SỰ CẦN THIẾT CẢI TIẾN DỮ LIỆU ĐẶC TẢ TRÊN TRANG WEB CỦA TỔNG CỤC THỐNG KÊ

*Lê Thủy Tiên**

1. Khái niệm về dữ liệu đặc tả và cách phân loại dữ liệu đặc tả thống kê

Dữ liệu đặc tả là thông tin cần thiết có thể sử dụng để diễn giải số liệu thống kê. Dữ liệu đặc tả mô tả dữ liệu bằng cách đưa ra các khái niệm, tổng mẫu, quy mô mẫu, đối tượng, biến, phương pháp luận và chất lượng. (Nguồn: cơ sở dữ liệu khái niệm định nghĩa của Eurostat). Để hình dung rõ hơn về dữ liệu đặc tả ta có thể lấy ví dụ: Nếu chỉ đưa ra một con số 98 triệu có lẽ mỗi chúng ta sẽ có một dự đoán của riêng mình. Người sẽ nói đây ắt hẳn là dân số của Việt Nam vào năm 2023, người thì cho rằng có lẽ đây là thu nhập bình quân đầu người năm của người dân Hà Nội? Tuy nhiên, khi con số này được gắn với một bảng biểu có tiêu đề: Dân số Việt Nam vào ngày 1/1/2023 thì tất cả mọi người đều hiểu con số này nói đến tổng dân số của Việt Nam vào năm 2023. Các thông tin từ tiêu đề bảng biểu, đơn vị biểu thị trong bảng biểu, thời gian, khái niệm làm rõ… được coi là dữ liệu đặc tả để chúng ta có thể hiểu được số liệu gắn với nó.

Ở Việt Nam, tại Nghị định số 64/2007/NĐ-CP về ứng dụng CNTT trong hoạt động của cơ quan nhà nước đã đưa ra khái niệm dữ liệu đặc tả như sau: "*Dữ liệu đặc tả (Metadata)*: là những thông tin mô tả các đặc tính của dữ liệu như nội dung, định dạng, chất lượng, điều kiện và các đặc tính khác nhằm tạo thuận lợi cho quá trình tìm kiếm, truy nhập, quản lý và lưu trữ dữ liệu.

Dữ liệu đặc tả thống kê có 02 chức năng cơ bản. Một là, xác định duy nhất và chính thức nội dung, các mối liên kết giữa các đối tượng và quy trình trong hệ thống thông tin thống kê. Hai là, quyết định các tham số kỹ thuật liên quan.

Trong thống kê, dữ liệu đặc tả được phân chia theo nhiều cách khác nhau.

Cơ quan thống kê New Zealand phân dữ liệu đặc tả thành:

Dữ liệu đặc tả khái niệm: *Dữ liệu đặc tả quy trình*; *Dữ liệu đặc tả tác nghiệp*; *Dữ liệu đặc tả hệ thống*; *Dữ liệu đặc tả bộ dữ liệu*.

Với Thống kê Anh quốc, các dữ liệu đặc tả sau được quan sát:

Dữ liệu đặc tả cấu trúc: Loại dữ liệu đặc tả này xác định dữ liệu và bộ dữ liệu như: các biến, vị trí định danh/chính trong file, loại dữ liệu, giá trị hợp lệ, tham chiếu đến danh mục mã kèm theo, mô hình dữ liệu.

Dữ liệu đặc tả quản trị: Mô tả quy trình và thông tin phụ trợ, đảm bảo việc xử lý đúng và quản trị dữ liệu.

Dữ liệu đặc tả tham chiếu: Mô tả nội dung, phương pháp luận, chất lượng của dữ liệu, nguồn gốc dữ liệu: Nguồn, người cung cấp; mô tả chung/ mục đích của bộ dữ liệu, tần suất, người sở hữu…

Dữ liệu đặc tả tác nghiệp, dữ liệu đặc tả mô tả, d*ữ liệu đặc tả quy trình*.

Cơ quan thống kê châu Âu phân biệt 02 loại dữ liệu đặc tả được sử dụng trong hệ thống thống kê châu Âu như sau: dữ liệu đặc tả cấu trúc và dữ liệu đặc tả tham chiếu.

