thay vì toàn bộ tổng thể mục tiêu.

Các mô-đun dữ liệu nghiên cứu được tạo sẵn có hiệu quả về thời gian và chi phí để sản xuất. Một thư mục tệp duy nhất được lưu trữ trong môi trường an toàn là đủ để đáp ứng nhu cầu của tất cả người dùng của một mô-đun cụ thể. Dữ liệu bao gồm một mô-đun tạo sẵn chỉ được biên dịch một lần và được lưu vào một thư mục chuyên dụng mà người dùng được ủy quyền có thể truy cập được. Do đó, không cần sao chép tệp dữ liệu mỗi khi thông tin trong một mô-đun cụ thể được yêu cầu hoặc cập nhật.

Thống kê Phần Lan được truyền cảm hứng để phát triển một thiết kế mô-đun cung cấp cho các nhà nghiên cứu khả năng truy cập dễ dàng nhưng linh hoạt vào dữ liệu vi mô bí mật bằng kinh nghiệm của họ với Dữ liệu nhân viên-nhân viên theo chiều dọc của Phần Lan (FLEED) và nhiều mô-đun dữ liệu kinh doanh khác nhau. Bộ dữ liệu này kéo dài hơn 28 năm (từ 1988 đến 2016) bao gồm tất cả những người từ 15 đến 70 tuổi sống ở lục địa Phần Lan. Hơn 160 biến dựa trên đăng ký trong bộ dữ liệu FLEED được sử dụng rộng rãi chứa thông tin về các đặc điểm cơ bản của cá nhân, hộ gia đình, điều kiện sống, việc làm, các mối quan hệ, thời gian thất nghiệp, thu nhập và giáo dục. Để đáp ứng các yêu cầu cấp quốc gia và EU về bảo vệ dữ liệu, tập dữ liệu lớn này sau đó đã được chia thành các tập con nhỏ hơn theo chủ đề. Do đó, các nhà nghiên cứu đã bị ngăn cản việc tiếp cận thông tin không cần thiết cho các mục đích cụ thể của họ. Các tập con của tập dữ liệu FLEED trước đây hiện được cung cấp dưới dạng các mô-đun dữ liệu nghiên cứu được tạo sẵn của FOLK.

2. Mô-đun dữ liệu nghiên cứu được tạo sẵn tại Cơ quan Thống kê Phần Lan

2.1. Các yếu tố cơ bản của mô-đun dữ liệu nghiên cứu tạo sẵn

Cơ quan Thống kê Phần Lan thiết kế các mô-đun dữ liệu nghiên cứu tạo sẵn trên cơ sở nhu cầu của các nhà nghiên cứu. Các hồ sơ hành chính được yêu cầu thường xuyên nhất về nhân khẩu học, xã hội và doanh nghiệp

được chuyển thành các mô-đun dữ liệu tạo sẵn. Phản hồi trực tiếp từ cộng đồng nghiên cứu là điều tối quan trọng trong việc xác định dữ liệu tương ứng với nhu cầu hiện tại và tương lai của nhà nghiên cứu. Một số nhà nghiên cứu chủ yếu làm việc trong các tổ chức khác giữ vị trí bán thời gian tại Dịch vụ Nghiên cứu tại Cơ quan Thống kê Phần Lan. Cùng với các nhà khoa học và chuyên gia dữ liệu riêng của Cơ quan Thống kê Phần Lan, những nhà nghiên cứu này đóng góp kiến thức chuyên môn theo chủ đề cụ thể của họ vào việc phát triển các mô-đun dữ liệu nghiên cứu đã được tạo sẵn mới. Họ cũng đóng vai trò là cầu nối giữa Cơ quan Thống kê Phần Lan với tư cách là nhà sản xuất dữ liệu và cộng đồng nghiên cứu với tư cách là người sử dụng dữ liệu. Ngoài ra, Cơ quan Thống kê Phần Lan khám phá sở thích của các nhà nghiên cứu bằng cách thực hiện khảo sát khách hàng và phân tích xu hướng trong bộ dữ liệu nghiên cứu được thiết kế riêng. Gần đây, các nhà nghiên cứu ngày càng yêu cầu quyền truy cập vào dữ liệu thô càng tốt và càng nhanh càng tốt sau khi phát hành.

