

# QUẢN LÝ RỦI RO VÀ CÔNG BỐ DỮ LIỆU VI MÔ: CÂN BẰNG RỦI RO TIẾT LỘ VÀ SỬ DỤNG DỮ LIỆU

*Sonia Whiteley & Eric Skuja*  
*Trung tâm Nghiên cứu Xã hội, Úc*

## **Tóm tắt**

Quản lý hiệu quả dữ liệu vi mô chấp nhận một sự thật là không có “sự cố định một chiều” sẽ giải quyết một cách thấu đáo tất cả các rủi ro tiềm ẩn đi kèm với công bố thông tin ghi chép cơ sở. Đó cũng là trường hợp điển hình mà các phương pháp khác nhau được sử dụng để tối đa hóa việc bảo mật dữ liệu có những rủi ro khác, và tất cả các phương pháp bảo mật dữ liệu sẽ hạn chế tính hữu dụng của dữ liệu cơ bản [1]. Chúng tôi đã phát triển một phương pháp để quản lý và công bố dữ liệu sử dụng dạng cân bằng có lưu tâm đến việc nhận dạng lại tiềm năng hoặc những mối đe dọa rõ ràng và ưu tiên việc sử dụng dữ liệu đồng thời giải quyết tích cực các trường hợp rủi ro để xảy ra hơn.

Phương pháp tiếp cận cân bằng quản lý rủi ro này sẽ được bàn luận liên quan đến cuộc Tổng Điều tra Phát triển trẻ thơ Úc (Australian Early Development Census-AEDC) được tiến hành định kỳ ba năm một lần để đo lường sự phát triển của trẻ khi chúng bắt đầu bước vào năm học chính thức đầu tiên. Dữ liệu AEDC được thu thập thông qua hình thức bảng kiểm (checklist) trực tuyến do giáo viên làm để đo lường 5 lĩnh vực phát triển của trẻ em. Các vấn đề liên quan đến rủi ro và cộng đồng nghiên cứu, các chiến lược giảm thiểu rủi ro quản lý dữ liệu, các phương thức truy cập dữ liệu vi mô và những tác động của việc chia sẻ rủi ro giữa các học giả và những người quản lý dữ liệu sẽ được khám phá.

**Từ khóa:** Quản lý rủi ro, sử dụng dữ liệu, dữ liệu điều tra, tổng điều tra, giáo dục mầm non, quản lý dữ liệu, bảo mật, dữ liệu vi mô

## **1. Điều tra Phát triển trẻ thơ Úc (AEDC)**

### **1.1. Về AEDC**

Ba năm một lần, các giáo viên hoàn thành một bảng kiểm cho mọi trẻ em Úc hiện đang theo học năm đầu tiên của chương trình toàn thời gian. Có khoảng 100 câu hỏi trong bảng kiểm bao gồm năm lĩnh vực lý thuyết về sự của phát triển trẻ em:

- Sức khỏe thể chất và tinh thần;
- Năng lực xã hội;

- Độ chín tình cảm;

- Kỹ năng ngôn ngữ và nhận thức (dựa trên cơ sở nhà trường);

- Các kỹ năng giao tiếp và kiến thức tổng quát.

Tổng điều tra Phát triển trẻ thơ Úc được thực hiện ở cấp quốc gia lần đầu tiên vào năm 2009, lần thứ hai vào năm 2012. Các công việc chuẩn bị đã được tiến hành cho việc thu thập dữ liệu AEDC 2015. Khoảng 290.000 bảng kiểm đã được hoàn

thành ở mỗi cuộc điều tra, tương đương với trên 96% trẻ thuộc phạm vi điều tra. Có ba hệ thống trường học tại Úc là hệ thống trường học của Chính phủ, tổ chức Công giáo và tổ chức Độc lập, cả ba hệ thống trường học này đều tích cực tham gia vào các cuộc điều tra AEDC.

Dữ liệu điều tra AEDC được thu thập trực tuyến, dựa trên hiểu biết và quan sát của giáo viên với trẻ trong lớp. Mỗi bảng kiểm mất ít hơn 20 phút để hoàn thành và kinh phí tài trợ cho giáo viên có sẵn cho tất cả các trường tham gia.