Dữ liệu đặc tả cấu trúc: được sử dụng để xác định số liệu thống kê, ví dụ tiêu đề, phụ đề, mô tả ngắn, tên các chiều, tên biến…

Dữ liệu đặc tả tham chiếu: mô tả các khái niệm, định nghĩa, phương pháp luận sử dụng

* *Vụ Phương pháp chế độ và Quản lý chất lượng thống kê*

trong thu thập và sản xuất số liệu, cung cấp thông tin về chất lượng số liệu.

2. Thực trạng xây dựng, sử dụng dữ liệu đặc tả tại cơ quan thống kê:

Thời gian qua, hệ thống cơ sở dữ liệu nói chung và dữ liệu đặc tả thống kê nói riêng đã giúp người sử dụng tin tiếp cận với số liệu thống kê cho nhu cầu sử dụng của mình, phần nào hiểu được thông tin thống kê thông qua các dữ liệu giải thích mô tả số liệu, tuy nhiên vẫn có thể thấy các tồn tại cần khắc phục:

a. Cơ sở dữ liệu đặc tả mặc dù đã hình thành nhưng chưa thể đầy đủ, chưa thể hoàn thiện để làm cơ sở thống nhất đồng bộ dữ liệu

Hiện tại, Tổng cục Thống kê đã hình thành các kho cơ sở dữ liệu các cuộc Tổng điều tra và các phần mềm khai thác dữ liệu vi mô như:

- Kết quả Tổng điều tra dân số và nhà ở các năm 1999, 2009;
- Tổng điều tra cơ sở kinh tế, hành chính sự nghiệp các năm 2007, 2012, 2017, 2021…
- Tổng điều tra nông thôn, nông nghiệp và thủy sản các năm 2006, 2011, 2016…
- Khảo sát mức sống dân cư
- Điều tra Lao động việc làm
- Cơ sở dữ liệu về đăng ký doanh nghiệp-sản phẩm chia sẻ dữ liệu hành chính giữa TCTK và Tổng cục Thuế.
- Trang dữ liệu tóm tắt quốc gia (NSDP) được Tổng cục Thống kê phối hợp với Bộ Tài chính và Ngân hàng Nhà nước xây dựng chia sẻ số liệu về tài chính và kinh tế vĩ mô theo các định dạng Excel và SDMX đồng thời cung cấp thông tin về dữ liệu đặc tả liên quan.
- Hệ thống dữ liệu không gian dân số và phát triển được cập nhật và duy trì tại Vụ Thống kê Dân số và Lao động với sự hỗ trợ kỹ thuật của tổ chức UNFPA.
- Cơ sở dữ liệu về SDG áp dụng chuẩn SDMX đảm bảo sự so sánh quốc tế của số liệu Việt Nam đã hình thành dù mới ở bước sơ khai, song đây sẽ là bước tạo nền móng để cơ sở dữ liệu cấp quốc gia hài hòa với cơ sở dữ liệu toàn cầu về các chỉ tiêu phát triển bền vững.

Dữ liệu thống kê và cả dữ liệu đặc tả hiện cũng được quản lý theo nhiều hệ quản trị dữ liệu và phân tán khắp nơi trong ngành Thống kê từ Trung ương đến địa phương (bao gồm cả một số dữ liệu đã tập trung về Trung ương nhưng vẫn chưa thống nhất, thiếu tính liên kết theo không gian và thời gian).

b. Dữ liệu đặc tả không được cập nhật thường xuyên, chưa hoàn toàn theo chuẩn quốc tế

Mục Dữ liệu đặc tả trên trang web của Tổng cục Thống kê thông tin còn sơ sài, chủ yếu tập trung vào giải thích thuật ngữ thống kê các chuyên ngành. Theo đúng khái niệm của dữ liệu đặc tả, mục này nên bao gồm các bảng danh mục thống kê, phương pháp luận và quy trình thống kê.

Chế độ báo cáo thuộc dữ liệu đặc tả tham chiếu. Thông tin phần này trên trang web đã cũ từ những năm 2008, chưa có cập nhật mới. Nhiều thông tin đã lỗi thời qua 2 lần sửa đổi Luật Thống kê (2015 và 2021). Cụ thể:

Chế độ báo cáo áp dụng cho Cục Thống kê (tên mới hiện nay là Chế độ báo cáo ngành

thống kê). Phiên bản trên cơ sở dữ liệu đã cũ từ năm 2008, chưa có phiên bản năm 2018.