Mỗi mô-đun tạo sẵn do Cơ quan Thống kê Phần Lan phát hành thuộc một bộ sưu tập theo chủ đề và xoay quanh một chủ đề cụ thể trong chủ đề đó. Các chủ đề chung bao gồm dân số, giáo dục, hoạt động kinh doanh, thu nhập, thị trường lao động, thương mại quốc tế, giao thông, nhà ở cũng như tiền lương và tiền công. Để kiểm soát quyền truy cập của người dùng vào dữ liệu vi mô bí mật, mỗi mô-đun bao gồm một tập hợp các biến tiêu chuẩn có số lượng bị hạn chế. Mặc dù số lượng biến là một tập hợp con của dữ liệu nguồn, nhưng tất cả các đơn vị thống kê ban đầu đều được chứa trong các mô-đun dữ liệu được tạo sẵn. Do đó, các mô-đun đại diện cho toàn bộ tập hợp trong nguồn hành chính.

Tất cả các mô-đun dữ liệu tạo sẵn đều có tên bao gồm tiền tố và nhãn. Tiền tố đề cập đến chủ đề và nhãn mô tả chủ đề của mô-đun chi tiết hơn. Các tên mô-đun này được sử dụng rộng rãi trong cộng đồng nghiên cứu. Hiện tại, Cơ quan Thống kê Phần Lan có tổng cộng 56 mô-đun dữ liệu nghiên cứu tạo sẵn được cung cấp. Bảng 1 cung cấp tóm tắt về các mô-đun này.

Danh mục siêu dữ liệu trên trang web Cơ quan Thống kê Phần Lan, được gọi là Taika, bao gồm tài liệu toàn diện mô tả các mô-đun dữ liệu được tạo sẵn và các biến trong đó (xem Thống kê Phần Lan, Taika - danh mục dữ liệu nghiên cứu). Tài liệu có sẵn bằng tiếng Phần Lan và trong một số trường hợp bằng tiếng Anh, ít nhất là ở định dạng viết tắt. Dữ liệu và các mô tả biến có thể được duyệt riêng và được truy vấn cho các cụm từ tìm kiếm và / hoặc các thuộc tính nhất định. Do đó, tài liệu dữ liệu có thể được điều chỉnh cho phù hợp với nhu cầu của một người và được tải xuống ở định dạng tệp do một người lựa chọn. Danh mục siêu dữ liệu không chỉ hữu ích cho người dùng hiện tại của các mô-đun tạo sẵn mà còn cho những người dùng trong tương lai. Danh mục cho phép các nhà nghiên cứu làm quen với các mô-đun tạo sẵn và xác định những mô-đun phù hợp với mục đích của họ trước khi họ nộp đơn xin cấp phép sử dụng dữ liệu thống kê.