### **1.2. AEDC và quản lý dữ liệu**

Chính phủ Úc luôn cam kết tạo điều kiện truy cập dữ liệu AEDC cho các mục đích lập chính sách, kế hoạch hóa và nghiên cứu. Một giao thức dữ liệu và một chính sách kết nối dữ liệu đã có ngay từ khi bắt đầu thu thập dữ liệu để đưa ra hướng dẫn sử dụng hợp lý dữ liệu.

Để phù hợp với phương pháp quản lý rủi ro hiện do Cục Thống kê Úc (ABS) khuyến cáo liên quan đến công bố dữ liệu vi mô, một bộ sưu tập các File ghi chép cơ sở bảo mật (Confidentialised Unit Record Files - CURF) đã được tạo ra cho năm 2009. Do có được đặc điểm phân tách địa lý chi tiết, tập tin được tách thành một CURF nghiên cứu và một CURF địa lý. Dữ liệu đã bị xáo trộn trong mỗi tập tin do một nhân viên hợp đồng thực hiện, tuy nhiên bản chất và mức độ chính xác của sự xáo trộn không được biết. Các trường chính liên quan đến sự hiểu biết các vấn đề phát triển của trẻ, như giới tính, đã thay đổi.

Trong giai đoạn sản xuất dữ liệu của lần thu thập năm 2012, rõ ràng là các CURF không nhất thiết phải phù hợp với mục đích. Các cơ quan chính phủ đã sử dụng các CURF AEDC cho mục đích lập chính sách và kế hoạch hóa, mặc dù dữ liệu không

còn nguyên vẹn. Các nhà nghiên cứu đã công bố các phát hiện từ các dự án của họ là các ấn phẩm AEDC chính thức có mâu thuẫn, và những khác biệt này có thể quy cho là do sử dụng các CURF. Để giải quyết vấn đề này, và những vấn đề khác liên quan, một cách tiếp cận khác để quản lý rủi ro dữ liệu vi mô đã được khám phá.

## **2. Rủi ro và cộng đồng nghiên cứu**

### **2.1. Cách tiếp cận truyền thống để quản lý rủi ro**

Cách tiếp cận truyền thống để quản lý rủi ro được liên kết với công bố dữ liệu vi mô có xu hướng tập trung vào "tình huống xấu nhất có thể xảy ra", nơi các chủ sở hữu hoặc người quản lý dữ liệu chịu trách nhiệm chính đối với việc xác định và giảm thiểu mọi nguy cơ tiềm ẩn liên quan đến an toàn, sự riêng tư và bảo mật dữ liệu [2]. Mô hình này được củng cố bằng các giả định rằng người sử dụng dữ liệu về cơ bản không có chuyên môn, không thể tin được, được đào tạo về dữ liệu không đầy đủ và, trong những trường hợp đặc biệt, có ý định lợi dụng sử dụng sai dữ liệu. Ví dụ, các kịch bản rủi ro xấu nhất được xem xét bởi Văn phòng Thống kê Vương quốc Anh bao gồm các cuộc tấn công chính trị, liên quan tới các bộ dữ liệu cá nhân, các nhà báo, và những người hàng xóm tò mò [3].

Để giải quyết các mối đe dọa rõ ràng tiềm ẩn được trình bày bởi những tình huống này, những người chủ hoặc người quản lý dữ liệu được yêu cầu phải "bảo vệ dữ liệu" từ những người sử dụng dữ liệu và đảm bảo rằng danh tính của một cá nhân hay một thuộc tính của một nhóm không thể được xác định dù cố ý hoặc vô ý. Phương pháp kịch bản xấu nhất của quản lý dữ liệu không giành ưu tiên cho sự hữu dụng hay tính chính xác của dữ liệu ghi chép cơ sở. Những câu trả lời cá nhân được bảo mật bằng cách

bỏ hoặc thay đổi thông tin ban đầu, và các quy tắc được sử dụng để thay đổi dữ liệu không được tiết lộ để ngăn chặn dữ liệu nguồn được tái tạo thông qua thiết bị đối chiếu.

