VAI TRÒ CỦA DỮ LIỆU ĐẶC TẢ TRONG THỐNG KÊ

Cathryn Dippo, Văn phòng Thống kê Lao động, và Bo Sundgren, Cục Thống kê Thụy Điển

Tóm tắt:

Dữ liệu đặc tả đóng vai trò thiết yếu trong việc phát triển và sử dụng thông tin thống kê. Việc sản xuất thông tin yêu cầu dữ liệu và dữ liệu đặc tả phải được xem xét như một tổng thể chứ không phải cá thể riêng biệt; do đó, quản lý dữ liệu đặc tả phải được xem như một phần không thể tách rời của sản xuất thống kê. Hơn nữa, do dữ liệu đặc tả cung cấp cơ sở cho sự hiểu biết của con người về dữ liệu nên các khía cạnh nhận thức của nó cũng phải được đề cập.

Từ khóa: Thông tin, sử dụng, người sử dụng, phổ biến, quản lý.

Khái niệm "dữ liệu đặc tả" và các khái niệm liên quan như "siêu thông tin", "siêu cơ sở dữ liệu" và "hệ thống siêu thông tin" được định nghĩa lần đầu bởi Sundgren (1973). Một định nghĩa rất ngắn chỉ ra dữ liệu đặc tả là "dữ liệu về dữ liệu", nghĩa là loại dữ liệu cấp hai; cf Froeschl (1997). Các nhà khoa học máy tính thường giới hạn ý nghĩa của dữ liệu đặc tả trong các mô tả chính quy về cách dữ liệu được đánh và định dạng. Mặt khác, các nhà khoa học thông tin và các nhà phát triển hệ thống cũng nhấn mạnh tầm quan trọng của dữ liệu đặc tả là sự mô tả về ý nghĩa hoặc nội dung ngữ nghĩa của dữ liệu; những mô tả này có thể được cấu trúc thêm bớt cũng như tăng giảm mức độ chính quy; chúng thường là các bản mô tả tự do.

Thống kê nhà nước có lẽ là lĩnh vực đầu tiên nhận ra tầm quan trọng của dữ liệu đặc tả, nhưng thậm chí cũng đã mất khoảng hai thập kỷ (và một số dự án không thành công) cho đến khi thực sự đạt được một số tiến bộ. Trong những năm 1980 và 1990, Phòng Thống kê của Liên hợp quốc/ECE đã tổ chức một số cuộc họp về các hệ thống siêu thông tin (METIS). Một Hướng dẫn đã được hình thành như một kết quả hữu hình; Sundgren (1993). Năm 1993, Cơ quan Thống kê châu Âu (Eurostat) đã tổ chức một cuộc hội thảo về dữ liệu đặc tả thống kê thu hút rất nhiều sự chú ý cũng như một số lượng lớn người tham gia. Năm 1994, hội nghị Compstat đã tổ chức một phiên thảo luận về dữ liệu đặc tả thống kê; Sundgren (1994).

Chỉ đến gần đây các khu vực khác của xã hội, bao gồm cả khu vực kinh doanh tư nhân, mới cảm nhận được nhu cầu về một cách tiếp cận dữ liệu đặc tả toàn diện và nghiêm túc hơn. Ở một mức độ nào đó, những nhu cầu này đã được kích hoạt bởi sự quan tâm của các công ty và tổ chức trong việc tái sử dụng dữ liệu hoạt động của họ cho các mục đích chiến lược hơn bằng cách tổ chức dữ liệu trong cái được gọi là kho dữ liệu và sử dụng các kỹ thuật mới như On-Line Analytical Processing (OLAP) và khai thác dữ liệu. Việc sử dụng dữ liệu thứ cấp phát sinh từ các quá trình hoạt động của tổ chức rõ ràng có rất nhiều điểm chung với việc sản xuất và sử dụng các thống kê nhà nước (phần lớn dựa vào dữ liệu hoạt động do hệ thống hành chính của xã hội tạo ra). Trong

cả hai trường hợp dữ liệu đặc tả đều đóng vai trò thiết yếu giúp bù đắp khoảng cách về thời gian và không gian giữa nguồn và việc sử dụng dữ liệu; ví dụ: Người sử dụng dữ liệu lịch sử thậm chí còn chưa được sinh ra tại thời điểm mà dữ liệu họ quan tâm được thu thập và lưu trữ.

Các công cụ mạnh mẽ như cơ sở dữ liệu và Internet đã tăng cường liên kết và chia sẻ dữ liệu giữa các nhóm người sử dụng đang phát triển nhanh chóng thuộc nhiều loại khác nhau. Sự phát triển này đã làm nổi bật tầm quan trọng của dữ liệu đặc tả bởi những dữ liệu có sẵn một cách dễ dàng mà không có dữ liệu đặc tả phù hợp đôi khi có thể mang lại nhiều bất lợi hơn là lợi ích. Không nhà sản xuất dữ liệu nào muốn mạo hiểm việc người sử dụng, khi không có dữ liệu đặc tả phù hợp, vô tình hoặc cố ý làm sai lệch dữ liệu để phù hợp với mục đích của bản thân. Ngay cả khi dữ liệu được đi kèm với dữ liệu đặc tả hoàn chỉnh và có chất lượng cao thì việc sử dụng sai là không thể tránh khỏi hoàn toàn, nhưng nếu xảy ra, ít nhất cũng có cơ sở thông tin khách quan để tranh luận.

Các mô tả dữ liệu đặc tả vượt ra khỏi hình thức và nội dung thuần túy của dữ liệu. Dữ liệu đặc tả cũng được sử dụng để mô tả các thực tế hành chính về dữ liệu, chẳng hạn như ai là người đã tạo ra chúng, và chúng được tạo ra khi nào. Những dữ liệu đặc tả đó có thể tạo điều kiện cho việc tìm kiếm và định vị dữ liệu hiệu quả. Các loại dữ liệu đặc tả khác mô tả quá trình đằng sau dữ liệu, cách dữ liệu được thu thập và xử lý, trước khi chúng được liên kết hoặc lưu trữ trong cơ sở dữ liệu. Mô tả hoạt động của quá trình thu thập đằng sau dữ liệu (bao gồm, ví dụ, các câu hỏi cho người trả lời) thường hữu ích hơn là khái niệm trừu tượng về quan điểm "lý tưởng" đằng sau dữ liệu.

Có một số ví dụ về tiêu chuẩn dữ liệu đặc tả hiện có. Ví dụ, Dublin Core (xem http://purl.org/metadata/dublin_core) là tập hợp gồm 15 phần tử dữ liệu đặc tả nhằm mục đích tìm kiếm tài nguyên điện tử. Hiện có các tiêu chuẩn nội dung dữ liệu đặc tả cho nhiều đối tượng, bao gồm dữ liệu sinh học và không gian địa lý (http://www.fgdc.gov/metadata/contstan.html).

Việc phát triển các tiêu chuẩn chung cho dữ liệu đặc tả chính quy và mang tính kỹ thuật thường đỡ phức tạp hơn phát triển các tiêu chuẩn chung cho dữ liệu đặc tả ít chính quy và mang tính nội dung. Vì vậy, hầu hết các nỗ lực chuẩn hóa chung đều quan tâm đến khái niệm dữ liệu đặc tả chính quy của các nhà khoa học máy tính, trong khi việc tiêu chuẩn hóa dữ liệu đặc tả mang tính nội dung phụ thuộc nhiều hơn vào bối cảnh cụ thể hoặc tính đa dạng của dữ liệu, và do đó thường được thực hiện trong các lĩnh vực ứng dụng cụ thể, như sinh học, địa lý hoặc thống kê.

