Hỗ trợ các nhà nghiên cứu và các tổ chức trong việc khai thác cơ sở dữ liệu hành chính cho mục đích thống kê: CHIẾN LƯỢC CỦA CƠ QUAN THỐNG KÊ QUỐC GIA Ý

Giovanna D'Angiolini, Pierina De Salvo và Andrea Passacantilli, Cơ quan Thống kê quốc gia Ý

Tóm tắt

Bài viết trình bày về chiến lược của Cơ quan Thống kê quốc gia Ý (Istat) nhằm hỗ trợ cả những người sử dụng truyền thống cũng như những người mới sử dụng nguồn dữ liệu hành chính cho mục đích thống kê. Một chiến lược như vậy bao gồm một số hoạt động nhằm mục đích liệt kê những nguồn dữ liệu hành chính sẵn có, ghi rõ nội dung thông tin và chất lượng của nguồn dữ liệu hành chính, tạo điều kiện thuận lợi để điều chỉnh nội dung của những nguồn dữ liệu hành chính này cho phù hợp với các tiêu chuẩn thống kê. Những thông tin thu thập được về các nguồn dữ liệu hành chính sẵn có được phổ biến cho người sử dụng thống kê tiềm năng thông qua một hệ thống quản lý siêu dữ liệu dựa trên web chuyên dụng được gọi là DARCAP. Ngoài ra, để hỗ trợ các phân tích chuyên sâu về chất lượng của các nguồn dữ liệu hành chính quan trọng nhất, chúng tôi đang nghiên cứu *Khung đánh giá chất lượng mới đối với các nguồn dữ liệu hành chính*.

Từ khóa: Nguồn dữ liệu hành chính, tài liệu dữ liệu hành chính, chất lượng dữ liệu hành chính, mô hình nguồn dữ liệu, sản xuất số liệu thống kê.

1. Chiến lược của cơ quan thống kê quốc gia Ý hỗ trợ sử dụng các nguồn dữ liệu hành chính cho mục đích thống kê: Sự hình thành và hoạt động

Ngày nay, có rất nhiều cơ quan thống kê quốc gia khai thác dữ liệu hành chính cho mục đích thống kê, nhằm nâng cao chất lượng của các sản phẩm đầu ra thống kê, giảm bớt gánh nặng cho người trả lời và để giảm thiểu tối đa chi phí [2] [8]. Vì vậy việc xác định nội dung của các nguồn dữ liệu hành chính sẵn có và đánh giá chất lượng của các dữ liệu hành chính đã thu thập được là mối quan tâm tất yếu đối với các cơ quan thống kê quốc gia. Tuy nhiên trong viễn cảnh này, tài liệu của các nguồn dữ liệu hành chính được tạo ra khi cần thiết, và chất lượng dữ liệu hành chính nói chung được đánh giá từ quan điểm của quá trình sản xuất dữ liệu của từng cơ quan thống kê quốc gia cụ thể mà trong đó dữ liệu hành chính có liên quan, như dữ liệu đầu vào hoặc dữ liệu phụ [3] [9].

Viễn cảnh truyền thống này đang phát triển một cách nhanh chóng. Ngày nay, việc sử dụng nguồn dữ liệu hành chính trong quy trình sản xuất dữ liệu của các cơ quan thống kê không chỉ là việc sử dụng thống kê đối với dữ liệu hành chính nữa. Nhờ có sự lan rộng của phương pháp tiếp cận về kho dữ liệu trong những năm gần đây mà ngày càng nhiều tổ chức phi thống kê đã và đang triển khai các hệ thống hỗ trợ việc ra quyết định của riêng mình - những hệ thống này khai thác dữ liệu hành chính nhằm theo dõi bối cảnh và tác động của những hoạt

động của tổ chức. Những hệ thống như vậy trên thực tế sử dụng các kỹ thuật thống kê cho dù mục đích của chúng không phải là sản xuất số liệu thống kê.

Việc sử dụng dữ liệu hành chính hỗ trợ ra quyết định đòi hỏi dữ liệu được khai thác phải có chất lượng tốt khi được xem là công cụ đo lường của các hiện tượng trong thế giới thực, nghĩa là nó đòi hỏi dữ liệu phải có chất lượng tốt xét từ quan điểm thống kê. Như một kết quả tất yếu, việc quan tâm tới vấn đề chất lượng dữ liệu hành chính đang lan rộng trong một số cộng đồng nghiên cứu như cộng đồng nghiên cứu cơ sở dữ liệu [1]. Gần đây, tầm nhìn về dữ liệu mở càng làm củng cố hơn xu hướng này.

Trong viễn cảnh mới như vậy đòi hỏi các cơ quan thống kê quốc gia phải đóng một vai trò quan trọng mới. Các cơ quan thống kê quốc gia phải suy nghĩ và đưa ra được các hướng dẫn, các phương pháp và công cụ nhằm hỗ trợ cho bất cứ người sử dụng nào cần để khai thác dữ liệu hành chính để có được hiểu biết về các hiện tượng trong thế giới thực.

Điều đặc biệt ở Ý là nhiều cơ quan thuộc Hệ thống thống kê quốc gia Ý (chẳng hạn như các cơ quan chính phủ cần theo dõi các tác động của những chính sách mà họ ban hành), đang xây dựng các kho dữ liệu lớn mà trong đó có thể bao gồm các cơ sở dữ liệu hành chính do chính họ sở hữu cùng với dữ liệu điều tra và các cơ sở dữ liệu hành chính bên ngoài.