Bảng 1. Thống kê Các mô-đun dữ liệu sẵn sàng cho nghiên cứu hiện có của Phần Lan

Tiền tố của bộ dữ liệu	Chủ đề mô-đun của bộ dữ liệu	Các chủ đề đã chọn của mô-đun (Nguồn dữ liệu)	Tổng thể mục tiêu	Số mô-đun
FOLK	Dân số	Thu nhập Giáo dục Gia đình Bằng cấp Sống cùng nhau	Cá nhân	12
EDUC	Giáo dục	Tuyển sinh Học sinh, sinh viên Bằng cấp	Cá nhân	3
FIRM	Hoạt động kinh doanh	Hoạt động kinh doanh Báo cáo tài chính	Doanh nghiệp	23
FLOWN	Sở hữu	Thông tin cổ đông Thông tin cổ tức	Doanh nghiệp, cá nhân	1
SES	Thu nhập	Cơ cấu thống kê thu nhập Cấu trúc hài hòa của thống kê thu nhập	Cá nhân	1
TAX	Thu nhập	Tiền lương, lương hưu và trợ cấp (Sổ đăng ký thu nhập quốc gia)	Cá nhân, doanh nghiệp	3
INFRA	Dữ liệu không gian	Dữ liệu không gian của cư dân	Cá nhân, tòa nhà	2
Dữ liệu từ các tổ chức khác	Thị trường lao động Thương mại quốc tế Giáo dục Vận tải	Số liệu thống kê về dịch vụ việc làm (Bộ Kinh tế và Việc làm) Thương mại quốc tế về hàng hóa (Hải quan Phần Lan) Kết quả xét tuyển (Hội đồng xét tuyển)	Cá nhân, doanh nghiệp	11

2.2. Liên kết các mô-đun dữ liệu tạo sẵn

Các mô-đun dữ liệu nghiên cứu được tạo sẵn có thể được sử dụng để xây dựng các bộ dữ liệu phong phú tương ứng với các nhu cầu cụ thể của một dự án nghiên cứu. Các cơ quan chức năng khác nhau ở Phần Lan sử dụng mã nhận dạng thống nhất cho cá nhân, doanh nghiệp và tổ chức cũng như các tòa

nhà và nhà ở. Do đó, các mô-đun dữ liệu tạo sẵn có thể được liên kết với nhau bằng các mã nhận dạng giả bắt nguồn từ các mã nhận dạng trực tiếp này. Loại liên kết dữ liệu này không khả thi với dữ liệu vi mô dựa trên khảo sát ngoại trừ các nghiên cứu theo thời gian. Một khi dữ liệu được hài hòa, các mã nhận dạng thống nhất cho phép các cá nhân và doanh nghiệp được theo dõi trong suốt vòng đời của chúng. Sự hài hòa dữ liệu thường bao gồm việc đặt tên các biến một cách nhất quán theo thời gian và làm cho các phân loại được sử dụng phổ biến nhất có thể so sánh được.

Mã nhận dạng thống nhất giúp có thể kết hợp các mô-đun tạo sẵn ngay cả với dữ liệu cấp đơn vị từ các nguồn nằm ngoài tầm kiểm soát của Cơ quan Thống kê Phần Lan, chẳng hạn như dữ liệu về việc sử dụng các dịch vụ chăm sóc sức khỏe hoặc mua thuốc theo toa. Những dữ liệu này được gửi đến Cơ quan Thống kê Phần Lan để lấy bút danh trước khi cung cấp cho các nhà nghiên cứu. Dữ liệu khảo sát do Cơ quan Thống kê Phần Lan hoặc chính các nhà nghiên cứu thu thập có thể được tích hợp với dữ liệu đăng ký cấp đơn vị ở một mức độ hạn chế. Người trả lời khảo sát cần được thông báo trước khi phỏng vấn hồ sơ hành chính nào sẽ được kết hợp với thông tin họ cung cấp cho cuộc khảo sát (xem ví dụ: Törmälehto 2008).