Nhưng thuật ngữ "dữ liệu đặc tả" có nghĩa là gì trong lĩnh vực thống kê nhà nước của chúng ta? Mặc dù định nghĩa trong từ điển - "dữ liệu về dữ liệu" - ngắn gọn và chính xác nhưng nó không bao hàm được tính cụ thể và ngữ cảnh cần thiết để truyền đạt ý nghĩa. Vì vậy, một vài năm trước, các thành viên của Diễn đàn Mở về Dữ liệu đặc tả đã phát triển định nghĩa sau:

"Dữ liệu đặc tả thống kê mô tả hoặc ghi lại dữ liệu thống kê, ví dụ như dữ liệu vi mô và dữ liệu vĩ mô, hoặc các dữ liệu đặc tả khác. Dữ liệu đặc tả thống kê tạo điều kiện cho việc chia sẻ, truy vấn và nhận thức về dữ liệu thống kê trong suốt thời gian tồn tại của dữ liệu". Định nghĩa này khá chính xác và ngắn gọn; hơn nữa, nó bao hàm một số bối cảnh. Nhưng liệu nó có đủ để chuyển tải ý

nghĩa sao cho những người sử dụng khác nhau, nhưng có thể hiểu được một cách tương đương? Có lẽ là không.

Để rõ ràng hơn việc định nghĩa dữ liệu đặc tả thống kê, chúng ta phải thảo luận về vai trò cơ bản của dữ liệu đặc tả. Dữ liệu đặc tả cung cấp bối cảnh cho dữ liệu; nếu không có dữ liệu đặc tả thì dữ liệu cũng không có ý nghĩa. Suy nghĩ theo toán học, dữ liệu kết hợp với dữ liệu đặc tả như một tập hợp sản xuất thông tin. Ví dụ, số 4.1 chỉ là một con số cho đến khi người ta nói rằng nó là ước tính chính thức về tỷ lệ thất nghiệp điều chỉnh theo mùa ở Mỹ trong tháng công bố tức tháng 5 năm 2000 của Văn phòng Thống kê Lao động vào ngày 3 tháng 6 năm 2000.

Tùy thuộc vào mục đích sử dụng con số 4.1 và kiến thức chung của bạn, các dữ liệu đặc tả nêu trên có thể đủ hoặc có thể không. Nếu bạn có kiến thức chung về thống kê và khái niệm về tính không chắc chắn, bạn có thể sẽ muốn biết thêm về khoảng tin cậy ước tính hoặc hệ số biến thiên. Nếu bạn là một nhà phân tích chính sách, bạn có thể sẽ muốn biết thêm về các định nghĩa chi tiết được sử dụng để phân loại những người có việc làm, thất nghiệp hoặc không trong lực lượng lao động. Nếu bạn có kiến thức về các phương pháp điều tra, bạn có thể sẽ muốn biết thêm về tỷ lệ phản hồi hoặc thậm chí cả hình thức và chuỗi các câu hỏi được sử dụng. Và đây mới chỉ là một sự khởi đầu nhỏ của các mô tả dữ liệu đặc tả cho con số 4.1 này.

Mục tiêu của chúng tôi trong bài luận này là để chỉ ra độ rộng của các định nghĩa gắn liền với thuật ngữ dữ liệu đặc tả trong bối cảnh thống kê nhà nước và các cơ quan sản xuất ra chúng. Trước tiên, chúng tôi trả lời các câu hỏi tại sao, ai, cái gì, khi nào, ở đâu và như thế nào của dữ liệu đặc tả thống kê. Chúng tôi chỉ ra rằng cần có một sự đa dạng về quan điểm để mô tả dữ liệu đặc tả thống kê. Trong phần 2 sẽ thảo luận về mối quan hệ giữa dữ liệu đặc tả và chất lượng. Trong hai phần cuối của bài luận, chúng tôi mô tả một số nỗ lực nghiên cứu đa ngành đang được thực hiện tại Văn phòng Thống kê Lao động và Cục Điều tra Dân số Mỹ và Cục Thống kê Thụy Điển. Kết quả của các dự án này sẽ giúp chúng tôi làm rõ định nghĩa dữ liệu đặc tả thống kê theo tính chất đa dạng về người sử dụng và cách sử dụng.

1. Định nghĩa dữ liệu đặc tả thống kê: Tại sao? ai? cái gì? khi nào? ở đâu? như thế nào?

Một cái nhìn sâu sắc qua nhiều năm phân tích, thảo luận và thử nghiệm chỉ ra rằng các vấn đề dữ liệu đặc tả thống kê cần phải được xử lý theo nhiều khía cạnh: Tại sao? ai? cái gì? khi nào? ở đâu? như thế nào? Đây sẽ là chủ đề của phần này. Một cái nhìn quan trọng khác cho rằng dữ liệu đặc tả của một tổ chức phải được coi là một hệ thống. Nếu không, nó sẽ không thể đáp ứng tất cả các nhu cầu quan trọng cho dữ liệu đặc tả với thời gian và nguồn lực sẵn có. Chủ đề này sẽ được xử lý trong phần 4.

1.1. Tại sao cần có dữ liệu đặc tả thống kê?

Dữ liệu đặc tả thống kê có nhiều mục đích. Mục đích đầu tiên và cơ bản nhất là để giúp người sử dụng diễn giải, hiểu và phân tích dữ liệu thống kê (dữ liệu vi mô, dữ liệu vĩ mô hoặc các dữ liệu đặc tả thống kê khác), kể cả khi họ không tham gia vào quy trình sản xuất đằng sau dữ liệu thống kê. Nói theo cách khác, dữ liệu đặc tả thống kê nên giúp người sử dụng chuyển đổi dữ liệu thống kê thành thông tin (Xem Hand (1993) để biết thêm về phần thảo luận xuất sắc "Dữ liệu, dữ liệu đặc tả và thông tin").

Thông tin chỉ có trong bộ não của con người và chỉ có thể được truyền đạt và chia sẻ giữa người với người bằng phương pháp trình bày dữ liệu. Thông tin có thể được trình bày thông qua dữ liệu theo nhiều cách: Ngôn ngữ nói hoặc viết, hình ảnh, trình bày theo cách thức điện tử, cử chỉ và ngôn ngữ cơ thể, v.v...

Dữ liệu đặc tả thống kê cũng giúp người sử dụng xác định, định vị và thu thập các dữ liệu thống kê có thể có liên quan đến mục đích thông tin của người sử dụng. Tìm kiếm thông tin thống kê, đặc biệt trong thời đại Internet, là một nhiệm vụ đã bắt đầu nhận được sự chú ý của cộng đồng khoa học thông tin (xem phần 3), nhưng rất nhiều vấn đề đã được phát hiện không có cách khắc phục dễ dàng. Một tập hợp các vấn đề quan trọng và dai dẳng liên quan đến khái niệm và thuật ngữ, ví dụ: Sự khác nhau những khái niệm của nhà sản xuất và người sử dụng và sự thật rằng thuật ngữ kỹ thuật có thể có nhiều định nghĩa mâu thuẫn (thậm chí trong một tổ chức). Dữ liệu đặc tả có thể giúp giải quyết những vấn đề này.

Dữ liệu đặc tả thống kê, đặc biệt dữ liệu đặc tả về quy trình, được sử dụng để mô tả và cung cấp ý kiến đánh giá liên quan đến các quy trình chi tiết và các bước xảy ra trong một chuỗi sản xuất thống kê, các quy trình hoạt động cũng như các quy trình thiết kế và quy hoạch. Các dữ liệu đặc tả này là không thể thiếu đối với người đánh giá quy trình sản xuất thống kê, bao gồm cả nhà sản xuất. Hầu hết các phương pháp cải tiến quy trình, bao gồm cả phương pháp của Deming (1982), được xây dựng dựa trên sự sẵn có của dữ liệu đặc tả hoặc dữ liệu về quy trình sản xuất. Mô tả quy trình cùng loại cũng có thể có giá trị cho mục đích giảng dạy và đào tạo, ví dụ như giới thiệu nhân viên mới hoặc cải thiện hiệu suất của nhân viên hiện có.