Tuy nhiên, thường thì khả năng các nguồn dữ liệu hành chính trở thành nguồn thông tin thống kê bị hạn chế, do thiếu các thông tin phù hợp về nội dung và chất lượng của những nguồn dữ liệu hành chính này hoặc do chúng sử dụng các định nghĩa, phân loại và quy tắc quản lý dữ liệu có tính định hướng hành chính.

Để xử lý những hạn chế như vậy, Istat đã cam kết thực hiện một chiến lược chung nhằm làm cho những nguồn dữ liệu hành chính sẵn có trở nên dễ hiểu và dễ sử dụng hơn [5].

Nói chung, chiến lược của Istat nhằm mục đích:

• Thu thập thông tin về các nguồn dữ liệu hành chính sẵn có và tạo ra các tài liệu chuẩn về nội dung thông tin và chất lượng của nguồn dữ liệu hành chính

• Điều chỉnh (khi có thể) nội dung của các nguồn dữ liệu hành chính sẵn có thông qua việc sử dụng các định nghĩa, phân loại thống kê và quy tắc quản lý dữ liệu chuẩn.

Cung cấp cho người sử dụng những kiến thức phù hợp về nội dung và chất lượng của các nguồn dữ liệu hành chính là bước đi đầu tiên nhằm đẩy mạnh việc khai thác thống kê của họ. Để thực hiện một nhiệm vụ như vậy, Istat đang triển khai một số hoạt động có tính hệ thống về tài liệu có liên quan đến các loại nguồn dữ liệu hành chính khác nhau.

Các cơ quan chính phủ trung ương quản lý các hệ thống thông tin lớn cấu thành từ nhiều nguồn dữ liệu hành chính được duy trì và khai thác thông qua các thủ tục hành chính. Trong bối cảnh đó, các chuyên gia của Istat cùng với các chuyên gia của nguồn dữ liệu cùng thực hiện một cuộc điều tra riêng một cách có hệ thống về từng nguồn dữ liệu hành chính và các biểu mẫu hành chính có liên quan của nguồn dữ liệu. Một cuộc điều tra về nguồn dữ liệu hành chính là một hoạt động thu thập và phân tích tài liệu tuân theo một mẫu tiêu chuẩn để thu thập các thông tin có thể so sánh được về nội dung và chất lượng của nguồn dữ liệu, như được trình bày trong phần 2 dưới đây. Các thông tin đã thu thập

được quản lý bởi các công cụ của một hệ thống quản lý siêu dữ liệu dựa trên web chuyên dụng được gọi là DARCAP (Lưu trữ tư liệu hành chính công) để phổ biến tới bất kỳ người sử dụng thống kê tiềm năng của các nguồn dữ liệu hành chính. Hệ thống DARCAP được minh họa ngắn gọn trong phần 3.

Cuộc điều tra chuyên dụng như vậy cho phép chúng tôi lập các tài liệu một cách kỹ lưỡng về nội dung thông tin của những nguồn dữ liệu hành chính sẵn có, nhưng họ chỉ thu thập được một số lượng hạn chế những đánh giá định tính về chất lượng những nguồn dữ liệu đó. Đối với những nguồn dữ liệu phức tạp và quan trọng nhất, người sử dụng thống kê có thể phải cần đến các thông tin bổ sung về chất lượng của nguồn dữ liệu. Nhằm hỗ trợ cho việc phân tích chất lượng một cách chuyên sâu, chúng tôi đang nghiên cứu một Khung đánh giá chất lượng mới đối với các nguồn dữ liệu hành chính, Khung này được mô tả ngắn gọn trong phần 4.

Không giống như các cơ quan Trung ương, các cơ quan chính quyền địa phương thường xuyên quản lý rất nhiều các cơ sở dữ liệu hành chính độc lập nhằm hỗ trợ một lượng lớn các nhiệm vụ hành chính không đồng nhất có liên quan tới nhiều chủ đề, trải rộng từ quản lý môi trường tới giám sát nhân sự. Để có được những hiểu biết về các nguồn dữ liệu hành chính như vậy, Istat cùng với các cơ quan đại điện cho chính quyền địa phương tổ chức các cuộc điều tra chuyên dụng về các nguồn dữ liệu hành chính. Những cuộc điều tra này liệt kê các nguồn dữ liệu hành chính hiện có và phân loại chúng theo các chủ đề. Ngoài ra, các cuộc điều tra này còn thu thập một số thông tin khác của từng nguồn dữ liệu hành chính, chẳng hạn như đối tượng quan sát chính và các biến. Các thông tin đã thu thập cũng sẽ được lưu trữ vào hệ thống DARCAP.

Tất cả những hoạt động mô tả ở trên được định hướng nhằm cung cấp cho những người sử dụng thống kê tiềm năng của các nguồn dữ liệu hành chính những thông tin phù hợp về nội dung và chất lượng của các nguồn dữ liệu đó.

Istat cũng đang triển khai các hoạt động khác nhằm tạo điều kiện thuận lợi để điều chỉnh nội dung của các nguồn dữ liệu hành chính. Hoạt động này là sự giám sát các thay đổi và dự án đổi mới liên quan đến các nguồn dữ liệu hành chính và các biểu mẫu có liên quan của chúng. Điều đáng chú ý là theo Luật Thống kê Ý thì các quan sở hữu các nguồn dữ liệu hành chính cần phải tuân thủ các khuyến nghị của Istat liên quan tới biểu mẫu và các nguồn dữ liệu họ đang quản lý, tuy nhiên trên thực tế thì điều này rất khó để thực thi. Hoạt động này nhằm khắc phục vấn đề nêu trên.