2.3 Ứng dụng để sử dụng các mô-đun dữ liệu tạo sẵn

Các mô-đun dữ liệu tạo sẵn do Cơ quan Thống kê Phần Lan phát hành cung cấp cho các nhà nghiên cứu khả năng truy cập tương đối nhanh chóng vào dữ liệu vi mô bí mật. Vì các mô-đun này được cung cấp ở định dạng chuẩn và các tệp dữ liệu luôn sẵn có trong môi trường an toàn, nên việc xử lý đơn xin cấp phép sẽ mất ít thời gian hơn so với dữ liệu nghiên cứu được thiết kế riêng. Theo thông lệ hiện tại, giấy phép bao gồm tất cả các bản cập nhật trong tương lai cho các mô-đun dữ liệu tạo sẵn được chỉ định trong đó miễn là giấy phép còn hiệu lực (tối đa năm năm với tùy chọn gia hạn). Do đó, các nhà nghiên cứu không cần phải đăng ký riêng giấy phép mới để có quyền truy cập vào bản phát hành mới nhất cho các mô-đun dữ liệu theo ý của họ. Tuy nhiên, đơn xin cấp phép mới được yêu cầu bất cứ khi nào các mô-đun tạo sẵn mới hoặc dữ liệu khác được thêm vào một dự án nghiên cứu hiện có.

Bởi vì các mô-đun dữ liệu tạo sẵn chứa thông tin về tổng số mẫu của tổng thể mục tiêu, các đơn xin cấp phép và các kế hoạch nghiên cứu đi kèm được sàng lọc cẩn thận. Ứng dụng phải chỉ định các mô-đun dữ liệu tạo sẵn được yêu cầu và các dữ liệu khác cùng với mục đích của dự án nghiên cứu. Khi quyền truy cập vào dữ liệu vi mô chứa thông tin về người được yêu cầu, ứng dụng phải cung cấp lý do cho nhu cầu sử dụng thông tin về tổng dân số. Các lý do đủ tiêu chuẩn bao gồm ví dụ như theo dõi các gia đình theo thời gian hoặc sản xuất các nhóm đối chứng cho dân số nghiên cứu. Nếu việc sử dụng dữ liệu tổng thể là không hợp lý, thay vào đó, một mẫu tương ứng với nhu cầu của thiết kế nghiên cứu sẽ được cung cấp.

3. Các biện pháp bảo vệ dữ liệu và bảo mật dữ liệu

3.1. Các điều kiện tiên quyết chung để phát hành dữ liệu tạo sẵn

Các mô-đun dữ liệu nghiên cứu được tạo sẵn cung cấp sự thỏa hiệp giữa khả năng truy

cập dễ dàng vào dữ liệu vi mô bí mật và việc tuân thủ khuôn khổ pháp lý về xử lý thông tin cấp đơn vị. Các điều kiện để sản xuất và sử dụng các mô-đun dữ liệu tạo sẵn được xác định trong Quy định chung về bảo vệ dữ liệu (GDPR) và Đạo luật bảo vệ dữ liệu quốc gia bổ sung (1050/2018) cũng như trong Đạo luật thống kê quốc gia (280/2004) . Các thỏa thuận pháp lý này xác định thông tin cấp độ đơn vị chi tiết cao có trong đăng ký hành chính có thể được phát hành và cách thức truy cập vi dữ liệu này (xem Thống kê Phần Lan 2013 và UNECE 2007 để biết các mô tả trước đây về luật pháp quốc gia về phát hành vi dữ liệu ở Phần Lan).

Vì bộ dữ liệu tạo sẵn bao gồm các bộ biến tiêu chuẩn được tạo sẵn mà không cần chỉnh sửa, nội dung thông tin của mỗi mô-đun được xem xét kỹ lưỡng trong giai đoạn lập kế hoạch để ngăn các nhà nghiên cứu tiếp cận với thông tin không cần thiết. Đề xuất về mô-đun dữ liệu hoàn thiện sẵn sàng được đệ trình để đánh giá cho ủy ban đạo đức thống kê của Cơ quan Thống kê Phần Lan, bao gồm đại diện của các bộ phận khác nhau. Quyết định cuối cùng về việc phát hành một mô-đun dữ liệu nghiên cứu đã sẵn sàng mới do các giám đốc phụ trách đưa ra.