Dữ liệu đặc tả thống kê ghi lại các cuộc điều tra, hệ thống sản xuất và công cụ sản xuất hiện có theo cách mà các nhà thiết kế các cuộc điều tra và hệ thống sản xuất mới có thể sử dụng các tài nguyên và kinh nghiệm này. Do đó, dữ liệu đặc tả thống kê có thể được sử dụng trong các cơ sở tri thức và hệ thống dựa trên tri thức (ví dụ: Các hệ thống chuyên gia) và cho các mục đích quản lý tri thức, nói chung, liên quan đến việc thiết kế và vận hành các cuộc điều tra thống kê và hệ thống sản xuất. Ví dụ, việc xây dựng một bản câu hỏi khảo sát mới nhằm cung cấp thông tin về chăm sóc sức khoẻ cho trẻ em nghèo đói sẽ rất khó khăn nếu nhà phát triển không được tiếp cận bộ câu hỏi tiêu chuẩn để phân loại gia đình nghèo đói.

Dữ liệu đặc tả thống kê mô tả dữ liệu thống kê theo cách giúp nó có thể được xử lý bằng phần mềm máy tính. Những dữ liệu đặc tả thống kê này cần được cấu trúc và chuẩn hóa hơn là những dữ liệu đặc tả phục vụ người sử dụng dữ liệu thống kê.

Do đó, vai trò chính của dữ liệu đặc tả thống kê là tạo điều kiện thuận lợi và chia sẻ. Dữ liệu đặc tả rất cần thiết cho việc diễn giải số liệu thống kê. Các kiến thức mới thu được từ việc diễn giải thống kê có thể giúp cải tiến sản xuất (chi phí thấp hơn hoặc chất lượng tốt hơn) hoặc nâng cao hiểu biết về một số hiện tượng trên thế giới. Hơn nữa, dữ liệu đặc tả là dữ liệu cho các nhà thiết kế các cuộc điều tra. Việc biên soạn và lưu trữ của nó giúp các nhà thiết kế đưa ra các quy trình đo lường mới thông qua việc tái sử dụng hoặc học hỏi từ kinh nghiệm quá khứ.

1.2. Ai là người sử dụng dữ liệu đặc tả thống kê?

Phân theo nghĩa rộng, có hai đối tượng sử dụng dữ liệu đặc tả thống kê - nhà sản xuất và người sử dụng số liệu thống kê. Khi

nhắc đến các nhà sản xuất, chúng tôi muốn nói đến các nhà thiết kế các quy trình thu thập dữ liệu, những người thu thập, xử lý và đánh giá dữ liệu, ví dụ: Tất cả nhân viên trong các cơ quan thống kê và các đơn vị tham gia quá trình sản xuất số liệu thống kê đều đóng một vai trò nhỏ trong việc phát triển, sản xuất và đánh giá thống kê. Nhóm người sử dụng bao gồm các công chức, chính trị gia, nhà phân tích chính sách, nhà khoa học xã hội, nhà phân tích tài chính, học sinh và giáo viên các cấp, nhà báo và các công dân quan tâm.

Những người sử dụng khác nhau có những yêu cầu khác nhau về dữ liệu thống kê và dữ liệu đặc tả. Họ cũng khác nhau về tri thức và khả năng. Do đó, cần phải tính đến nhu cầu của nhiều đối tượng sử dụng khác nhau khi thiết kế dữ liệu đặc tả thống kê và các hệ thống dữ liệu đặc tả thống kê.

Các nhà sản xuất thống kê cũng có thể trở thành người sử dụng. Tuy nhiên, có một sự khác biệt quan trọng giữa "người sử dụng dữ liệu thống kê nội bộ" và người sử dụng dữ liệu thống kê bên ngoài cần phải được tính đến khi thiết kế dữ liệu đặc tả và hệ thống dữ liệu đặc tả. Một người sử dụng là nhà sản xuất có nghĩa là họ đã có sự hiểu biết liên quan cần thiết do đã tham gia vào việc thiết kế và vận hành các quy trình sản xuất thống kê. Do đó, một người sử dụng nội bộ là nhà sản xuất sẽ không có nhu cầu tương tự đối với dữ liệu đặc tả như người sử dụng bên ngoài - người đã không tham gia thiết kế và sản xuất dữ liệu thống kê.

1.3. Dữ liệu đặc tả thống kê là gì?

Một định nghĩa đơn giản và cơ bản của dữ liệu đặc tả là dữ liệu mô tả các dữ liệu khác. Do đó, dữ liệu đặc tả thống kê là dữ liệu mô tả dữ liệu thống kê. Dữ liệu đặc tả thống kê cũng có thể mô tả quy trình thu thập, xử lý hoặc tạo ra dữ liệu thống kê; những dữ liệu đặc tả này còn được gọi là dữ liệu quy trình. Cuối cùng, thuật ngữ "dữ liệu đặc tả thống kê" cũng có thể được sử dụng để mô tả các tài nguyên và công cụ có ích trong sản xuất thống kê, ví dụ: Phân loại và tiêu chuẩn thống kê, đăng ký và phương pháp thống kê, quy trình và phần mềm thống kê.

Do nhu cầu về dữ liệu đặc tả của người sử dụng rất khác nhau nên định nghĩa về một tập hợp dữ liệu đặc tả cần thiết và đầy đủ cũng biến đổi theo người sử dụng và cách sử dụng. Ví dụ: Người sử dụng tìm kiếm một con số quy định cho một hợp đồng hoặc hợp đồng thuê chỉ cần một tập hợp dữ liệu đặc tả tối thiểu - đủ để xác định con số cụ thể cần thiết. Mặt khác, các nhà thiết kế khảo sát đánh giá chất lượng dữ liệu từ các quy trình thu thập dữ liệu khác đòi hỏi rất nhiều dữ liệu đặc tả. Nghĩa là, ví dụ, nếu người trả lời được lựa chọn trong các phương thức trả lời (ví dụ: Mail, touchtone, internet) thì người đánh giá cần phải biết chi tiết về từng phương thức (ví dụ: Bố cục vật lý hoặc loại giọng nói, phương tiện định vị) đồng thời hiểu cách mỗi người trả lời tương tác với phương thức họ đã chọn (ví dụ: Phản hồi các dữ liệu bị mất, sao lưu hoặc dừng). Do không có mô hình chi tiết và nhân quả của phương thức sai số ngoài chọn mẫu, không có cách nào để xác định ít nhất đủ tập hợp dữ liệu đặc tả cần thiết để đánh giá các thiết kế thay thế hoặc định lượng chất lượng của một thiết kế cụ thể. Hậu quả là quan điểm về dữ liệu đặc tả của một nhà thiết kế hoặc người đánh giá bị hạn chế bởi khả năng xác định dữ liệu đặc tả liên quan của người đó.

Một ví dụ khác: Một nhà báo sẽ không có khả năng cũng như sự kiên nhẫn để lĩnh hội một lượng lớn dữ liệu đặc tả chi tiết và dựa trên lý thuyết; thay vào đó, anh/cô ấy

cần được cung cấp một dữ liệu đặc tả mạnh mẽ và dữ liệu được trình bày một cách dễ hiểu để tránh những diễn giải sai lệch tồi tệ nhất. Mặt khác, một nhà khoa học xã hội thậm chí có thể thắc mắc về các giả định của nhà sản xuất thống kê ban đầu và đưa ra các kết quả thống kê mới dựa trên các giả định khác. Kiểu người sử dụng thứ hai sẽ cần quyền truy cập vào tất cả các giả định và tình huống liên quan khác trong việc thu thập, chuẩn bị dữ liệu và các quy trình ước tính đã được thiết kế và vận hành bởi nhà sản xuất thống kê.