Đối với các nguồn dữ liệu hành chính quan trọng nhất, cơ quan sở hữu nguồn dữ liệu được yêu cầu phải thông báo tới Istat mỗi khi họ có kế hoạch thay đổi nội dung nguồn thông tin. Một thông báo như vậy có liên quan tới tất cả các loại thay đổi, thay đổi định kỳ các hình thức thu thập biểu mẫu bảng kê cũng như các dự án đổi mới lớn chẳng hạn như là một kho dữ liệu mới.

Trên cơ sở các thông báo đã nhận được, Istat sẽ đưa ra các phản hồi và khuyến nghị phù hợp. Ví dụ về các khuyến nghị đó là: sử dụng các bảng phân loại chính thức thay thế cho các bảng phân loại không chính thức, cải tiến hệ thống mã định danh, cải tiến các thủ tục kiểm soát chất lượng.

Hệ thống DARCAP cung cấp cho các cơ quan sở hữu các nguồn dữ liệu hành chính một hệ thống phụ chuyên dụng nhằm hỗ trợ các hoạt động thông báo thay đổi. Tất cả các thông báo nhận được cùng với các khuyến nghị có liên quan của họ đều được lưu trữ vào hệ thống DARCAP. Ngoài ra, các chuyên gia của Istat còn có thể phân tích nội dung thông tin của các biểu mẫu và các nguồn dữ liệu hành chính mới được tạo ra, giống như họ làm với các nguồn dữ liệu và biểu mẫu hiện có.

Tất cả các hoạt động mô tả ở trên được điều phối bởi một Ủy ban hài hòa các biểu mẫu hành chính *(Committee for Harmonizing Administrative Forms)*, mà các thành viên của Ủy ban này do cơ quan thống kê quốc gia Ý và những cơ quan sở hữu các nguồn dữ liệu hành chính quan trọng nhất chỉ định. Ủy ban này được một mạng lưới các chuyên gia hỗ trợ.

2. Hoạt động điều tra các nguồn dữ liệu hành chính: biên soạn tài liệu về nội dung và chất lượng của các nguồn dữ liệu hành chính sẵn có

Việc điều tra về một nguồn dữ liệu hành chính được thực hiện bằng các công cụ phân tích tài liệu sẵn có và phỏng vấn các chuyên gia về nguồn dữ liệu của cơ quan sở hữu nguồn dữ liệu cũng như những người sử dụng nguồn dữ liệu. Các tài liệu đã thu thập được sau đó được cấu trúc theo cấu trúc cơ sở dữ liệu của DARCAP, để được lưu trữ vào cơ sở dữ liệu đó.

Hoạt động điều tra bao gồm ba hành động: 1) Xác định một mô tả chung của nguồn dữ liệu, (2) Phân tích và ghi lại nội dung thông tin của nguồn dữ liệu, (3) Thu thập thông tin về chất lượng dữ liệu của nguồn dữ liệu.

1) Xác định một mô tả chung của nguồn dữ liệu: chúng tôi định rõ tên gọi và mục đích của nguồn dữ liệu hành chính, cơ quan sở hữu và các cơ quan quản lý khác, các luật quy định nó và các luật quy định khác, các thủ tục hành chính có liên quan, tập hợp các biểu mẫu hành chính hoặc các công cụ khác đang được sử dụng để làm dữ liệu đầu vào duy trì nguồn dữ liệu hành chính.

2) Phân tích và ghi lại nội dung thông tin của nguồn dữ liệu: hoạt động ghi chép tài liệu nhằm mục đích tạo ra một tiêu chuẩn kỹ thuật chuẩn và có thể so sánh được nội dung của mỗi nguồn dữ liệu hành chính sẵn có trong mối quan hệ với các đối tượng quan sát được ở thế giới thực. Theo cách sử dụng rộng rãi, chúng tôi gọi đặc điểm kỹ thuật như vậy của nội dung thông tin là mô hình tài liệu nguồn dữ liệu hành chính

Chính xác hơn, một mô hình nguồn dữ liệu hành chính là một mô tả có cấu trúc nội dung thông tin của chính nó, dựa trên mô hình khái niệm tiêu chuẩn. Để xác định một mô hình khái niệm như vậy, chúng tôi đã phân tích vòng đời của dữ liệu hành chính và chọn ra các loại đối tượng khác nhau trong thế giới thực có liên quan tới chúng, và chúng tôi đặt những đối tượng đó trong mối quan hệ tương ứng với các đối tượng khác hiện có liên quan tới bất kỳ một thống kê nào, đó là các tập hợp và các biến số. Mô hình khái niệm của chúng tôi được định hướng để hướng tới hỗ trợ việc khai thác về thống kê đối với các nguồn dữ liệu hành chính, tuy nhiên nó có thể dễ dàng được chuyển đổi sang thành các ngôn ngữ và mô hình khái niệm đa mục đích phục vụ cho tiêu chuẩn kỹ thuật của mô hình tài liệu nguồn dữ liệu hành chính [4].

Các nguồn dữ liệu hành chính thu thập các thông tin về một số loại đối tượng trong thế giới thực

để nhằm hỗ trợ các hoạt động hành chính. Trước hết bất kỳ hoạt động hành chính nào đều đòi hỏi phải thu thập dữ liệu về những thực thể mà các hoạt động đó nhắm đến. Các thực thể này là tập hợp con của hai tổng thể dân cư, một mặt các thực thể này thực hiện các hoạt động kinh tế, mặt khác chúng là tập hợp con của tổng thể dân cư có liên quan như các hộ gia đình, đơn vị lãnh thổ của doanh nghiệp. Ngoài ra, các thông tin được thu thập về tập hợp các sự kiện cụ thể đó có thể liên quan đến thực thể và là mối quan tâm nhằm tới mục đích của các hoạt động hành chính. Tổng thể dân cư và tập hợp các sự kiện đã quan sát được liên kết với nhau bằng các mối quan hệ. Đối với cả tổng thể dân cư và tập hợp các sự kiện đã quan sát, các thông tin phù hợp được thu thập về các đặc điểm của chúng, có thể được thay đổi theo thời gian.