Phương pháp quản lý rủi ro dữ liệu vi mô này có thể tạo ra cảm giác an toàn giả như, trong hầu hết các trường hợp, không thể bảo mật hoàn toàn một file ghi chép cơ sở. Sử dụng bảo mật như là trung tâm của một chiến lược quản lý rủi ro có thể dẫn đến sự tự mãn về phía các nhà quản lý dữ liệu và người sử dụng dữ liệu khi có niềm tin rằng chính dữ liệu vi mô đang được bảo vệ một cách thực chất. Bất kỳ một sự tự mãn nào đều có thể dẫn đến việc giảm sự tập trung đối với việc sử dụng dữ liệu thích hợp và các biện pháp bảo mật dữ liệu cần thiết.

### ***2.2. Những phương pháp khác để quản lý rủi ro***

Một phương pháp khác để quản lý rủi ro là chia sẻ trách nhiệm về sử dụng và báo cáo dữ liệu vi mô thích hợp với cộng đồng nghiên cứu [4]. Giả định cơ bản của phương pháp này là các thành viên của cộng đồng nghiên cứu không có dụng ý chủ động sử dụng sai dữ liệu vi mô mà họ yêu cầu. Kinh nghiệm quốc tế không cung cấp một bằng chứng nào cho thấy các nhà nghiên cứu cố tình lạm dụng dữ liệu bí mật để xác định và tiết lộ thông tin cá nhân [2].

Trong khi việc sử dụng sai dữ liệu vi mô có chủ ý của các nhà nghiên cứu hình như không phải là trường hợp rủi ro chủ chốt, thì có những chỉ dẫn rõ ràng những thực tế đó liên quan đến xử lý, lưu trữ và xuất bản dữ liệu không đáp ứng được các điều khoản cụ thể của thỏa thuận quyền sử dụng. Các ví dụ về các hành vi không phù hợp liên quan đến dữ liệu vi mô ẩn danh, gồm có:

- Sử dụng dữ liệu vi mô cho một dự án hoặc mục đích chưa được phê duyệt;

- Cho phép những người dùng tin chưa được phép truy cập dữ liệu;

- Lưu trữ dữ liệu vi mô trong môi trường không an toàn như lưu giữ trong USB;

- Không áp dụng các quy tắc kiểm soát công khai (bỏ các ô có số liệu nhỏ) trong các xuất bản phẩm.

Những hành vi này tạo nên các "tình huống rủi ro thực tế" mà về thực chất có nhiều khả năng hơn tình trạng sử dụng sai dữ liệu ác ý. Có khả năng là các thành viên của cộng đồng nghiên cứu tin rằng không có một nguyên nhân nào về sự quan tâm đến những hành vi này ít chuyên môn hơn bởi vì, là các nhà nghiên cứu, họ thực sự đáng tin cậy và, trong trường hợp các file đã bảo mật, thì dữ liệu vi mô đã được bảo vệ nếu có gì đó sai.

### **3. Các chiến lược giảm thiểu rủi ro quản lý dữ liệu**

Phương pháp đề xuất nhằm giảm thiểu rủi ro liên quan đến tiếp cận dữ liệu vi mô AEDC được xây dựng trên cơ sở các thủ tục và hướng dẫn hiện hành. Tiêu điểm là về sử dụng các chiến lược đa cấp thông qua việc quản lý dữ liệu và các nhóm nghiên cứu bên ngoài đồng thời cung cấp các mức hỗ trợ thích hợp nhằm ngăn chặn và quản lý bất kỳ tình huống rủi ro tiềm ẩn nào. Các chiến lược giảm thiểu rủi ro liên quan gồm có:

- Đánh giá tất cả các dự án để khẳng định mục tiêu nghiên cứu chính;

- Đánh giá tất cả các dự án để đảm bảo đúng theo các yêu cầu về sử dụng, lưu trữ và xuất bản;

- Hạn chế quyền truy cập đối với những người sử dụng dữ liệu được phép và thích hợp;

- Cung cấp siêu dữ liệu chi tiết;

- Nặc danh hóa dữ liệu vi mô;

- Kiểm tra kết quả dữ liệu vi mô (dạng biểu)

được sản xuất từ dữ liệu vi mô;

- Cung cấp các dịch vụ hỗ trợ người sử dụng dữ liệu và duy trì một cộng đồng nghiên cứu được tham gia;

- Cung cấp đánh giá rủi ro định lượng của file dữ liệu cho các nhà nghiên cứu.