1.4. Dữ liệu đặc tả được sử dụng khi nào?

Việc sản xuất thông tin thống kê là một quá trình phức tạp. Không có bất kỳ nỗ lực thu thập dữ liệu mới hoặc sửa đổi dữ liệu hiện có nào đang diễn ra độc lập. Dữ liệu đặc tả dưới dạng kinh nghiệm đi trước, dù được ghi lại hay từ kiến thức cá nhân, được tất cả những người liên quan trong việc tạo và sử dụng thông tin thống kê từ giai đoạn lập kế hoạch ban đầu sử dụng thông qua việc sử dụng sản phẩm. Việc người thiết kế hoặc thực hiện một quy trình cụ thể được tiếp cận với nhiều dữ liệu đặc tả liên quan hơn sẽ cho khả năng cao cho ra các đặc điểm kỹ thuật hoặc kết quả với chất lượng tốt hơn. Dữ liệu đặc tả càng được liên kết với các mẩu dữ liệu hoặc thống kê cụ thể thì càng nhiều người tìm kiếm thông tin sẽ tìm thấy số lượng thích hợp và sử dụng nó ngay lúc này, vào ngày mai hoặc vài thế kỷ nữa tính từ bây giờ.

1.5. Dữ liệu đặc tả được sử dụng ở đâu?

Việc sử dụng từ "dữ liệu đặc tả", trái ngược với tài liệu, là một việc quan trọng. Từ tài liệu có nguồn gốc ngữ nghĩa theo phương thức dựa trên vật chất, chủ yếu là giấy nhưng cũng có thể là đá và kim loại (đồng xu). Hơn nữa, tài liệu thường gắn liền với hoạt động viết. Dữ liệu đặc tả như một phần của thông tin thống kê không giới hạn trong việc viết trên giấy. Bản đồ, đồ thị, ảnh chụp màn hình máy tính, chương trình máy tính, mã biên soạn, tài liệu scan và cơ sở dữ liệu đều là các thành phần của dữ liệu đặc tả. Một số chỉ tồn tại trong không gian ảo. Có thể chắc chắn rằng việc sử dụng dữ liệu đặc tả không giới hạn trong các toà nhà với bốn bức tường và một cái nóc (ví dụ: Văn phòng, phòng học, nhà ở); những người thu thập dữ liệu tại hiện trường về cây trồng, chất lượng nước và không khí, cá và động vật hoang dã, v.v... là những người sử dụng dữ liệu đặc tả nặng. Khi mà chúng ta đang tiến gần hơn đến môi trường kỹ thuật trong sản xuất và sử dụng thông tin thống kê thì những nơi mà dữ liệu đặc tả được sử dụng sẽ chỉ bị giới hạn bởi các điều kiện vật lý cản trở việc sử dụng máy tính.

1.6. Dữ liệu đặc tả được sử dụng như thế nào?

Dữ liệu đặc tả là một công cụ giúp hiểu và lĩnh hội. Nó cung cấp ý nghĩa cho các con số. Ở mức cơ bản nhất, dữ liệu đặc tả cho phép diễn giải một con số. Điều đó có nghĩa con số 4.1 không có ý nghĩa gì nếu không có dữ liệu đặc tả. Dữ liệu đặc tả cũng là một công cụ giúp diễn giải, sử dụng dữ liệu để đưa ra các suy luận và tạo điều kiện cho việc thu thập kiến thức mới. Dữ liệu đặc tả giúp người tìm kiếm thông tin tìm ra dữ liệu và xác định xem liệu nó có phù hợp với vấn đề đang đặt ra hay không, tức là xác định tính thích hợp của nó cho việc sử dụng. Dữ liệu đặc tả giúp các nhà thiết kế phát triển các quá trình mới tốt hơn và giúp người thực hiện đáp ứng các quy định của quá trình, ví dụ: Bằng cách thông báo về các phương pháp và công cụ liên quan, cách thức sử dụng và kinh nghiệm từ các ứng dụng trước đó.

Dữ liệu đặc tả cũng là một công cụ để sửa đổi các quá trình làm việc nhằm nâng cao chất lượng dữ liệu hoặc giảm chi phí. Ghi chép các thủ tục liên quan đến những cái có hiệu quả và những cái không hiệu quả sẽ giúp những người khác đưa ra các lựa chọn tốt hơn và tránh các sai lầm. Việc tái sử dụng dữ liệu đặc tả từ những lần thực hiện trước đó có thể giúp giảm chi phí (ví dụ: Các công cụ thu thập dữ liệu điện tử, phần mềm cho chọn mẫu hoặc tính tỷ trọng, biên soạn tài liệu sổ tay hướng dẫn người phỏng vấn).

1.7. Kết luận

Tóm lại, vai trò của dữ liệu đặc tả khá phổ biến. Bất kỳ và tất cả các định nghĩa có thể phù hợp với các trường hợp cụ thể. Vì vậy, làm thế nào để chúng ta có thể quyết định tập hợp dữ liệu đặc tả nào thích hợp cho một trường hợp cụ thể? Hãy nghiên cứu. Trong hai phần cuối của bài luận này, chúng tôi sẽ mô tả các dự án nghiên cứu gần đây và đang được tiến hành nhằm thông tin cho nhà sản xuất về quá trình cung cấp dữ liệu đặc tả cho người sử dụng. Nhưng trước tiên, hãy xem một ví dụ minh họa và một cuộc thảo luận về dữ liệu đặc tả và chất lượng.

2. Dữ liệu đặc tả và chất lượng

Dữ liệu đặc tả đóng một vai trò quan trọng trong việc liên kết phép đo khảo sát và cải tiến chất lượng quy trình (Dippo 1997). Có mối quan hệ hai chiều giữa dữ liệu đặc tả và chất lượng. Một mặt, dữ liệu đặc tả mô tả chất lượng thống kê. Mặt khác, dữ liệu đặc tả chính là thành phần chất lượng giúp cải thiện tính sẵn có và khả năng tiếp cận của dữ liệu thống kê.

2.1. Thống kê chất lượng tốt có những đặc trưng gì?

Thứ nhất, thống kê tốt phải phù hợp với vấn đề của người sử dụng. Điều này phải được đánh giá bởi người sử dụng trong một tình huống sử dụng cụ thể. Một thống kê tương tự có thể rất phù hợp trong một tình huống sử dụng nhưng lại ít nhiều không liên quan trong một tình huống sử dụng khác. Tính liên quan là một vấn đề khó trong thống kê nhà nước, vì các thống kê được sản xuất cho nhiều người sử dụng và mục đích sử dụng trong một thời gian dài, được gọi là thống kê đa năng. Nhằm cho phép nhiều người sử dụng, hiện tại và trong tương lai, đánh giá mức độ liên quan của các thống kê nhất định trong các tình huống sử dụng khác nhau, cần cung cấp rất nhiều dữ liệu đặc tả về ý nghĩa của dữ liệu được thu thập ban đầu (có thể từ các nguồn khác nhau) và cách dữ liệu này được xử lý trong quy trình sản xuất ban đầu.

Thứ hai, thống kê tốt phải đúng một cách hợp lý (chính xác), nghĩa là chúng không được có lỗi nghiêm trọng. Tối thiểu phải biết (và ghi lại) nguồn sinh lỗi, và khi có thể, cần ước tính kích cỡ lỗi. Tăng cường độ chính xác của dữ liệu đặc tả phải là một phần không thể tách rời trong chương trình làm việc của các nhà sản xuất thống kê.