Ví dụ, Bộ Giáo dục công liên tục thu thập thông tin về các sinh viên, các trường học và các trường đại học cùng với các đặc điểm của chúng cũng như về tập hợp các sự kiện như mức độ đăng ký tham dự khóa học, các kỳ thi, mức độ thu nhập với các đặc điểm của chúng. Mỗi thành tố của những đối tượng quan sát này có các đặc điểm định tính hoặc đặc điểm định lượng như ngày tháng năm sinh, nơi cư trú, ngày nhập học, điểm thi, cũng như mối quan hệ với các thành tố trong các đối tượng quan sát khác.

Vì vậy, chúng tôi ghi chép lại các tổng thể quan sát được – những tổng thể tương ứng với các tập hợp đó là mục tiêu của các thủ tục hành chính, và tập hợp các sự kiện có liên quan của chúng, mỗi tổng thể này gắn với một định nghĩa về nó. Chúng tôi cũng ghi chép lại các đặc điểm chính được sở hữu bởi những thành tố đơn lẻ thuộc những tập hợp nhất định gắn với các định nghĩa của nó, và các phân loại có liên quan (danh sách các phương thức) đối với các đặc điểm định tính. Từ quan điểm thống kê, các đặc điểm định lượng và các đặc điểm định tính cùng với các phân loại có liên quan được xem như là các biến số.

Công việc mô tả khái niệm của nội dung của nguồn dữ liệu hành chính tạo ra phân loại của mô hình nguồn dữ liệu, trong đó bao gồm: các tập hợp chính mà có thể là tổng thể dân cư hay tập hợp các sự kiện, các đặc điểm chính của tổng thể dân cư hoặc của tập hợp các sự kiện, và cũng có thể là các mối quan hệ có liên quan tới tổng thể dân cư và tập hợp các sự kiện.

Kết quả của công việc này là một mạng lưới của các tổng thể dân cư hoặc tập hợp của các sự kiện chủ yếu, liên kết bởi mối quan hệ 1-1 hoặc 1-nhiều. Trong đó mỗi tập hợp có các định nghĩa riêng và đặc điểm có liên quan riêng.

Một phân tích sâu hơn về nguồn dữ liệu hành chính có thể dẫn đến việc chọn ra nhiều hơn các tổng thể dân cư hoặc tập hợp các sự kiện mà trong đó có sự kết hợp giữa các đặc điểm và mối quan hệ của chúng và liên kết với các tập hợp chủ yếu thông qua tập hợp con hoặc các mối quan hệ phân vùng. Một mối quan hệ con chỉ đơn giản là sự liên kết với hai tập hợp khi một tập hợp thu được một phần của các thành tố của tập hợp kia. Một mối quan hệ phân vùng liên kết một tập hợp với nhiều tập hợp cùng chia cắt nó, đó là: mỗi thành tố của tập hợp được phân vùng thuộc về một và chỉ một trong những tập hợp phân vùng.

3) Thu thập thông tin về chất lượng dữ liệu của nguồn dữ liệu: với một bảng hỏi chuyên dụng, chúng tôi phỏng vấn các chuyên gia của nguồn dữ liệu để thu thập thông tin được sử dụng cho một đánh giá ban đầu về chất lượng của nguồn dữ liệu.

Với mục đích này, chúng tôi hỏi các chuyên gia của nguồn dữ liệu về các thông tin có liên quan tới mỗi tổng thể dân cư hoặc một tập hợp các sự kiện. Đối với mỗi tổng thể dân cư, chúng tôi ghi chép lại các sự kiện đưa vào và đưa ra và cách mà theo đó việc đăng ký chúng ảnh hưởng tới mức độ bao phủ của tổng thể dân cư. Đối với mỗi tập hợp của các sự kiện, chúng tôi ghi lại cách mà theo đó các sự kiện riêng lẻ được ghi vào nguồn dữ liệu và phân bổ thời gian của các sự kiện cũng như các vấn đề liên quan như: phạm vi đăng ký, cụ thể là khả năng đăng ký hiệu quả tất cả các sự kiện đơn lẻ được mong đợi, sự sai lệch có hệ thống của đăng ký liên quan tới các mục đích của thủ tục đăng ký hành chính, đăng ký kịp thời, cụ thể là khoảng thời gian chậm trễ giữa sự xuất hiện của sự kiện và việc đăng ký nó.

Các vấn đề chính và các biện pháp can thiệp có thể có liên quan tới các định nghĩa của tập hợp, sự phù hợp của các phân loại đã sử dụng và sự tương thích của chúng với các phân loại chuẩn, các mã định danh có thể đã được sử dụng để liên kết chính xác với các nguồn dữ liệu khác cũng đã được đánh giá. Đối với các nguồn dữ liệu hành chính nói chung, các vấn đề chính và các biện pháp can thiệp có thể có liên quan tới việc sử dụng phù hợp cho mục đích thống kê và sự phổ biến kịp thời của nó cũng được đánh giá cùng với các chiến lược đổi mới có liên quan.

Theo cách như vậy, chúng tôi thu được đánh giá định tính ban đầu về chất lượng của nguồn dữ liệu. Nhằm mục đích phân tích sâu hơn về chất lượng của các nguồn dữ liệu hành chính, điều đó rất hữu ích và cần thiết để tính toán các chỉ tiêu bằng số phù hợp với tiêu chuẩn. Như đã được mô tả trong phần 4, Khung đánh giá chất lượng đối với các nguồn dữ liệu hành chính đã chỉ rõ các khái niệm, phương pháp và các chỉ tiêu cụ thể phục vụ cho việc đánh giá chất lượng chuyên sâu.