Để đáp ứng các yêu cầu về bảo mật dữ liệu và bảo vệ dữ liệu, các số nhận dạng trực tiếp được loại trừ khỏi các mô-đun dữ liệu nghiên cứu được tạo sẵn. Các mã nhận dạng duy nhất, chẳng hạn như mã nhận dạng cá nhân, mã nhận dạng doanh nghiệp cũng như mã cư trú cho các tòa nhà và nhà ở, được thay thế bằng mã nhận dạng giả. Bởi vì các mô-đun dữ liệu tạo sẵn chứa một số lượng biến hạn chế, tập dữ liệu cuối cùng được liên kết cùng với các số nhận dạng giả này tuân theo nhu cầu GDPR để giảm thiểu dữ liệu: các nhà nghiên cứu chỉ sử dụng các mô-đun có liên quan cho dự án của họ. Mặc dù việc đặt bút danh đảm bảo rằng các chủ thể dữ liệu của các mô-đun không thể được xác định trực tiếp, chúng vẫn có thể được xác định một cách gián tiếp. Do đó, dữ liệu tạo sẵn chỉ có thể được sử dụng thông qua hệ thống truy cập từ xa, nơi có thể kiểm soát quá trình xử lý dữ liệu.

Để bảo vệ hơn nữa quyền riêng tư của các chủ thể dữ liệu, thông tin nhạy cảm sẽ bị xóa khỏi các mô-đun dữ liệu được tạo sẵn. Giống như thông tin đặc biệt, dữ liệu cá nhân thuộc một số danh mục đặc biệt nhất định cũng bị loại trừ khỏi các mô-đun này. Việc xử lý thông tin thuộc các loại này thường bị cấm. Dữ liệu như vậy tiết lộ nguồn gốc chủng tộc và dân tộc, ý kiến chính trị, tôn giáo hoặc niềm tin triết học, tư cách thành viên công đoàn cũng như khuynh hướng hoặc hoạt động tình dục. Các danh mục dữ liệu cá nhân đặc biệt cũng bao gồm thông tin liên quan đến sức khỏe cũng như dữ liệu di truyền và sinh trắc học trực tiếp nhận dạng người đó. Cơ quan Thống kê Phần Lan không duy trì hồ sơ của những dữ liệu này. Tuy nhiên, dữ liệu sức khỏe cũng như di truyền và sinh trắc học có sẵn thông qua các cơ quan chức năng khác và có thể được xử lý và liên kết với bộ dữ liệu tạo sẵn của Cơ quan Thống kê Phần Lan nếu GDPR, luật Châu Âu hoặc luật pháp quốc gia đưa ra một ngoại lệ đối với lệnh cấm.

3.2. Giải pháp an toàn để sử dụng dữ liệu được tạo sẵn

Vì các chủ thể dữ liệu có thể được xác định gián tiếp từ dữ liệu vi mô bí mật, các mô-đun tạo sẵn chỉ có thể truy cập được

thông qua hệ thống truy cập từ xa FIONA mà Cơ quan Thống kê Phần Lan đã phát triển cùng với công ty nhà nước Trung tâm Khoa học CNTT. Hệ thống này cung cấp một cách an toàn để xử lý dữ liệu nhạy cảm. FIONA là một môi trường khép kín không có kết nối internet bên ngoài và các nhà nghiên cứu không có khả năng lấy ra hoặc tải lên bất kỳ dữ liệu hoặc bất kỳ tệp nào vào hệ thống. Chỉ có thể sử dụng FIONA trong khu vực EU và các nước thứ ba mà ủy ban EU đã cho phép chuyển dữ liệu vi mô và từ các địa chỉ IP cấp tổ chức mà Cơ quan Thống kê Phần Lan đã chấp nhận là địa chỉ an toàn. Trong các tình huống không thể truy cập FIONA, các nhà nghiên cứu có thể đến thăm phòng thí nghiệm nghiên cứu tại các cơ sở của CƠ quan Thống kê Phần Lan ở Helsinki.