Thứ ba, thống kê tốt phải kịp thời và cập nhật. Dữ liệu đặc tả được quản lý tốt có thể giúp giảm thời gian trễ giữa công tác thiết kế và công tác thực hiện bằng cách giảm thời gian phát triển thông qua việc sử dụng lại (ví dụ: Các thành phần phần mềm, câu hỏi, qui trình). Hơn nữa, bằng cách quản lý dữ liệu đặc tả như một phần của quá trình sản xuất, tính kịp thời và chất lượng của các sản phẩm phổ biến có thể được cải thiện.

Thứ tư, thống kê tốt phải được xác định rõ để dễ dàng so sánh với các thống kê khác mà người sử dụng cần trong một tình huống sử dụng nhất định, ví dụ: Các thống kê tương tự liên quan đến một vùng/quốc gia

khác, một khoảng thời gian hoặc một ngành công nghiệp. Khả năng so sánh chỉ có thể được xác nhận thông qua dữ liệu đặc tả chính xác. Do đó, cần quản lý dữ liệu đặc tả về thay đổi hệ thống phân loại và địa lý cùng với các liên kết giữa dữ liệu và dữ liệu đặc tả. Nếu không, người sử dụng có thể hiểu sai rằng những khác biệt là do thay đổi trong hiện tượng được đo lường chứ không phải là do khác biệt trong phạm vi địa lý hoặc phân loại.

Thứ năm, thống kê tốt phải có sẵn, dễ truy xuất, diễn giải và phân tích. Dữ liệu đặc tả tốt tạo điều kiện khám phá tài nguyên, đặc biệt là thông qua internet. Do đó, các tiêu chuẩn nội dung dữ liệu đặc tả như Dublin Core và Sáng kiến Tài liệu Dữ liệu (DDI) là rất cần thiết. Ủy ban DDI đã đưa ra định nghĩa về Định dạng tài liệu (DTD) để "đánh dấu" các tập lệnh cho các tập hợp dữ liệu vi mô. DTD sử dụng Ngôn ngữ Đánh dấu Mở rộng (XML) - đó là ngôn ngữ của một vùng được đánh dấu phổ biến hơn, SGML. DDI đã được sử dụng trong các dự án quốc tế lớn như dự án Công cụ và Nguồn Khoa học Xã hội Mạng lưới châu Âu (NESSTAR). (Xem http://www.icpsr.umich.edu/DDI/intro.html.)

2.2. Vai trò của dữ liệu quy trình trong các tuyên bố chất lượng

Việc tuyên bố chất lượng dữ liệu thống kê không dễ dàng như việc tuyên bố chất lượng hàng hóa vật chất, ví dụ như một chiếc xe hơi. Trong trường hợp sau, các quy mô thứ tự (từ 1 đến 5) thường được sử dụng để chỉ ra chất lượng tốt/xấu của một số "tính năng" quan trọng của hàng hoá. Đối với dữ liệu thống kê, không có nhiều tính năng tuyệt đối có thể được đánh giá theo cùng một cách cho tất cả người sử dụng và mục đích sử dụng. Có nhiều tính năng khác, phải được đánh giá bởi người sử dụng, tính đến mục đích sử dụng cụ thể trong tầm tay. Nhằm cho phép người sử dụng thực hiện đánh giá trong một tình huống sử dụng cụ thể, nhà sản xuất dữ liệu và dữ liệu đặc tả thống kê phải cung cấp các mô tả khá chi tiết về các quá trình đằng sau dữ liệu, ví dụ:

- Những câu hỏi nào đã được đưa ra, và chúng được hỏi như thế nào?
- Các câu trả lời đã được kiểm tra khả năng lỗi và nhầm lẫn như thế nào?
- Những quy tắc nào đã được sử dụng để nhập và mã hóa dữ liệu?
- Sự khác biệt giữa các khái niệm mục tiêu và các khái niệm được đo lường là gì?
- Các trường hợp không phản hồi được xử lý như thế nào?
- Những giả định ước tính và thủ tục ước tính nào đã được sử dụng?

Do đó, việc sản xuất dữ liệu đặc tả thống kê chất lượng tốt đòi hỏi phải có sự cam kết từ nhà sản xuất thống kê, một cam kết gắn liền với cam kết sản xuất dữ liệu chất lượng tốt.

3. Các hoạt động nghiên cứu tại Văn phòng Thống kê Lao động[1]: Nghiên cứu người sử dụng

Các hoạt động nghiên cứu liên quan đến dữ liệu đặc tả tại Văn phòng Thống kê Lao động đang tập trung vào người sử dụng. Các hoạt động nghiên cứu bao gồm nghiên cứu người sử dụng và tổ chức tri thức bởi các nhà khoa học thông tin, nghiên cứu nhận thức bởi các nhà tâm lý học nhận thức và

[1] *John Bosley và Fred Conrad của Văn phòng Thống kê Lao động đã tham gia chuẩn bị phần này của bài luận.*

kiểm tra khả năng sử dụng bởi các nhà tâm lý học nhân tố con người.

3.1. Nghiên cứu người sử dụng

Việc hiểu được người sử dụng của bạn là ai cũng như mong muốn và kiến thức chuyên môn của họ là rất quan trọng đối với việc thiết kế một trang web có thể sử dụng và hữu ích có đủ dữ liệu đặc tả nhằm làm hài lòng người sử dụng. Trong vài năm gần đây, Marchionini và Hert (1997) đã nghiên cứu người sử dụng của ba trang web thống kê nhà nước: Văn phòng Thống kê Lao động (BLS), Khảo sát Dân số hiện tại (kết hợp giữa Cục điều tra dân số và BLS) và FedStats (liên doanh 14 Cơ quan thống kê là thành viên của Hội đồng Liên ngành về Chính sách Thống kê). Trong năm đầu tiên, mục tiêu của họ là xác định những người đã sử dụng các trang này, những loại công việc họ đã thực hiện trên trang web, những chiến lược họ đã sử dụng để tìm thông tin thống kê và đưa ra các khuyến nghị giúp cải tiến thiết kế. Họ đã sử dụng nhiều phương pháp khác nhau để tiến hành điều tra, trong đó có nhiều phương pháp tương tự như các phương pháp đã được các nhà khoa học hành vi sử dụng trong việc phát triển và thử nghiệm các bảng khảo sát, ví dụ: Các cuộc phỏng vấn, các nhóm tập trung và phân tích nội dung. Một kết quả nghiên cứu của họ là sự phát triển của phép phân loại các hành động của người sử dụng dựa trên truy vấn.

Một khuyến nghị quan trọng rút ra từ nghiên cứu này là sự cần thiết phải xem xét lại giao diện trang web BLS (thứ phản ánh tổ chức theo định hướng chương trình của BLS) để đáp ứng tốt hơn nhu cầu của người sử dụng với chuyên môn và nhu cầu đa dạng. Dựa trên các kết quả này, Marchionini (1998) tiến hành thiết kế và thử nghiệm giao diện thay thế. Các thiết kế lặp lại được dựa trên bốn nguyên tắc thiết kế: Coi người sử dụng là trung tâm, các giao diện khác nhau cho các nhóm người sử dụng khác nhau (không phải các giao diện thích ứng với người sử dụng cá nhân), thông tin phong phú và hiển thị không gian.

Hert (1998), trong nghiên cứu tiếp theo của mình về người sử dụng thông qua các cuộc phỏng vấn với các bên trung gian, đã tìm ra một số vấn đề liên quan đến dữ liệu đặc tả, ví dụ như thiếu kiến thức về cách dữ liệu được thu thập, thiếu khả năng toán học và thống kê và thiếu hiểu biết liên quan đến quá trình nghiên cứu hoặc bản chất của lỗi. Về mặt lịch sử, các bên trung gian đã cung cấp các kiến thức cần thiết để giải quyết những thiếu sót này; tuy nhiên, để phổ biến qua internet, trang web phải cung cấp các dịch vụ dựa trên dữ liệu đặc tả đang được cung cấp bởi các bên trung gian. Các dịch vụ đó có thể là hướng dẫn, kịch bản và trợ giúp trực tuyến dựa trên ngữ cảnh.