3. Quản lý và phổ biến các thông tin thu thập được về các nguồn dữ liệu hành chính sẵn có: hệ thống DARCAP

Như chúng tôi đã giới thiệu, DARCAP (Lưu trữ tư liệu hành chính công) là một hệ thống quản lý thông tin đựa trên web nhằm hỗ trợ cuộc điều tra các nguồn dữ liệu hành chính và các sáng kiến về tài liệu khác để cung cấp cho những người sử dụng tiềm năng các nguồn dữ liệu hành chính với các tài liệu có tổ chức về nội dung và tính năng của chúng [6].

Công cụ này cũng hỗ trợ cho các cơ quan sở hữu các nguồn dữ liệu hành chính trong việc gửi cho Istat các thông báo của họ về bất kỳ sự thay đổi nào mà trong đó có thể tác động tới các nguồn dữ liệu hành chính do họ quản lý hoặc các mẫu biểu hành chính có liên quan, và các tài liệu khuyến nghị của Istat phát hành.

Một cách chính xác hơn, DARCAP gồm 3 hệ thống con:

• DARCAP-Documenta: Nó cung cấp cho các chuyên gia của Istat các chức năng phụ vụ cho việc lập các tài liệu về nội dung thông tin và chất lượng của các nguồn dữ liệu hành chính quan trọng nhất được quản lý bởi các cơ quan hành chính trung ương, bằng cách lưu trữ các kết quả của các hoạt động điều tra chuyên dụng (đã được mô tả ở phần 2 nói trên). Ngoài ra, nó cũng cung cấp các chức năng để lưu trữ các kết quả của điều tra về các nguồn dữ liệu hành chính được quản lý bởi các cơ quan hành chính địa phương;

• DARCAP-Innova: Nó cung cấp cho các cơ quan hành chính các chức năng thông báo tới Istat mỗi lần họ có kế hoạch thay đổi các mẫu biểu hoặc các nguồn dữ liệu hành chính do họ quản lý. Nó cho phép các chuyên gia của Istat đưa ra các phản hồi về các dự án đổi mới đã được thiết kế và đưa ra những khuyến nghị phù hợp. Ngoài ra, nó còn cho phép các chuyên gia của Istat ghi chép lại nội dung thông tin về các mẫu biểu hoặc các nguồn dữ liệu hành chính mới được thiết kế khi cần thiết, bằng cách sử dụng các chức năng chuyên dụng của hệ thống con DARCAP-Documenta;

• DARCAP Consultazione: Đây là hệ thống con yêu cầu thông tin, nhằm mục đích phổ biến những thông tin thu thập được về các nguồn dữ liệu hành chính sẵn có tới những người sử dụng thống kê tiềm năng.

Đặc biệt, DARCAP Consultazione cung cấp cho người sử dụng cuối cùng hai môi trường riêng biệt phục vụ để truy cập tài liệu của các dự án đổi mới hoặc điều hướng một cách lần lượt thông qua tài liệu của các nguồn dữ liệu hành chính hoặc các mẫu biểu.

Truy cập tài liệu của các dự án đổi mới: đó là khả năng có thể tìm kiếm một dự án đổi mới theo tên dự án và tên của tổ chức và hiển thị tất cả đặc trưng chung hoặc cụ thể của bất kỳ dự án đổi mới nào, bao gồm tài liệu của các nguồn dữ liệu hành chính mới được thiết kế hoặc các mẫu biểu khi nó được tạo ra, cũng như các khuyến nghị của Istat.

Điều hướng thông qua tài liệu của các nguồn dữ liệu hành chính hoặc mẫu biểu hiện có: môi trường này cung cấp cho người dùng cuối cùng hai chức năng tìm kiếm khác nhau.

Chức năng tìm kiếm đầu tiên là tìm kiếm một nguồn dữ liệu hành chính hoặc một mẫu biểu hành chính theo tên và các tiêu chuẩn khác (tên và tiêu chuẩn khác phụ thuộc vào loại hình của cơ quan sở hữu nguồn dữ liệu).

Tìm kiếm theo tên đòi hỏi một chuỗi đặc điểm kỹ thuật. Đối với những nguồn dữ liệu hành chính hay các mẫu biểu hành chính thuộc quyền sở hữu của các cơ quan trung ương thì các tiêu chuẩn tìm kiếm là: thời hạn hiệu lực, loại nguồn dữ liệu, tên cơ quan quản lý. Đối với những nguồn dữ liệu hành chính hoặc mẫu biểu hành chính thuộc sở hữu của các cơ quan chính quyền địa phương thì các tiêu chuẩn tìm kiếm là: thời hạn hiệu lực, tên và loại cơ quan quản lý, vùng, loại thủ tục hành chính có liên quan, lĩnh vực chủ đề chung và lĩnh vực chủ đề cụ thể. Tiêu chuẩn sau đó tương ứng với một phân loại chính thức của nội dung chủ đề của nguồn dữ liệu hành chính. Danh sách các lựa chọn phù hợp được hiển thị cho mỗi tiêu chuẩn. Hệ thống hiển thị danh sách các nguồn dữ liệu hành chính hoặc mẫu biểu hành chính thỏa mãn các điều kiện cụ thể, trong đó người sử dụng cuối cùng có thể lựa chọn.