Trước khi các nhà nghiên cứu có thể có quyền truy cập vào hệ thống truy cập từ xa và dữ liệu đã tạo sẵn, họ phải điền vào một cam kết truy cập từ xa mà họ cam kết với các điều khoản và điều kiện của Dịch vụ Nghiên cứu Thống kê Phần Lan. Trong cam kết, ví dụ như báo cáo các địa chỉ và mô tả các biện pháp bảo mật dữ liệu của không gian làm việc là bắt buộc. Cam kết được ký bởi đại diện từ tổ chức của nhà nghiên cứu, người chịu trách nhiệm về việc sử dụng quyền truy cập từ xa của nhà nghiên cứu. Trong trường hợp sử dụng sai mục đích, Cơ quan Thống kê Phần Lan được phép ngắt truy nhập vào hệ thống truy cập từ xa.

3.3. Nguyên tắc của người dùng để bảo vệ dữ liệu bí mật

Các nhà nghiên cứu sử dụng dữ liệu được đặt bút danh phải có nghĩa vụ giữ bí mật và phải cam kết không tiết lộ cho những người khác ngoài những thông tin được đề cập trong giấy phép về bất kỳ thông tin nào mà họ có nghĩa vụ giữ bí mật. Các nhà nghiên cứu cũng phải đảm bảo rằng kết quả đầu ra không tiết lộ thông tin cấp đơn vị hoặc chứa thông tin cho phép xác định chủ thể dữ liệu.

Có một số hướng dẫn cho các loại tập dữ liệu và kết quả đầu ra khác nhau. Quy tắc chính trong việc bảo vệ các bảng tần số và cường độ là giá trị ngưỡng là ba, có nghĩa là đầu ra chỉ có thể được xuất bản nếu nó dựa trên ít nhất ba quan sát (với một vài ngoại lệ, chẳng hạn như giá trị ngưỡng là mười cho sẵn sàng -tạo dữ liệu dựa trên Đăng ký thu nhập). Ngoài ra còn có các quy tắc khác ngoài giá trị ngưỡng này, chẳng hạn như quy tắc thống trị (1.75), được sử dụng với dữ liệu doanh nghiệp cập nhật nhất. Về các bộ dữ liệu nhạy cảm nhất, có nhiều quy tắc bảo vệ dữ liệu cụ thể hơn cần được xem xét. Ví dụ: các giá trị ngưỡng nhất định áp dụng cho các đầu ra được xuất bản từ dữ liệu không gian. Khi mô-đun tạo sẵn bao gồm dữ liệu được cơ quan có thẩm quyền khác cho phép, Cơ quan Thống kê Phần Lan tuân theo các quy tắc bảo vệ dữ liệu của tổ chức đó.

Các nhà nghiên cứu được yêu cầu tuân theo các yêu cầu bảo vệ dữ liệu khi làm việc trong hệ thống truy cập từ xa. Tất cả việc truyền dữ liệu giữa FIONA và các nhà nghiên cứu đều thông qua một giao thức kiểm tra đầu ra, được thực hiện thủ công bởi nhân viên của Dịch vụ Nghiên cứu. Trước khi kết quả đầu ra có thể được gửi đến các nhà nghiên cứu, chúng sẽ được kiểm tra xem có khả năng vi phạm bí mật hay không. Vì quá trình đánh giá tiết lộ sử dụng các nguồn lực của Thống kê Phần Lan và dẫn đến sự chậm trễ cho các nhà nghiên cứu, các tùy chọn để

tự động hóa một phần giao thức hiện đang được khám phá.

4. Chi phí sản xuất và sử dụng dữ liệu nghiên cứu tạo sẵn

Cơ quan Thống kê Phần Lan tính phí cho cả việc xử lý đơn xin cấp phép sử dụng dữ liệu thống kê và dữ liệu nghiên cứu của chính nó. Các khoản phí bổ sung được thu cho việc sử dụng hệ thống truy cập từ xa hàng năm. Việc xử lý đơn xin cấp phép có tốc độ cố định và bao gồm một giờ làm việc chuẩn bị. Những giờ làm việc chuẩn bị bổ sung được tính theo giờ. Phí xử lý được thu ngay cả khi đơn đăng ký không được chấp thuận.