Mỗi chiến lược giảm thiểu rủi ro này được trình bày trong phần dưới đây.

### ***3.1. Đánh giá các dự án để khẳng định mục tiêu nghiên cứu chính***

Tất cả các yêu cầu truy cập liên quan đến công bố dữ liệu vi mô được đánh giá để khẳng định rằng mục đích chính của dự án là thực hiện hoặc hỗ trợ nghiên cứu xã hội hoặc nghiên cứu chính sách. Mục tiêu nghiên cứu chính có thể được đưa ra từ một phạm vi/ ứng dụng hoặc một phạm vi nghiên cứu thuần túy. Các mục đích nghiên cứu không nhất thiết phải mới hoặc thống nhất với mục đích chính.

Kinh nghiệm quốc tế cho thấy đánh giá dự án trước khi công bố dữ liệu vi mô thường được thực hiện thông qua đánh giá của đồng nghiệp hoặc ủy ban. Các khuyến nghị liên quan đến việc công bố dữ liệu vi mô thường được phê duyệt ở cấp tương đương với Thông kê trưởng hoặc phó (giống như Tổng cục trưởng hoặc Phó Tổng cục trưởng thông kê). Một ủy ban thích hợp là phù hợp cho mục đích này.

### ***3.2. Hạn chế quyền truy cập đối với những người sử dụng dữ liệu phù hợp***

Những người sử dụng dữ liệu phù hợp có mục đích nghiên cứu chính, có các công cụ cần thiết và được đào tạo để làm việc với dữ liệu vi mô phức tạp. Người sử dụng dữ liệu không sẵn lòng hoặc không chứng tỏ được họ là thành viên của cộng đồng nghiên cứu hoặc tại sao dự án đề xuất có liên quan với nhiệm vụ ở nơi làm việc của họ lại không được

xem là những người sử dụng dữ liệu thích hợp. Người sử dụng dữ liệu dường như không có các công cụ hoặc được tập huấn cần thiết để phân tích dữ liệu vi mô. Ở những nơi mà người sử dụng dữ liệu không cho thấy có đủ các công cụ hoặc được đào tạo cần thiết để phân tích dữ liệu vi mô, thì sự bổ sung các chương trình phân tích hoặc phát triển kỹ năng cần thiết là một điều kiện của việc công bố dữ liệu. Sinh viên đại học hoặc sau đại học phải hỏi xin sử dụng dữ liệu vi mô AEDC cùng với người hướng dẫn của họ.

### ***3.3. Đánh giá dự án đảm bảo đúng các yêu cầu về sử dụng, lưu trữ và xuất bản***

Tất cả các yêu cầu truy cập có liên quan đến công bố dữ liệu vi mô cần cung cấp thông tin chi tiết về tại sao người sử dụng dữ liệu phải tuân thủ các quy định liên quan đến việc sử dụng, lưu trữ và xuất bản dữ liệu vi mô. Hình như là người sử dụng dữ liệu có thể không chú ý đến các phương pháp xử lý dữ liệu thực tế tốt nhất, cái gì tạo nên sự vi phạm các yêu cầu liên quan và các kết quả của sự vi phạm đó cần được làm rõ.

Thông tin này có trong một số tài liệu, Hướng dẫn sử dụng dữ liệu và Hướng dẫn hỏi xin dữ liệu vi mô cũng như hợp đồng bản quyền, tuy nhiên rất khó để xác định xem các nhà nghiên cứu đọc và hiểu được tầm quan trọng của những yêu cầu này. Các hình thức truy cập dữ liệu hiện nay cho phép người sử dụng cho biết họ phân nào thể nào với các quy định sử dụng, lưu trữ và xuất bản. Điều gì tạo nên hành vi không phù hợp hoặc vi phạm quy định liên quan đến dữ liệu vi mô AEDC cũng cần được người sử dụng dữ liệu nhận biết trước khi sản xuất file dữ liệu.