Chức năng tìm kiếm thứ hai là tìm kiếm một nguồn dữ liệu hành chính hoặc một mẫu biểu hành chính theo nội dung thông tin: cho một chuỗi đặc điểm kỹ thuật, hệ thống hiển thị tất cả các tập hợp, các đặc điểm và các phân loại mà tên của chúng có chứa chuỗi ký tự cụ thể, và với mỗi loại có chứa đựng các nguồn dữ liệu hành chính hoặc các mẫu biểu hành chính như vậy, người dùng cuối cùng có thể lựa chọn giữa các loại đó.

Một lần nữa những người sử dụng cuối cùng lựa chọn một mẫu biểu hành chính hoặc một nguồn dữ liệu cụ thể mà họ có thể duyệt thông qua các tài liệu có liên quan của nó. Chính xác hơn, họ truy cập:

• Tên, mô tả và hiệu lực thời hạn, và một danh sách đơn giản của các tập hợp quan sát được, các đặc điểm và các phân loại;

• Một bài trình bày bằng đồ thị của mô hình nguồn dữ liệu, cụ thể là mạng lưới các tập hợp và mối quan hệ giữa chúng, với mỗi tập hợp, khả năng xem được danh sách các đặc điểm của nó cùng với các phân loại có liên quan và mạng lưới các tập hợp là tập hợp con của nó.

• Các tính năng chung khác như: các cơ quan sở hữu và các cơ quan quản lý khác, các thủ tục hành chính có liên quan và các luật quy định, đối với các nguồn dữ liệu hành chính các mẫu biểu hành chính đầu vào, dữ liệu hoặc các nguồn dữ liệu hành chính khác, và các thông tin khác bao gồm các tài liệu kèm theo và địa chỉ của các trang web.

Đối với các nguồn dữ liệu hành chính, có thể tải về một tài liệu dưới dạng pdf trong đó bao gồm bảng hỏi để điền của hiệu lực về chất lượng nguồn dữ liệu hành chính, bảng hỏi này thu thập các thông tin về một số khía cạnh như: việc sử dụng thực tế hoặc tiềm năng của nguồn dữ liệu hành chính, các thủ tục thu thập thông tin và mức độ bao phủ ước tính của các tập hợp quan sát được.

Trong phiên bản 2 của DARCAP, đối với các mẫu biểu hành chính, có thể xem nội dung thông tin liên quan đến các phần khác nhau tạo thành cấu trúc của chúng. Nó cũng có thể làm nổi bật một phần trong giao diện và mở một cửa sổ với phân loại của nội dung thông tin chi tiết của nó.

4. Đánh giá sâu về chất lượng của nguồn dữ liệu hành chính: Khung đánh giá chất lượng dữ liệu hành chính

Khung đánh giá chất lượng đối với các nguồn dữ liệu hành chính là công cụ phương pháp luận của Istat để đánh giá chất lượng của những nguồn dữ liệu hành chính sẵn có [7].

Như chúng ta đã thấy xu hướng là sự phát triển rộng khắp của kho dữ liệu và sự gia tăng sử dụng các nguồn dữ liệu hành chính cho các mục đích phi hành chính bắt buộc các cơ quan thống kê quốc gia phải chịu trách nhiệm thực hiện nhiệm vụ phương pháp luận kết hợp mới, cụ thể là để xác định một tập hợp đầy đủ và linh hoạt của tiêu chuẩn và quy trình đánh giá chất lượng có thể lặp lại đối với các nguồn dữ liệu hành chính, giống như họ đang làm với các cuộc điều tra [5].

Do vậy, khung đánh giá chất lượng đối với các nguồn dữ liệu hành chính xác định khung các chỉ tiêu chất lượng hợp lý để điều chỉnh bất kỳ một ai đó bên ngoài hay bên trong một cơ quan thống kê quốc gia, đặc biệt là bản thân chủ sở hữu các nguồn dữ liệu hành chính, nhằm đánh giá chất lượng của bất kỳ nguồn dữ liệu hành chính sẵn có nào.

Để đáp ứng được nhu cầu như vậy, chúng tôi đã dựa trên một khung phân tích kỹ lưỡng với các mục tiêu và đặc điểm riêng biệt của quá trình thu thập dữ liệu hành chính và các tác động của chúng về chất lượng của dữ liệu thu thập được.

Một phân tích như vậy đã được thực hiện đối với từng loại khác nhau của các đối tượng được quan sát hình thành nên bất kỳ mô hình nguồn dữ liệu nào [6]. Cách tiếp cận của chúng tôi là sáng tạo bởi vì mô tả của nội dung của một nguồn dữ liệu dựa trên mô hình dữ liệu không phải là một thực hành thường thấy giữa các nhà thống kê mặc dù trên thực tế tài liệu dữ liệu dựa trên mô hình dữ liệu là một thực hành phổ biến. Bằng cách giữ các chỉ tiêu đã được đề xuất cho mô hình nguồn dữ liệu, chúng tôi đảm bảo một hệ thống đặc điểm kỹ thuật của các chỉ tiêu và chúng tôi cung cấp các chuyên gia đánh giá chất

lượng với những định hướng cho việc lựa chọn giữa các biến có thể tính toán cũng như cho việc giải thích các chỉ tiêu được tính toán.

Khung được sắp xếp theo cấu trúc do cơ quan thống kê Hà Lan đề xuất, trong đó phân biệt ba quan điểm khác nhau về chất lượng, cụ thể là quan điểm về nguồn dữ liệu, quan điểm về siêu dữ liệu, và quan điểm về dữ liệu. Mỗi quan điểm này được gọi là "đa chiều" bao gồm một số khía cạnh, các phương pháp và chỉ tiêu chất lượng.