Thiết kế mô-đun là một cách hiệu quả về thời gian và chi phí để cung cấp cho các nhà nghiên cứu quyền truy cập vào dữ liệu vi mô. Hiệu quả này được phản ánh ở mức giá tương đối rẻ được tính cho các mô-đun dữ liệu tạo sẵn. Bởi vì mô-đun dữ liệu tạo sẵn chỉ được phát hành một lần ở định dạng tiêu chuẩn, chi phí sản xuất và bảo trì được phân bổ đồng đều cho tất cả người dùng được ủy quyền của mô-đun, cả những người hiện tại và trong tương lai. Mặt khác, chi phí biên soạn một bộ dữ liệu được thiết kế riêng chỉ do dự án nghiên cứu cần dữ liệu tùy chỉnh phát sinh.

Phương pháp sản xuất cũng cho phép Cơ quan Thống kê Phần Lan ấn định trước giá của các mô-đun dữ liệu tạo sẵn, điều này trái ngược với các tập dữ liệu nghiên cứu được thiết kế riêng. Do đó, chi phí dữ liệu cho một dự án nghiên cứu chủ yếu dựa vào các mô-đun dữ liệu tạo sẵn là khá dễ đoán. Hiện tại, phí dữ liệu chỉ được thu một lần, sau khi giấy phép người dùng đã được cấp và bao gồm các bản cập nhật dữ liệu trong tương lai. Hơn nữa, phí dữ liệu bao gồm hướng dẫn về việc sử dụng dữ liệu nghiên cứu đã tạo sẵn và bất kỳ bảo trì dữ liệu bổ sung nào. Hiện tại, có bốn loại giá khác nhau đại khái dựa trên kích thước của mô-đun dữ liệu và tần suất cập nhật. Mặt khác, chi phí kiểm tra đầu ra được chi trả bởi phí hàng năm tính cho việc sử dụng hệ thống truy cập từ xa. Phí xử lý đơn xin cấp phép, dữ liệu và sử dụng hệ thống truy cập từ xa có thể thay đổi định kỳ.

5. Kết luận

Để đáp ứng yêu cầu của các nhà nghiên cứu, Cơ quan Thống kê Phần Lan phát hành các mô-đun dữ liệu nghiên cứu được tạo sẵn từ nhiều chủ đề khác nhau. Các mô-đun cung cấp khả năng truy cập tương đối dễ dàng nhưng linh hoạt vào dữ liệu vi mô hành chính cho các mục đích được quy định trong luật pháp quốc gia, đó là các nghiên cứu khoa học và điều tra thống kê về điều kiện xã hội. Chính sách phát hành dữ liệu này không ảnh hưởng đến tính bảo mật của dữ liệu. Nó cũng hiệu quả về chi phí và thời gian cho cả Cơ quan Thống kê Phần Lan với tư cách là nhà cung cấp dữ liệu và các nhà nghiên cứu với tư cách là người sử dụng dữ liệu. Các tệp dữ liệu cấu thành một mô-đun chỉ được hình thành một lần và được lưu trữ trong một môi trường an toàn, nơi việc xử lý dữ liệu cấp đơn vị có thể được kiểm soát một cách khá dễ dàng.

Dữ liệu tạo sẵn là nguồn dữ liệu đáng chú ý cho các nghiên cứu tập trung vào các cá nhân và doanh nghiệp. Hơn một trăm dự án nghiên cứu và bốn trăm nhà nghiên cứu hiện đang sử dụng dữ liệu tạo sẵn trong các nghiên cứu của họ. Vào năm 2020, Cơ quan Thống kê Phần Lan đã xử lý gần trăm ứng dụng liên quan đến các mô-đun tạo sẵn và số lượng các loại yêu cầu dữ liệu này đang tăng

lên. Hầu hết các dự án đều có quyền truy cập vào dữ liệu tổng dân số và dữ liệu tạo sẵn được sử dụng thường xuyên nhất là các mô-đun FOLK chứa nhiều loại dữ liệu liên quan đến dân số.