### ***3.4. Cung cấp siêu dữ liệu chi tiết***

Siêu dữ liệu thường được mô tả như là "dữ liệu về dữ liệu". Đó là những thông tin mô tả được các

thành viên của cộng đồng nghiên cứu sử dụng để hiểu được tất cả các thành phần của dữ liệu từ các mục đích bao quát của việc thu thập, đến quá trình để có được dữ liệu, thông qua việc tạo ra tập dữ liệu. Người sử dụng dữ liệu tiềm năng cần được biết về những hạn chế có thể của việc thu thập và cách thức

mà dữ liệu được sử dụng hoặc phân tích trước đây để hỗ trợ các vấn đề nghiên cứu chất lượng cao, yêu cầu truy cập dữ liệu và kết quả dữ liệu. Có một số loại siêu dữ liệu khác nhau được cung cấp cho người sử dụng dữ liệu, xem tại bảng 1.

**Bảng 1:** Các loại siêu dữ liệu và thực trạng về tài liệu hỗ trợ

Loại siêu dữ liệu	Phạm vi	Tài liệu hỗ trợ
Siêu dữ liệu theo ngữ cảnh	Thông tin về mục đích thu thập dữ liệu và sự cam kết về thu thập hiện nay của cộng đồng nghiên cứu	Giao thức dữ liệu AEDC Hướng dẫn sử dụng dữ liệu AEDC
Paradata	Thông tin định tính và định lượng về quá trình thu thập dữ liệu  Thông tin về kinh nghiệm làm việc với dữ liệu của người sử dụng	Báo cáo kỹ thuật AEDC  Thông tin về kinh nghiệm người sử dụng hiện không có
Siêu dữ liệu chất lượng	Đánh giá về phạm vi, tính đầy đủ và độ chính xác của dữ liệu	Hướng dẫn sử dụng dữ liệu AEDC Báo cáo kỹ thuật AEDC
Siêu dữ liệu khái niệm	Mô tả chi tiết các yếu tố dữ liệu hiện có	Từ điển dữ liệu AEDC
Siêu dữ liệu cấu trúc	Thông tin về cấu hình file và quan hệ giữa các file dữ liệu	Giao thức dữ liệu AEDC Hướng dẫn sử dụng dữ liệu AEDC
Siêu dữ liệu truy cập	Dữ liệu được truy cập như thế nào, vào lúc nào và những ai là người dùng hợp pháp	Giao thức dữ liệu AEDC Hướng dẫn sử dụng dữ liệu AEDC

Cung cấp tài liệu hỗ trợ chi tiết cho người sử dụng dữ liệu làm tối đa hóa khả năng rằng yêu cầu truy cập dữ liệu sẽ chính xác và tập trung [5]. Người sử dụng dữ liệu hiện đang được tiếp cận với thông tin về những hạn chế của dữ liệu và cần phải có đầy đủ thuộc tính về mục đích của dữ liệu và quá trình thu thập để hỗ trợ việc giải thích chính xác các kết quả đầu ra.

### **3.5. Nặc danh dữ liệu vi mô**

Nặc danh dữ liệu vi mô liên quan đến việc loại bỏ tất cả các yếu tố dữ liệu có thể xác định rõ ràng một cá nhân. Trong trường hợp dữ liệu vi mô AEDC, tên đầy đủ, địa chỉ và các tọa độ địa lý chứa dữ liệu có thể nhận biết một cách rõ ràng một người. Có một số biên số khác đã được xem xét kết hợp để nhận biết một bản ghi duy nhất trong một bộ dữ liệu.



Để bảo toàn tính chính xác và chất lượng của dữ liệu vi mô, chỉ có những thay đổi nhỏ về dữ liệu được đưa ra, bao gồm:

- Kết hợp các biến nhân khẩu học có số lượng nhỏ các quan sát trong một số phân tổ trả lời nào đó (ví dụ như quốc gia nơi sinh được ghép thành Úc và Khác);

- Loại bỏ biến dữ liệu Cộng đồng địa phương. Biến này cung cấp thông tin địa phương được đặt mục tiêu cao và rất có thể dẫn đến xác định vô ý các cá nhân. Khi biến này không phù hợp với hệ thống phân tổ địa lý được thừa nhận, thì điều không chắc là chúng sẽ được người sử dụng dữ liệu yêu cầu.