Chế độ báo cáo áp dụng cho Bộ, ngành: Phiên bản ở trên trang web Tổng cục Thống kê đã cũ, từ năm 2008 theo Quyết định số 111/2008/QĐ-TTg của Thủ tướng Chính phủ. Phiên bản hiện tại là Chế độ báo cáo quốc gia theo Nghị định số 60/2018/ NĐ-CP ngày 20 tháng 4 năm 2018. Hiện tại, do luật Thống kê mới sửa đổi năm 2021, một dự thảo chế độ báo cáo cấp quốc gia đang được xây dựng, cần có kế hoạch cập nhật lên cơ sở dữ liệu của ngành.

Chế độ báo cáo áp dụng cho doanh nghiệp đã bỏ từ năm 2015 từ khi Luật Thống kê ban hành. Tuy nhiên, trên trang web Tổng cục Thống kê vẫn tồn tại mục này.

Mục Hệ thống tài khoản quốc gia hiện tại không thể truy cập được có thể do lỗi thiết kế hay lỗi lưu trữ.

c. Dữ liệu đặc tả chưa hoàn toàn gắn với bảng số liệu. Khi truy cập Tab số liệu trên trang web TCTK, người dùng không tiếp cận được số liệu ngay, phải thao tác theo PXWeb, một cách trình bày bảng số liệu không trực quan, rối mắt, gây khó hiểu cho người dùng. Truy xuất dạng file PXweb khiến người dùng tin không thể hiểu cặn kẽ, cụ thể về số liệu mình vừa xem. Số liệu một nơi, dữ liệu minh họa một nơi.

d. Dữ liệu đặc tả còn thiếu các thông tin liên quan. Trên trang web TCTK, phần danh mục Dữ liệu đặc tả phần bảng danh mục thiếu các thông tin sau:

+) Mã số của từng bảng phân loại;

+) Bảng tương thích giữa các phiên bản của một bảng phân loại;

+) Các định dạng khác nhau của một bảng phân loại; Một bảng phân loại có thể có định dạng đầy đủ hoặc rút gọn.

+) Gia đình bảng phân loại. Gia đình bảng phân loại là hệ thống các bảng phân loại có liên quan với nhau về mặt nội dung và mã. Ví dụ Liên hợp quốc đã xây dựng một gia đình các bảng phân loại như Hệ thống phân ngành, Bảng phân loại sản phẩm, các bảng danh mục chi tiêu theo mục đích…

Đối với điều tra hầu hết các thông tin chuẩn về dữ liệu đặc tả tham chiếu theo SDMX là chưa đầy đủ. Theo chuẩn SDMX, dữ liệu đặc tả đối với quy trình (điều tra) gồm các nội dung sau: (1) Địa chỉ liên hệ; (2) Cập nhật dữ liệu đặc tả; (3) Trình bày thống kê; (4) Đơn vị đo; (5) Thời kì tham chiếu; (6) Nhiệm vụ thể chế; (7) Tính bảo mật; (8) Chính sách phổ biến; (9) Tần suất công bố; (10) Định dạng công bố; (11) Khả năng tiếp cận tài liệu; (12) Quản lí chất lượng; (13) Tính liên quan; (14) Tính chính xác, độ tin cậy; (15) Tính kịp thời, độ trễ; (16) Tính bảo mật; (17) Tính chặt chẽ; (18) Chi phí; (19) Chỉnh sửa số liệu; (20) Quy trình thống kê; (21) Bình luận; (22) Các dữ liệu đặc tả liên quan.

Tuy nhiên, hầu hết các nội dung dữ liệu đặc tả tham chiếu của các cuộc điều tra của TCTK giới hạn trong quyết định về phương án điều tra của từng cuộc điều tra, các nội dung dữ liệu đặc tả về chất lượng đều chưa có.

3. Đề xuất và kiến nghị

Phát triển, xây dựng và duy trì hệ thống dữ liệu đặc tả cần gắn liền với Chiến lược tầm nhìn phát triển của toàn ngành thống kê để đảm bảo hệ thống dữ liệu đặc tả được xây dựng đáp ứng các hướng phát triển của ngành thống kê nói riêng và nhu cầu thông tin nói chung.

Dữ liệu đặc tả khi xây dựng nên được gắn với số liệu. Số liệu được giải thích và được hiểu tường minh nhờ có dữ liệu đặc tả. Tuy nhiên, với cách sắp xếp dữ liệu đặc tả và số liệu trên trang web hiện tại của TCTK ta thấy hai mảng này tách rời nhau. Nên xây dựng một linked point (điểm kết nối) giữa số liệu và dữ liệu đặc tả.