Trong đa chiều về nguồn dữ liệu, các khía cạnh chất lượng liên quan đến nguồn dữ liệu hành chính nói chung, chủ sở hữu của nguồn dữ liệu, và các điều kiện cung cấp. Đa chiều về siêu dữ liệu đặc biệt tập trung vào siêu dữ liệu liên quan tới các khía cạnh của nguồn dữ liệu hành chính. Nó liên quan tới sự tồn tại và tính đầy của của tài liệu và liên quan tới loại và cấu trúc của các mã định danh. Đa chiều về dữ liệu tập hợp tất cả các chỉ tiêu định lượng được tính toán từ dữ liệu và nhằm mục tiêu đo lường các phương diện chất lượng truyền thống cho dữ liệu thu thập được, chẳng hạn mức độ bao phủ của các tập hợp được quan sát và độ chính xác của các giá trị thu thập được cho các đặc điểm quan sát.

Đối với đa chiều về nguồn dữ liệu và siêu dữ liệu, Khung đề xuất một bộ các chỉ tiêu định tính tương tự như bộ chỉ tiêu đã được đề xuất trong dự án BLUE-ETS. Lưu ý rằng ngoài yêu cầu các chủ sở hữu dữ liệu hành chính phải xác nhận tính sẵn có của tài liệu nguồn dữ liệu hành chính, chúng tôi cũng cũng cung cấp cho họ các công cụ chuẩn phù hợp để quản lý những tài liệu như vậy, cụ thể là hệ thống DARCAP.

Đối với đa chiều về dữ liệu, hiện tại chúng tôi đang xác định một bộ chỉ tiêu cấu trúc hơn và phong phú hơn, hoàn toàn dựa trên mô hình nguồn dữ liệu hành chính. Nó bao gồm cả các chỉ tiêu định tính và định lượng.

Các chỉ tiêu định tính trong đa chiều về dữ liệu được xác định bằng cách khai thác hoạt động điều tra, nó đã thu thập một đánh giá chất lượng ban đầu riêng biệt cho mỗi tập hợp (các tổng thể dân cư và tập hợp các sự kiện) trong nguồn dữ liệu hành chính.

Với các chỉ tiêu định lượng, cụ thể là các chỉ tiêu được tính toán từ dữ liệu và vì vậy đòi hỏi tính sẵn có của bộ dữ liệu, chung phải được tính toán bởi chủ sở hữu dữ liệu hành chính cũng như bởi cơ quan thống kê quốc gia khi mà nó thu được bộ dữ liệu. Viễn cảnh tốt nhất là khi một thủ tục tính toán có tính chất cộng tác được áp dụng.

Để xác định các chỉ tiêu định lượng đó, trước tiên một mặt chúng tôi phải phân biệt giữa các sai sót có thể và mặt khác là các cách kiểm tra chúng. Các sai sót có thể được xác định liên quan đến các đối tượng đó có thể xuất hiện trong một mô hình nguồn dữ liệu hành chính theo cách sau đây.

Đối với mỗi đối tượng trong một mô hình dữ liệu, cụ thể là một tập hợp, một đặc điểm hay một mối quan hệ, chúng tôi có thể xây dựng các bảng kê liên quan tới các thành tố được quan sát. Các nguồn dữ liệu hành chính tiếp tục thu thập và lưu trữ dữ liệu mà trên thực tế kết hợp một cách phù hợp các bảng kê đó.

Ví dụ, giả sử rằng một học sinh mới đăng ký trong một danh sách đăng ký học sinh, đó là một thành tố mới được nhập vào tổng thể học sinh, một thành tố mới được nhập vào tập hợp các sự kiện đăng ký nhập học. Nếu học sinh mới được cấp một mã định danh n và việc đăng ký nhập học được cấp một mã định danh i, danh sách đăng ký học sinh chấp nhận hai bản ghi mới: 1) Một bản ghi kết nối

bảng kê học sinh (n) với bảng kê cư trú (n, Milan) và những bảng kê tương tự khác liên quan tới đặc điểm đã đăng ký của học sinh mới, 2) Một bản ghi khác kết nối bảng kê đăng ký nhập học (i) với các bảng kê đăng ký nhập học_sinh viên (i, n), đăng ký nhập học_khóa học (i, thống kê) và có thể các bảng kê khác liên quan tới các đặc điểm đã đăng ký của bản thân việc đăng ký nhập học.

Có thể xảy ra trường hợp một số bảng kê bị sai, và một số bảng kê đúng nhưng không có trong bộ dữ liệu. Do đó, bất cứ lúc nào chúng tôi cũng có thể có trong nguồn dữ liệu hành chính:

• *Các sai sót bao gồm*: các bảng kê sai (một cách chắc chắn hoặc tạp thời) chấp nhận trong nguồn dữ liệu

• *Các sai sót loại trừ*: các bảng kê đúng (chắc chắn hoặc tạm thời) loại trừ từ nguồn dữ liệu

Các sai sót khác có thể liên quan tới sự nhận dạng sai của các thành tố liên quan, bởi vì các vấn đề trong hệ thống mã định danh, như là: lỗi cú pháp trong nhận dạng, nhận dạng các phần tử không tồn tại, thiếu nhận dạng cho các phần tử hiện có, có nhiều hơn một định dạng cho mỗi phần tử, các phần tử chia sẻ định dạng.

Đối với mỗi tập hợp (tổng thể dân cư hoặc tập hợp các sự kiện), các sai sót bao gồm hoặc loại trừ lần lượt tương ứng với các sai sót bao phủ quá mức và sai sót bao phủ dưới mức, và bằng cách kết nối chúng với các sai sót nhận dạng chúng tôi thu được một bản ghi đặc điểm kỹ thuật của tất cả các sai sót có thể có liên quan tới tập hợp.