Căn cứ vào phạm vi an toàn được đưa ra và thấy người sử dụng dữ liệu chính không yêu cầu dữ liệu vi mô với ý định tiết lộ danh tính của các cá nhân, thì không cần bảo mật dữ liệu vi mô.

### ***3.6. Kiểm tra dữ liệu vi mô (đầu ra dạng biểu) được sản xuất từ dữ liệu vi mô***

Nếu dữ liệu vi mô được sử dụng để sản xuất dữ liệu vi mô, thông tin dạng biểu cụ thể hơn, vẫn có khả năng tạo ra các ô nhỏ làm lộ các đặc điểm của 3 hoặc ít hơn 3 trẻ, hoặc đặc điểm của các nhóm mặc dù có sự nặc danh hoặc bảo mật nào đó có thể xảy ra. Nếu đầu ra chính của dữ liệu vi mô AEDC là dữ liệu vi mô, thì dữ liệu vi mô là sản phẩm chính để đáp ứng yêu cầu hơn là công bố dữ liệu vi mô ẩn danh. Điều này đảm bảo rằng toàn bộ những kiểm tra không để lộ cần thiết đã được thực hiện trên dữ liệu vi mô trước khi công bố.

Trong những trường hợp mà dữ liệu vi mô là kết quả phân tích thứ cấp, thì người sử dụng dữ liệu có được sự xác nhận từ các nhà quản lý dữ liệu rằng đã tuân theo tất cả các quy tắc giữ kín liên quan. Các biểu số liệu được các nhà quản lý xem và ký tất

trước khi tài liệu liên quan được hoàn tất để xuất bản.

### ***3.7. Cung cấp các dịch vụ hỗ trợ sử dụng dữ liệu và duy trì cộng đồng nghiên cứu tham gia***

Thực tế quốc tế đã chứng tỏ rằng tối thiểu hóa các rào cản và giải thích rõ tại sao những hạn chế thực tế hoặc hạn chế rõ ràng tồn tại liên quan đến việc sử dụng, lưu trữ và phổ biến dữ liệu vi mô có khả năng quản lý hiệu quả rủi ro. Một số thực tế yếu kém được các nhà nghiên cứu chỉ ra có thể xuất phát từ nhận thức rằng người quản lý dữ liệu đang cố ngăn chặn việc truy cập hơn là khuyến khích sử dụng thích hợp. Sự liên lạc thường xuyên với người dùng dữ liệu AEDC thông qua cảnh báo thư điện tử, cập nhật, hội thảo và hội nghị giúp khuyến khích "các hành vi dữ liệu" phù hợp và hữu ích, và có thể tạo ra cảm giác về kết nối không nằm trong phạm vi cung cấp bộ dữ liệu. Khuyến khích người sử dụng dữ liệu tìm kiếm sự hỗ trợ từ các nhà quản lý dữ liệu và đồng cấp bản quyền nếu họ yêu cầu sự hỗ trợ chuyên môn có thể giảm thiểu rủi ro khi dữ liệu được sử dụng hoặc giải thích không phù hợp.

### ***3.8. Cung cấp đánh giá rủi ro định lượng của các tập dữ liệu cấp cho các nhà nghiên cứu***

Một báo cáo đánh giá rủi ro chính thức được tạo ra cho mỗi bộ dữ liệu có sẵn cho các nhà nghiên cứu. Bằng việc tập trung vào hai lĩnh vực quan tâm, (1) mức độ mà các nhà nghiên cứu có thể kết nối AEDC với các bộ dữ liệu khác và (2) nguy cơ về các ô kích thước nhỏ được xuất bản, ta có thể định lượng các quan tâm này nhờ xem xét trước khi quyết định công bố dữ liệu. Ví dụ sau minh họa một quá trình đánh giá rủi ro hai giai đoạn điển hình.