Dữ liệu tạo sẵn có thể được mong đợi sẽ trở thành cốt lõi của các dịch vụ dữ liệu vi mô do Cơ quan Thống kê Phần Lan cung cấp cho các nhà nghiên cứu. Để có thể đáp ứng nhu cầu gia tăng, nhiều nguồn lực hơn đã được phân bổ để duy trì các mô-đun đã sẵn sàng hiện tại và phát triển các mô-đun mới. Kết quả là, một nhóm riêng biệt với bảy thành viên chịu trách nhiệm về các hoạt động này đã được thành lập. Vì các dự án nghiên cứu đã mở rộng mục đích của chúng sang các hiện tượng mới, nên số lượng các mô-đun tạo sẵn được dự đoán sẽ tăng lên trong tương lai. Theo xu hướng hiện nay, các dự án nghiên cứu được kỳ vọng sẽ trở nên lớn hơn về số lượng nhà nghiên cứu cũng như số lượng mô-đun tạo sẵn mà họ muốn truy cập.

Theo phản hồi, các nhà nghiên cứu đánh giá cao các bản cập nhật thường xuyên mà họ có quyền truy cập tự động mà không cần quy trình ứng dụng. Họ cũng đánh giá rằng các mô-đun tạo sẵn phù hợp với nhu cầu của các loại nhà nghiên cứu khác nhau, cho dù họ là người mới bắt đầu hay những người đã có kinh nghiệm. Mặt khác, một số nhà nghiên cứu có kinh nghiệm cho rằng một số mô-đun nhất định đã quá xử lý và thay vào đó muốn có quyền truy cập vào dữ liệu thô. Trong những trường hợp này, có thể áp dụng cho dữ liệu được điều chỉnh bổ sung cho các mô-đun được tạo sẵn. Cơ quan Thống kê Phần Lan đặt mục tiêu tạo ra các mô-đun mới với nhiều thông tin chi tiết hơn để thỏa mãn những nhu cầu này. Ví dụ, trong năm qua, Cơ quan Thống kê Phần Lan đã giới thiệu một số mô-đun mới với ít dữ liệu được xử lý về thu nhập tiền lương và phúc lợi được trích xuất từ Đăng ký thu nhập. Các mô-đun mới này đã đóng một vai trò quan trọng trong các nghiên cứu gần đây về hậu quả kinh tế của đại dịch COVID-19.

Tài liệu tham khảo

1. Statistics Finland. Taika – research data catalogue. Available at <https://taika.stat.fi/en/>.

2. Statistics Finland (2013). "Development and challenges of on-line micro-data usage. United Nations Economic Commission for Europe Conference of European Statisticians, Geneva.

3. Statistics Finland (2004). *Use of Registers and Administrative Data Sources for Statistical Purposes. Best Practices of Statistics Finland*. Handbooks 45. Helsinki.

4. Törmälehto, V.-M. (2008). "Social statistics – integrated use of survey and administrative data at Statistics Finland." International Association for Official Statistics Conference on Reshaping Official Statistics, Shanghai. Available at <https://www.iaos-isi.org/papers/CS_26_3_Tehto.pdf>.

5. UNECE (2007). *Managing Statistical Confidentiality & Microdata Access. Principles and Guidelines of Good Practice.* Available at <https://unece.org/fileadmin/DAM/stats/publications/Managing.statistical.confidentiality.and.microdata.access.pdf>.

Đỗ Ngát (dịch)

Nguồn: https://unece.org/sites/default/files/2021-12/SDC2021_Day1_Kankaanranta_AD.pdf