3.2. Kiểm tra khả năng sử dụng

Việc kiểm tra phòng thí nghiệm khả năng sử dụng để đánh giá giao diện máy tính của con người phải được coi là một thành phần thiết yếu trong bất kỳ nỗ lực phát triển hệ thống nào. Công tác này mở rộng đến cả việc thiết kế các trang web thống kê và các cơ sở dữ liệu thống kê khác.

Công tác kiểm tra khả năng sử dụng các trang web thống kê thường bao gồm việc yêu cầu một nhóm người tham gia thử nghiệm thực hiện một số nhiệm vụ liên quan đến dữ liệu, chẳng hạn như lựa chọn và tải một hoặc nhiều biến thể bằng cách vận dụng các đối tượng xuất hiện trên một hoặc nhiều giao diện có thể truy cập tại trang web đang được giám sát. Trong các cuộc kiểm tra giao diện "thử nghiệm" không chính thức trước đó, những người tham gia có thể chỉ đơn

giản khám phá (các) giao diện và cho ý kiến về độ hữu ích của các tính năng khác nhau, cách họ muốn sắp xếp các đối tượng giao diện và mức độ cấu trúc trang web hợp lý với họ. Những đánh giá này được gửi lại cho các nhà thiết kế web, những người sau đó sẽ tiến hành điều chỉnh thiết kế và thực hiện lại các bài kiểm tra khả năng sử dụng. Khi thiết kế đến hạn, những người tham gia có thể thực hiện các nhiệm vụ đã được cấu trúc (lên kịch bản) nhằm thu thập dữ liệu hiệu suất có khả năng giám sát phân tích, ví dụ như thời gian trung bình mà một nhóm người sử dụng cần để hoàn thành một kịch bản được giao, tỷ lệ thời gian người sử dụng truy xuất dữ liệu mục tiêu.

Máy quay video có thể được sử dụng để ghi lại khuôn mặt của đối tượng (và các nhận xét bằng lời) và sự tương tác của họ với bàn phím và chuột, cuốn băng ghi hình sau đó sẽ được tích hợp với video từ màn hình hiển thị máy trạm. Các nhà nghiên cứu có thể quan sát thử nghiệm trực tiếp hoặc xem các đoạn video, thường là các đoạn video đã chỉnh sửa, để làm nổi bật các vấn đề thiết kế quan trọng. Thông thường sẽ có một buổi thảo luận sau khi hoàn thành nhiệm vụ để nhóm thử nghiệm có thể tìm hiểu các vấn đề chưa được giải quyết thỏa đáng qua dữ liệu quan sát với những người tham gia. Ví dụ, những người tham gia có thể được hỏi về những gián đoạn khó hiểu trong khi thực hiện nhiệm vụ được giao để ghi nhận các quan điểm chủ quan của họ về các nguyên nhân gây gián đoạn.

Một cách tiếp cận khác (không cần phải thực hiện trong phòng thí nghiệm) là kiểm tra xem người sử dụng nghĩ gì về các thông tin mà trang web có ý định cung cấp. Một cách để làm điều này là yêu cầu người sử dụng sắp xếp các thẻ có tên các chủ đề trang web thành các chồng và kiểm tra bằng mắt hoặc phân cụm phân tích các chồng này để xác định mức độ tương ứng giữa quan điểm của người sử dụng và các nhà thiết kế về cách thức thông tin được cấu trúc.

Các nhà nghiên cứu nhân tố con người tại BLS đã tiến hành một số thử nghiệm về khả năng sử dụng trên BLS internet và các trang mạng nội bộ, trang CPS và các giao diện dựa trên người sử dụng nguyên mẫu được thiết kế bởi Marchionini (1999) để thay thế cho trang chủ BLS hiện tại. Công tác này bao gồm việc sử dụng dữ liệu đặc tả trong phạm vi họ đánh giá khả năng của người sử dụng để lấy các tài liệu mô tả dữ liệu thực tế. Tuy nhiên, họ vẫn phải làm nhiều hơn nữa để cải thiện cấu trúc của các trang web nhằm giúp người sử dụng dễ dàng định vị và thu thập dữ liệu số. Cấu trúc của một trang web và việc thiết kế các trang web là các loại dữ liệu đặc tả; chúng cung cấp thông tin về vị trí và ngữ cảnh của dữ liệu.

3.3. Nghiên cứu nhận thức

Các thử nghiệm phòng thí nghiệm bao gồm các cuộc phỏng vấn yêu cầu nói ra suy nghĩ và các phương pháp nghiên cứu nhận thức khác có thể và nên được sử dụng để hiểu về các chiến lược của người dùng trang web trong việc thu thập thông tin và hiểu các thuật ngữ đang được sử dụng. Chỉ vậy thôi sao, liệu lượng dữ liệu đặc tả có được cung cấp đủ để giúp người sử dụng thu thập và hiểu những gì đang được trình bày?

Hert đã tiến hành một thí nghiệm với bốn biến của một chỉ số chủ đề A-Z. Cô nhận thấy rằng cấu trúc của các công cụ tổ chức hiện tại và thuật ngữ được sử dụng trong các công cụ này rất có vấn đề đối với người sử dụng. Do đó, cô đề nghị tăng cường chỉ số bằng cách thêm vào nhiều mục cho một chủ đề và các mục này sử dụng ngôn ngữ phổ biến.

Các nhà nghiên cứu BLS và Cục điều tra dân số (Census) đã tiến hành một số nghiên cứu thí điểm nhằm hướng tới phát triển các quy ước cho các tên gọi ngắn để khảo sát các biến. Các quy tắc và hướng dẫn xây dựng một quy ước đặt tên được cung cấp trong Phần 5 của ISO 11179, và một quy ước cụ thể được cung cấp trong một phụ lục thông tin đang được xem xét trong nghiên cứu này. Tuy nhiên, quy ước đặt tên đó được phát triển từ một mô hình dữ liệu không rõ ràng trong nghiên cứu về cách một phạm vi phổ rộng người sử dụng dữ liệu diễn giải các tên gọi hoặc các thành phần của chúng. Công tác thí điểm bao gồm việc tạo ra các tên biến ngắn dựa trên ngôn ngữ trừu tượng từ một câu hỏi khảo sát và các câu trả lời hợp lệ. Các quy tắc ngữ nghĩa và ngữ pháp khác nhau đã được sử dụng để tạo ra các tên biến, và một nhóm nhỏ (N = 15) người sử dụng dữ liệu đã xếp hạng mức độ diễn đạt hiệu quả ý nghĩa của câu hỏi tương ứng của biến. Phân tích những kết quả sơ bộ này cho thấy các biến đặt tên ngữ nghĩa hoặc ngữ pháp ảnh hưởng không nhiều đến khả năng hiểu câu hỏi. Mặt khác, thậm chí cả bài kiểm tra nhỏ này cũng chỉ ra rằng khó có thể tìm ra tên viết tắt "tốt" cho một số loại biến nhất định. Nghiên cứu sâu hơn sẽ tập trung vào việc kiểm tra và chỉnh lý kết quả sơ bộ về sau. Nghiên cứu bổ sung này cũng sẽ được thiết kế lại để người tham gia tích cực xây dựng tên cho các biến, sử dụng các quy trình được phát triển bởi các nhà biên soạn từ điển để xây dựng các từ điển thay vì chỉ đơn thuần phản ứng với các tên biến được tạo ra bởi nhóm nghiên cứu. Cách tiếp cận này được đưa ra bởi một nhà khoa học thông tin khác đang làm việc với BLS là Stephanie Haas (1999) của tổ chức UNC-Chapel Hill.