(1) Trong phần thứ nhất của đánh giá rủi ro, một nhà nghiên cứu đã yêu cầu có bộ dữ liệu AEDC rất lớn được chứa 250 biến của 560.000 trẻ em qua hai chu

kỳ điều tra AEDC - cơ bản gần hết dữ liệu AEDC. Trong số 250 biên AEDI, chỉ có 12 biên được cho là có khả năng có trong bộ dữ liệu hành chính mà các nhà nghiên cứu có thể truy cập. Mười hai biên đó bao gồm các định danh có thể như tuổi, giới tính, dân bản địa, ngôn ngữ, quốc gia nơi sinh và một vài biến địa lý rộng. Vì các biên rất cụ thể như: trường đã học hay thị trấn mà trẻ em đã sống không có trong bộ dữ liệu, nên sẽ rất khó để nhà nghiên cứu kết nối hai bộ dữ liệu đó chỉ sử dụng 12 biên này. Báo cáo của chúng tôi đã xác định số lượng các bản ghi duy nhất trong bộ dữ liệu AEDC mà nhà nghiên cứu có thể sử dụng trong một dự án kết nối dữ liệu “không được phép”. Nếu tỷ lệ các bản ghi duy nhất là quá cao, thì sẽ yêu cầu nhà nghiên cứu biện minh về việc đưa vào các biên cụ thể. Ngoài ra, có thể yêu cầu nhóm một số các biên rất phân tán thành các tổ hoặc nhấn mạnh việc loại bỏ của chúng khỏi bộ dữ liệu.

(2) Trong phần thứ hai của đánh giá, cùng 12 biên được liệt kê trong một bảng cho thấy tổng số các tổ trong mỗi biên và số lượng các tổ có 3 trẻ hoặc ít hơn. Ví dụ như trong một bộ dữ liệu, tuổi của trẻ em đã được nhóm thành 14 tổ. Một trong những tổ đó chỉ có 2 trẻ em. Ô này có thể vô tình xuất hiện trong một biểu của báo cáo đã xuất bản. Đó là một vấn đề đơn giản để giảm số lượng nhóm tuổi xuống 13. Các biên có số lượng lớn các tổ và một tỷ trọng lớn các ô nhỏ có thể nhận biết một cách dễ dàng trong bộ dữ liệu bất kỳ và có thể thực hiện các bước để làm giảm rủi ro cho việc xuất bản của họ.

Trong thực tế, việc đánh giá rủi ro đối với các yêu cầu quy mô lớn là một quá trình lặp đi lặp lại nhằm thích ứng với những nhạy cảm xung quanh việc phát hành dữ liệu vi mô về từng cá nhân và các nhu cầu nghiên cứu của người xin nghiên cứu.

#### 4. Phương thức truy cập

Cho đến nay, chỉ có một phương thức truy cập được hỗ trợ liên quan tới dữ liệu vi mô AEDC: những người sử dụng được cấp quyền truy cập sử dụng cho một người dùng, tập tin ẩn danh. Có một số các tùy chọn khác có thể được khám phá nếu các nhà nghiên cứu yêu cầu dữ liệu nhân khẩu học hoặc dữ liệu địa lý chi tiết, vượt quá mức độ rủi ro có thể chấp nhận được. Các phương thức truy cập thay thế tương tự khác gồm có:

- Các phòng thí nghiệm dữ liệu, nơi các thành viên của cộng đồng nghiên cứu truy cập dữ liệu tại một địa chỉ được phê duyệt. Kết quả đầu ra được kiểm tra bởi các nhà quản lý dữ liệu được phê chuẩn và chỉ được công bố nếu các tiêu chuẩn về giảm thiểu nguy cơ lộ thông tin được đáp ứng (ví dụ như các ô nhỏ trong đầu ra dữ liệu vi mô bị chặn).

- Truy cập từ xa cho phép các thành viên của cộng đồng nghiên cứu truy cập dữ liệu thông qua một máy chủ an toàn, sử dụng kết nối Internet an toàn.

- Việc thực hiện từ xa, nơi các thành viên của cộng đồng nghiên cứu gửi mã để các nhà quản lý dữ liệu mã hóa và công bố kết quả đầu ra sau khi đã được kiểm tra về nguy cơ lộ thông tin.