Để các đơn vị thống kê chuyên ngành không phải cung cấp nhiều lần cùng một loại dữ liệu đặc tả cho nhiều cơ quan khác nhau, TCTK nên xây dựng một cơ sở dữ liệu dùng chung (shared database) để các cơ quan khác nhau khi có nhu cầu về dữ liệu đặc tả chỉ cần kết nối vào cơ sở dữ liệu này là có thể thu nhận được thông tin mình cần. Ngoài ra, hệ thống dữ liệu đặc tả thống kê còn đang phân tán, nằm rải rác ở nhiều đơn vị, cần một hệ thống thông tin đầu vào tập trung.

Để tăng cường hiệu quả, dữ liệu đặc tả cần phải có cấu trúc và nhất quán giữa các cơ quan thống kê, làm cơ sở cho việc chuẩn hóa, đồng bộ dữ liệu thống kê giữa các hệ thống.Thực hiện nghiên cứu, đánh giá các tiêu chuẩn dữ liệu đặc tả quốc tế trong lĩnh vực thống kê nhằm phát triển dữ liệu đặc tả thống kê. Ngay từ giai đoạn ban đầu khởi tạo xây dựng hệ thống dữ liệu đặc tả, việc tuân thủ các chuẩn quốc tế về xây dựng, trao đổi dữ liệu sẽ đảm bảo sự thuận lợi sau này cho cơ quan trong kết nối, chia sẻ, kết xuất thông tin. Các chuẩn thống kê cần tuân thủ ngay khi xây dựng hệ thống dữ liệu đặc tả là SDMX, SIMS, GSBPM...

Hệ thống dữ liệu đặc tả của thống kê Việt Nam chưa có phần chỉ tiêu liên quan tới chất lượng thống kê. Đây là mảng quan trọng giúp người sử dụng tin nắm được mức chính xác của số liệu và cách thức thu thập, điều chỉnh số liệu, tạo sự minh bạch cho số liệu cung cấp từ cơ quan Tổng cục. Điều này còn đặc biệt quan trọng, khi TCTK vừa bổ sung thêm một chức năng mới trong hoạt động của mình đó là quản lý chất lượng thống kê./.

TÀI LIỆU THAM KHẢO

1. Cathryn S. Dippo and Bo Sundgre, The Role of Metadata in Statistics

2. Jeremy Leipzig,1, Daniel Nust, Charles Tapley Hoyt, Karthik Ram, Jane Greenberg, The role of metadata in reproducible computational research

3. UNECE, Statistical Metadata in a Corporate context: A guide for manager, 2009

4. Graeme Oakley, Australian Bureau of Statistics, Quality infrastructure system- a case study of an end to end application at the ABS, 2008

5. Cử nhân Nguyễn Thị Ngọc Vân, Nghiên cứu xây dựng nội dung, cấu trúc hệ thống cơ sở dữ liệu thống kê kinh tế xã hội tổng hợp quốc gia, TCTK, 2006

6. Luật Lưu trữ số 01/ 2013/ QH13 ngày 11/11/2011

7. Tiến sỹ Lê Mạnh Hùng, Nghiên cứu xây dựng kho dữ liệu đầu vào của Tổng cục Thống kê, TCTK, 2009

8. Quyết định số 01/2019/QĐ-TTg ngày 05/01/2019 của Thủ tướng Chính phủ ban hành Bộ tiêu chí chất lượng thống kê nhà nước đến năm 2030.

9. Luật Thống kê số 01/2015/QH12

10. Luật Giao dịch điện tử số 51/2005/ QH11

11. Nghị quyết số 52-NQ/TW của Bộ Chính trị

12. Luật Công nghệ thông tin số 67/ 2006/ QH11

13. Chỉ thị số 15/CT-TTg ngày 22/5/2012 của Thủ tướng Chính phủ

14. Chu Thị Thắm , Nghiên cứu về dữ liệu đặc tả dùng cho quản lý văn bản điện tử trong các cơ quan nhà nước, 2014

15. Trần Thị Thu Trang, Xu hướng trao đổi thông tin của thống kê hiện đại, Vụ Thống kê Tổng hợp - Tổng cục Thống kê, 2019