Một dự án đang thực hiện khác của Carol Hert và các nhân viên BLS và Census là nhằm xác định số lượng dữ liệu đặc tả tối thiểu mà người sử dụng dữ liệu cần để đưa ra các quyết định chính xác và tự tin về mức độ phù hợp của một biến khảo sát cụ thể với một phân tích đã được đưa vào kế hoạch. Công tác chuẩn bị cho nghiên cứu này bao gồm việc tạo ra một loạt các kịch bản nghiên cứu hợp lý có thể thực hiện bằng cách sử dụng dữ liệu từ một tập hợp dữ liệu BLS/Census phổ biến, Khảo sát Dân số hiện tại (CPS). Sau đó, một nhóm người kỳ cựu sử dụng dữ liệu CPS đã đạt được sự nhất trí về tập con của các biến CPS được coi là "tốt nhất" để trích xuất nhằm thực hiện một phân tích có thể đáp ứng mục tiêu của từng kịch bản. Những người sử dụng chuyên gia này cũng đã đề cử một tập hợp lớn hơn với tên gọi tương tự nhưng ít phù hợp hơn với các biến CPS cho mỗi kịch bản để bắt buộc những người tham gia nghiên cứu chọn các biến tốt nhất từ danh sách các dữ liệu cạnh tranh. Trong nghiên cứu thực tế, lượng dữ liệu đặc tả dành cho người tham gia về các danh sách biến sẽ được đặt ở ba cấp độ - tối thiểu, trung bình và dồi dào. Sự lựa chọn biến số "tốt nhất" của những người tham gia sẽ được so sánh theo cả ba cấp độ này để xác định tác động của việc có thêm dữ liệu đặc tả đến việc cải thiện lựa chọn chính xác so với đánh giá của các chuyên gia. Những người tham gia sẽ cung cấp dữ liệu về các yếu tố dữ liệu đặc tả mà họ thấy hữu ích nhất trong việc phân biệt các biến có liên quan với các lựa chọn kém phù hợp. Xu hướng nghiên cứu này sẽ tiếp tục với các nghiên cứu bổ sung nhằm xác định xem "điểm giảm dần lợi nhuận" cho dữ liệu đặc tả có thể được thiết lập gần đúng hay không, ngoài những thông tin bổ sung không cải thiện sự lựa chọn của người sử dụng trong các biến số cạnh tranh.

3.4. Kết luận

Như đã lưu ý trong phần 1.1, mục đích đầu tiên và quan trọng nhất của dữ liệu đặc tả là hỗ trợ người sử dụng dữ liệu thống kê. Nếu nhà sản xuất thống kê muốn xác định liệu mình có đang cung cấp dữ liệu đặc tả khả dụng, hữu ích và đủ hay không, thì họ phải tham gia vào các nghiên cứu người sử dụng. Các khía cạnh nhận thức của dữ liệu đặc tả và hầu hết các thành phần của các sản phẩm thống kê phổ biến (ví dụ: văn bản, bảng, biểu đồ, đồ thị, bản đồ) là một lĩnh vực đáng được các nhà sản xuất thống kê chú ý nhiều hơn nữa.

4. Các hoạt động nghiên cứu tại Cục Thống kê Thụy Điển: Quản lý dữ liệu đặc tả tích hợp

Rõ ràng là dữ liệu đặc tả thống kê có rất nhiều người sử dụng và cách sử dụng khác nhau cũng như quan trọng. Không còn nghi ngờ gì về nhu cầu và sự cần thiết phải có dữ liệu đặc tả thống kê.

Phía nguồn cung có nhiều vấn đề hơn. Ai sẽ là người cung cấp dữ liệu đặc tả khẩn cấp khi cần? Nhà cung cấp dữ liệu đặc tả thống kê cuối cùng không ai khác chỉ có thể là nhà sản xuất những dữ liệu thống kê sẽ được mô tả. Tuy nhiên, các nhà sản xuất thống kê không phải lúc nào cũng có động lực để sản xuất cả dữ liệu đặc tả. Trước hết, họ (thường) giả định (sai lầm) rằng họ biết tất cả những gì đáng để biết về những thông tin thống kê họ sản xuất. Những kiến thức này nằm trong não của họ và họ có rất ít lý do để ghi lại chúng cho những người khác có thể chia sẻ ở những nơi khác hoặc vào những thời điểm về sau. "Nếu có ai đó muốn biết thêm về các thống kê này, họ có thể thoải mái hỏi tôi" là một tuyên bố khá phổ biến của các nhà sản xuất thống kê. Tuy nhiên lời nhận xét này không tính đến thực tế rằng ngay cả các nhà sản xuất thống kê cũng chỉ có trí nhớ giới hạn và họ không phải lúc nào cũng sẵn sàng để phục vụ người sử dụng. Kể cả khi không tính đến việc này thì cũng khá phi thực tế để yêu cầu người sử dụng liên hệ với nhà sản xuất khi cần một số thông tin về ý nghĩa hoặc chất lượng của một số dữ liệu thống kê nhất định.

Điều quan trọng là phải tìm cách khuyến khích các nhà sản xuất thống kê cung cấp dữ liệu đặc tả tốt đi kèm với dữ liệu thống kê mà họ tạo ra. Cần có cả củ cà rốt và cây gậy *(một kiểu chính sách ngoại giao trong quan hệ quốc tế)*. Củ cà rốt có thể được dùng để chứng minh cho các nhà sản xuất rằng trong thực tế sẽ có cả các tình huống khi các nhà sản xuất thống kê cũng cần dữ liệu đặc tả, ví dụ như khi sắp thiết kế một cuộc điều tra thống kê mới và khi dữ liệu đặc tả (ví dụ: Loại và nhãn hiệu) cần phải được cung cấp cho một phần mềm. Cây gậy có thể được coi là một tiêu chuẩn tài liệu cần phải tuân thủ.

Đương nhiên, tiêu chuẩn đó nên được hỗ trợ bởi một công cụ thân thiện với người sử dụng để giúp công việc trở nên dễ dàng nhất có thể cho nhà sản xuất. "Sử dụng các công cụ chứ không phải các quy tắc" là một khẩu hiệu thường xuất hiện trong một số văn phòng thống kê[2].

Theo một cách lý tưởng thì các lưu trữ và hệ thống dữ liệu đặc tả khác nhau cùng tồn tại trong một tổ chức nên đóng vai trò là các thành phần tương thích của một tổng thể hoàn chỉnh, tức là một hệ thống siêu thông tin được tích hợp tốt về mặt khái niệm và kỹ thuật và không dư thừa có thể đáp ứng mọi

[2] *Chúng tôi tin rằng, người tạo ra khẩu hiệu này là Wouter Keller, Cục Thống kê Hà Lan*

nhu cầu dữ liệu đặc tả quan trọng của tổ chức và người sử dụng giúp tối thiểu hóa nỗ lực của con người. Trên thực tế, điều này có nghĩa là cần có một khuôn khổ khái niệm và cơ sở hạ tầng kỹ thuật chung cho tất cả các lưu trữ và hệ thống dữ liệu đặc tả. Việc thu thập một dữ liệu đặc tả nhất định sẽ diễn ra khi dữ liệu đặc tả xuất hiện tự nhiên lần đầu trong một quá trình thiết kế hoặc sản xuất. Không nên thu thập lại dữ liệu đặc tả nếu đã có dữ liệu đặc tả tương tự và nếu một dữ liệu đặc tả nhất định có thể bắt nguồn từ dữ liệu đặc tả hiện có thì công tác này cần được thực hiện tự động bằng các công cụ phần mềm. Các phần mềm và ứng dụng cần dữ liệu đặc tả phải có khả năng lấy và chuyển hóa dữ liệu đặc tả nhiều nhất có thể từ những nguồn hiện có bằng các công cụ tự động. Cần có một bộ phận dữ liệu đặc tả không dư thừa chủ chốt giúp chuyển hóa các dữ liệu đặc tả khác nhằm phục vụ các mục đích khác nhau trong một tổ chức thống kê và cho tất cả các hạng mục người sử dụng thống kê quan trọng, cả những người sử dụng cao cấp như các nhà nghiên cứu và những người sử dụng bình thường như các nhà báo và người dân trên phố.