Cung cấp các phương pháp tiếp cận khác để truy cập dữ liệu vi mô nằm ngoài phạm vi của bộ hiện tại của các hoạt động quản lý dữ liệu, tuy nhiên, nếu có yêu cầu, tất cả các phương thức nêu trên có thể được hỗ trợ. Tùy chọn hoặc các tùy chọn hữu hiệu nhất sẽ phụ thuộc phần lớn vào các yêu cầu của người sử dụng dữ liệu tiềm năng. Ví dụ, nếu việc tăng nền tảng kỹ năng và hiểu biết thống kê của người sử dụng dữ liệu là một ưu tiên, thì phòng thí nghiệm dữ liệu được hỗ trợ bởi các thành viên của đội quản lý dữ liệu có thể là một ưu tiên để xem xét. Hay là, hỗ trợ một cộng đồng học thuật cao có thể

tạo điều kiện tốt nhất sử dụng cách tiếp cận từ xa. Trong cả hai trường hợp, trọng tâm sẽ là việc đáp ứng những nhu cầu của cộng đồng nghiên cứu về dữ liệu vi mô chi tiết và chính xác cùng với giảm thiểu nguy cơ lộ thông tin cá nhân hoặc nhóm. Không phân biệt phương thức tiếp cận, người dùng dữ liệu sẽ vẫn yêu cầu có được sự chấp thuận cho việc truy cập dữ liệu vi mô AEDC sử dụng các thủ tục đã có.

### 5. Tác động ảnh hưởng

Từ khi thực hiện khung quản lý rủi ro mới này, đã không có vi phạm nào về báo cáo sử dụng không hợp lý dữ liệu AEDC. Các tổ chức và cơ quan xem xét việc công bố dữ liệu vi mô điều tra mẫu có thể nghiên cứu những tác động ảnh hưởng sau đây từ kinh nghiệm AEDC.

- Dữ liệu vi mô ẩn danh cho phép cải thiện tiện ích cho người sử dụng dữ liệu và không nhất thiết phải trình bày các mức độ lộ thông tin cao hơn so với một file ghi chép cơ sở được bảo mật.

- Dữ liệu vi mô ẩn danh đảm bảo rằng có "một phiên bản của sự thật" và rằng kết quả đầu ra được sản xuất bởi các nhà nghiên cứu sẽ nhất quán trong các cộng đồng nghiên cứu và chính sách. Báo cáo nhất quán của các số liệu chính rất quan trọng cho

việc nuôi dưỡng niềm tin về dữ liệu AEDC của bên liên quan và của công chúng.

- Các yêu cầu truy cập bất kỳ tập tin ghi chép cơ sở nào cần tuân theo cùng một format, các thủ tục chi tiết về đánh giá, quản lý và hoàn tất.

- Những lo ngại về lạm dụng không chú ý dữ liệu vi mô cần phải được truyền đạt rõ ràng tới cộng đồng nghiên cứu cho rằng nguy cơ của hành vi vi phạm không độc hại xuất hiện cao hơn so với các vi phạm độc hại có thể. Các nhà nghiên cứu cần phải nhận thức được cái gì cấu thành nên sự vi phạm và những hậu quả đưa lại của hành vi vi phạm đối với chính họ là các cá nhân cũng như các cơ quan hoặc tổ chức của họ.

- Nơi mà có bằng chứng xác thực rằng một người sử dụng dữ liệu tiềm năng có thể không có kỹ năng hay kinh nghiệm để phân tích và xử lý một cách thích hợp dữ liệu vi mô, thì việc đào tạo và hỗ trợ phù hợp là một điều kiện của việc công bố dữ liệu.

- Tất cả các yêu cầu truy cập dữ liệu chính đáng có thể điều tiết nhờ kết hợp việc dàn xếp liên quan đến các yếu tố dữ liệu cần thiết và việc đưa ra các phương thức truy cập được hỗ trợ (và trợ giúp).

### Tài liệu tham khảo:

- [1] F. Ritchie, "UK release practices for official microdata," *Statistical Journal of the IAOS* 26, pp. 103-111, 2009/2010.
- [2] T. Desai and F. Richie, "Effective Researcher Management," in *Joint UNECE/Eurostat work session on statistical data confidentiality*, 2009.
- [3] M. a. D. A. Elliot, "Disclosure Risk for Microdata: Report to the European Union ESP/204 62/DG III," 1998.
- [4] United Nations, "Managing Statistical Confidentiality & Microdata Access", United Nations, New York and Geneva, 2007.
- [5] W. G. A. H. A. Thomas, *Metadata standards to support controlled access to microdata*, Tarragona, Spain, 2011.