Đối với mỗi đặc điểm bắt buộc, chúng tôi có thể có một sai sót loại trừ, sai sót này tương ứng với một sai sót không phản hồi, cũng như một sai sót loại trừ và bao gồm kết hợp nếu phần tử là có liên kết với một mục sai trong phân loại hoặc một giá trị số sai tương ứng với một sai sót đo lường; đối với các đặc điểm không bắt buộc, chúng tôi có thể cũng có các sai sót bao gồm. Các sai sót nhận dạng có thể cũng có tác động đến các đặc điểm quan sát, khi một sự thay đổi trong một đặc điểm được đăng ký cho một phần tử đã có trong bộ dữ liệu, chẳng hạn một thị trấn nơi cư trú của một sinh viên. Các sai sót có thể liên quan tới các mối quan hệ được xác định một cách tương tự.

Các phương pháp kiểm tra chất lượng sẵn có chủ yếu là: tìm kiếm các sai sót hiển nhiên, như các mã nhận dạng trùng nhau, kết nối với các nguồn dữ liệu khác, sử dụng các ràng buộc logic (bắt buộc hoặc không tương thích giữa các bảng kê khác nhau), tính toán thời gian trễ giữa thời điểm xuất hiện của sự kiện và thời điểm đăng ký của chúng.

Cho đến nay, chúng tôi đã xác định được một khung các chỉ tiêu chất lượng liên quan tới mức bao phủ của các tập hợp và định danh của các thành phần bằng cách kết nối các sai sót có thể có một cách phù hợp và các phương pháp kiểm tra chất lượng. Hiện tại, chúng tôi đang phân tích các sai sót có thể có về các đặc điểm và các mối quan hệ để xác định hai khung chỉ tiêu chất lượng khác liên quan tới tất cả các loại không trả lời, sai sót đo lường, sai sót quan hệ.

Nên nhớ rằng, các chỉ tiêu mà chúng tôi đã đề xuất là có thể tính toán riêng biệt cho mỗi tập hợp, đặc điểm và mối quan hệ trong mô hình nguồn dữ liệu hành chính, nhằm hỗ trợ một cách có hiệu quả cho bất kỳ việc sử dụng thống kê nào của thông tin đã thu thập bởi bất cứ người sử dụng nào quan tâm.

5. Công việc hiện tại và tương lai

Hiện tại chúng tôi đang tiến hành điều tra nguồn dữ liệu về một bộ các nguồn dữ liệu hành chính quan trọng đầu tiên do các cơ quan chính phủ trung ương sở hữu và các mẫu biểu hành chính có liên quan của chúng. Chúng tôi cũng đã lưu trữ trong hệ thống DARCAP các kết quả của cuộc điều tra đầu tiên về các nguồn dữ liệu hành chính do các cơ quan chính quyền địa phương sở hữu. Chúng tôi dự định mở rộng hoạt động điều tra thông qua việc xử lý ngày càng nhiều các nguồn dữ liệu hành chính và khởi động hoạt động giám sát các thay đổi của các nguồn dữ liệu hành chính và các dự án đổi mới.

Ngoài ra, chúng tôi cũng đang tiến hành công việc xác định các chỉ tiêu trong đa chiều về dữ liệu trên cơ sở phân tích cẩn thận các sai sót có thể có dựa vào các đối tượng có thể xuất hiện trong mô hình nguồn dữ liệu hành chính. Cuối cùng, Khung đánh giá chất lượng đối với các nguồn dữ liệu hành chính sẽ bao gồm các chỉ tiêu định tính để đánh giá chất lượng sơ bộ trong đa chiều về nguồn dữ liệu và siêu dữ liệu cùng với một bộ chỉ tiêu phong phú gồm cả chỉ tiêu định tính và định lượng để đánh giá chất lượng chuyên sâu và tùy chỉnh trong đa chiều về dữ liệu. Công việc này cũng là một gợi ý cho một hướng nghiên cứu thú vị về chất lượng dữ liệu.

Tài liệu tham khảo:

[1] M. Benedikt, P. Bohannon, G. Bruns Data Cleaning for Decision Support. First Int'l VLDB Workshop on Clean Databases (2006)

[2] G.J. Brackstone, Issues in the use of administrative records for statistical purposes, Survey methodology (1987)

[3] P. Daas, S. Ossen, M. Tennekes, L.. Zhang, C. Hendriks, K. Foldal Haugen, F. Cerroni, G. Di Bella, T. Laitila, A. Wallgren, BLUE – ETS Deliverable 4.2 - Report on methods preferred for the quality indicators of administrative data sources (2011)

[4] G. D'Angiolini, Manuale per la documentazione di archivi, moduli e dataset nel sistema DARCAP, Istat document (2013)

[5] G. D'Angiolini, P. , De Salvo, A. Passacantilli, Istat's new strategy and tools for enhancing statistical utilization of the available administrative databases, European conference on quality in official statistics, Vienna (2014)

[6] G. D'Angiolini, P. De Salvo, A. Passacantilli, E. Patruno, T. Saccoccio, C. De Rosa, E. Valente, DARCAP: a tool for documenting the information content and the quality of the available administrative databases, European conference on quality in official statistics, Vienna (2014)

[7] G. D'Angiolini, P. , De Salvo, A. Passacantilli, F. Pogelli, Framework per la qualità degli archivi amministrativi, Istat document (2013)

[8] United Nations Economic Commission for Europe (UNECE), Using Administrative and Secondary Sources for Official Statistics: A Handbook of Principles and Practices, United Nations Publication (2011)

[9] R. Vis-Visschers, J. Arends-Tóth, Checklist for the Quality evaluation of Administrative Data Sources, Discussion paper by Statistics Netherlands (2009)