Nói cách khác, nhằm tạo điều kiện cho các công việc liên quan đến dữ liệu đặc tả của các nhà sản xuất thống kê ở mức tốt nhất có thể, người ta nên cung cấp các công cụ giúp thu thập dữ liệu đặc tả khi chúng xuất hiện lần đầu và một hệ thống quản lý dữ liệu đặc tả tích hợp tạo điều kiện cho việc chuyển đổi và tái sử dụng dữ liệu đặc tả hiện có cho các mục đích khác: Các giai đoạn khác trong chuỗi sản xuất, các sản phẩm phần mềm khác, các quy trình thống kê khác.

Khoảng năm 1990, Cục Thống kê Thụy Điển đã phát triển một khuôn khổ khái niệm tích hợp để mô tả có hệ thống và đầy đủ các cuộc khảo sát thống kê và sổ đăng ký quan trắc thống kê theo nghĩa rộng, bao gồm sổ đăng ký, hệ thống sản xuất thống kê dựa trên cơ sở các nguồn hành chính và hệ thống thống kê phụ như tài khoản quốc gia. Khuôn khổ khái niệm, được gọi là SCBDOK, được phát triển bởi Bengt Rosén (giáo sư thống kê) và Bo Sundgren (giáo sư tin học); xem Rosén & Sundgren (1991).

Khuôn khổ khái niệm SCBDOK sau đó được sử dụng làm cơ sở thiết kế một số lưu trữ và hệ thống dữ liệu đặc tả của Cục Thống kê Thụy Điển:

- Một hệ thống, còn được gọi là SCBDOK, nhằm ghi chép các đăng ký quan trắc cuối cùng, được lưu trữ để các nhà nghiên cứu và những người khác sử dụng trong tương lai. Hệ thống được dựa trên các mẫu tài liệu. Hầu hết các dữ liệu đặc tả theo yêu cầu của mẫu đều là các dữ liệu đặc tả văn bản tự do nhưng các tập con của dữ liệu đặc tả, theo định nghĩa của mẫu phụ METADOK, đều được định dạng như các bảng quan hệ, và cũng có thể được sử dụng tự động bởi các sản phẩm phần mềm được phát triển trong nội bộ hoặc thương mại nhằm sản xuất thống kê.

- Một khái niệm chất lượng tiêu chuẩn đã được phát triển dựa trên khuôn khổ khái niệm SCBDOK và được sử dụng để sản xuất các bản khai chất lượng tiêu chuẩn cho tất cả các văn phòng thống kê nhà nước ở Thụy Điển. Cũng giống như các tài liệu của SCBDOK, các bản khai chất lượng được cấu trúc bằng phương thức mẫu. Như một bước đầu trong việc sản xuất các bản khai chất lượng cao hơn, các bản mô tả sản phẩm ngắn gọn (khoảng 10 trang) đã được tạo ra, nhưng ý định bây giờ là tăng mức độ tham vọng.

• Cùng với lý thuyết phân loại được thêm vào, SCBDOK cũng đã hình thành cơ sở khái niệm cho cơ sở dữ liệu thống kê trung ương của Cục Thống kê Thụy Điển nhằm mục đích bao quát tất cả các tiêu chuẩn quốc gia và quốc tế, bao gồm cả các phiên bản hiện tại và lịch sử cũng như các phiên bản Thụy Điển và quốc tế (của các phân loại quốc tế).

• SCBDOK, METADOK, các bản khai chất lượng và cơ sở dữ liệu phân loại là tất cả các thành phần tích hợp của hệ thống dựa trên Internet nhằm phổ biến tất cả các thống kê Thụy Điển chính thức, "Cơ sở dữ liệu Thống kê Thụy Điển" được đưa vào hoạt động từ ngày 01/01/1997 và hiện đang có sẵn miễn phí; Cục Thống kê Thụy Điển (1995) và Sundgren (1997).

Hơn nữa, Cục Thống kê Thụy Điển là đơn vị chủ trì của một dự án nghiên cứu dữ liệu đặc tả với tên gọi Quản lý Siêu thông tin Tích hợp (IMIM), do Liên minh châu Âu tài trợ trong Chương trình Khuôn khổ Nghiên cứu và Phát triển lần thứ 4. Ngoài những kết quả khác, dự án IMIM đã cho ra một sản phẩm phần mềm với tên gọi BRIDGE (Rauch & Karge 1999), có khả năng chứa dữ liệu đặc tả từ nhiều nguồn và cung cấp dữ liệu đặc tả cho các sản phẩm phần mềm cũng như cho các mục đích "con người" khác nhau. Phần mềm BRIDGE được dựa trên một mô hình dữ liệu và một hệ thống quản lý cơ sở dữ liệu hướng tới đối tượng với tên gọi ODABA-2 vượt trội hơn so với mô hình dữ liệu quan hệ với công nghệ tiên tiến nhất nhằm quản lý dữ liệu đặc tả. Hệ thống BRIDGE hiện đang được sử dụng làm cơ sở để phân loại các cơ sở dữ liệu tại nhiều quốc gia châu Âu. Một giao diện dữ liệu đặc tả tiêu chuẩn với tên gọi ComeIn đã được phát triển để làm một cơ sở khác (ngoài ODABA-2 và BRIDGE) cho các giao diện lưu trữ dữ liệu đặc tả.

Cục Thống kê Thụy Điển vừa chủ trì một dự án nghiên cứu dữ liệu đặc tả khác - dự án METAWARE do Liên minh châu Âu tài trợ trong Chương trình Khuôn khổ Nghiên cứu và Phát triển lần thứ 5. Dự án này tập trung vào việc quản lý dữ liệu đặc tả liên quan đến kho dữ liệu.

Quý vị có thể tìm hiểu thêm thông tin về phát triển dữ liệu đặc tả tại Cục Thống kê Thụy Điển trong Sundgren (2000).

5. Kết luận

Dữ liệu đặc tả rất phổ biến đối với các quy trình sản xuất và diễn giải thống kê. Việc xác định dữ liệu đặc tả thống kê đòi hỏi phải có kiến thức về người sử dụng và cách sử dụng tiềm năng, và do đó rất khó thực hiện. Phạm vi ý nghĩa của nó rộng đến mức các nhà sản xuất dữ liệu đặc tả phải xem xét việc sản xuất theo cách thức tương tự như cách thức đã được sử dụng để sản xuất dữ liệu. Hơn nữa, phạm vi các hoạt động có trong khía cạnh nhận thức của phương pháp khảo sát phải được mở rộng sang sản xuất và sử dụng dữ liệu đặc tả. Công tác quản lý dữ liệu đặc tả phải được xem như là một phần không thể thiếu trong sản xuất thống kê và bản thân hệ thống quản lý dữ liệu đặc tả phải được thiết kế từ các thành phần tích hợp tốt, lưu trữ dữ liệu đặc tả cũng như các công cụ và ứng dụng phần mềm.

Hoàng Linh (lược dịch)

Nguồn:
https://www.bls.gov/ore/pdf/st000040.pdf