Các dự án về Dữ liệu mở của Thống kê Nhà nước tại Nhật Bản

ToshihikoAKATANI, Trung tâm Thống kê Quốc gia

Giới thiệu:

Về cung cấp các số liệu thống kê Nhà nước, việc sử dụng Internet đã trở nên phổ biến trong mỗi Văn phòng Thống kê Quốc gia (NSO). Tuy nhiên, trong bối cảnh sự phát triển nhanh chóng công nghệ thông tin và truyền thông (ICT), tiềm năng về cách thức mới để sử dụng dữ liệu thống kê đang mở rộng. Vì vậy, giải quyết vấn đề này là một trong những ưu tiên hàng đầu cho NSOs.

Cụ thể, thúc đẩy việc sử dụng các dữ liệu Nhà nước, tức là tạo điều kiện thuận lợi cho "Open data - Dữ liệu mở" - gần đây dự kiến sẽ là một công cụ quan trọng cho việc tạo ra các hoạt động kinh doanh và dịch vụ mới, thực hiện các dịch vụ công thông qua sự hợp tác giữa khu vực công và khu vực tư nhân, và đảm bảo tính minh bạch và trách nhiệm giải trình của Chính phủ. Trong tương lai, số liệu thống kê ở các lĩnh vực đóng vai trò như những người dẫn đầu giỏi nhất về Open data trong Chính phủ, không chỉ trong phạm vi mỗi quốc gia mà còn trên thế giới.

Cục Thống kê Nhật Bản (SBJ) và Trung tâm Thống kê Quốc gia (NSTAC) xây dựng một Dự án có tên là "Cổng kết nối nhằm cải thiện các dịch vụ thống kê để thân thiện với người dùng" (GAUSS) vào năm 2011, thúc đẩy sử dụng các số liệu thống kê của công chúng, khu vực tư nhân và ủng hộ dịch vụ tạo ra giá trị gia tăng và đổi mới doanh nghiệp. Dự án này có ý định nâng cấp các phương pháp phổ biến số liệu thống kê đồ sộ và đa dạng, trong tương lai công nghệ tiên tiến này sẽ vẫn được sử dụng.

Dự án bao gồm các chủ đề sau:

Dự án 1: Phát triển môi trường để nâng cao việc sử dụng các số liệu thống kê của giao diện lập trình ứng dụng (API).

Dự án 2: Cải thiện Thống kê hệ thống thông tin địa lý (GIS).

Bài viết này mô tả tổng quan từng dự án và cung cấp một cái nhìn chung về các công việc trong tương lai và tác động đối với NSO.

Từ khóa: dữ liệu mở (Open data), giao diện lập trình ứng dụng (API), hệ thống thông tin địa lý (GIS)

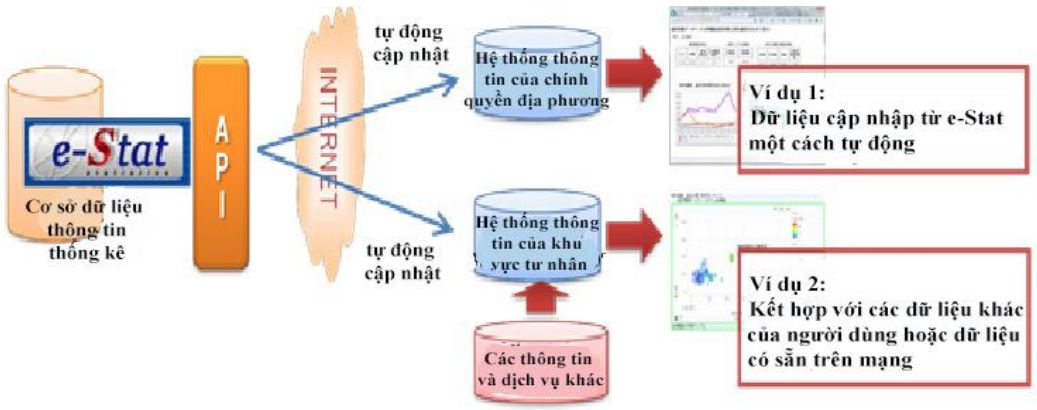
Dự án 1: Phát triển môi trường để nâng cao việc sử dụng các số liệu thống kê của API

Dự án đầu tiên là "Phát triển môi trường để nâng cao việc sử dụng các số liệu thống kê của API". API là viết tắt của "Application Programming Interface - Giao diện lập trình ứng dụng" và cho phép các thành phần phần mềm để tương tác với nhau.

Các hàm của API được sản xuất bởi SBJ và NSTAC, tự động thu hồi các dữ liệu thống kê được lưu trữ trong cơ sở dữ liệu của "e-Stat"

(trang thông tin điện tử số liệu thống kê Nhà nước của Nhật Bản; phát triển bởi SBJ và điều hành bởi NSTAC) thông qua chương trình. Trong quá khứ, người sử dụng e-Stat lấy dữ liệu thống kê bằng cách tìm kiếm thủ công và tải về các bảng số liệu.

Hình 1: Tổng quan về các chức năng của API

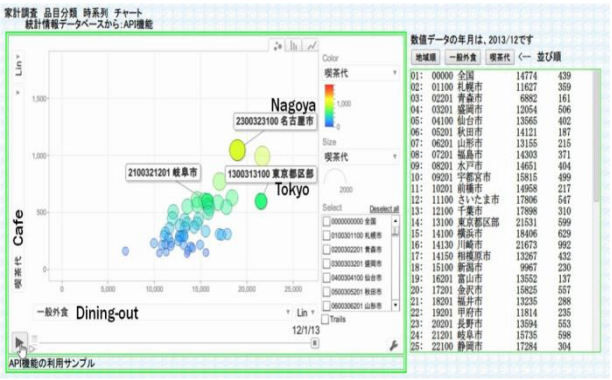


Với các chức năng của API, ví dụ, chính quyền địa phương và các công ty địa phương có thể sắp xếp một công việc với khối lượng lớn từ việc cập nhật dữ liệu thống kê trong hệ thống thông tin của họ từ e-Stat.

Hơn nữa, các hàm API sẽ cho nhiều cải tiến về phân tích dữ liệu, phối hợp các chương trình khác từ Internet.

Đây là một trong những ví dụ được cung cấp bởi các chức năng của API:

Hình 2: Ví dụ về phân tích được hỗ trợ bởi các chức năng của API



Ở Hình 2, các bong bóng đại diện cho số liệu thống kê của 52 thành phố lớn tại Nhật Bản, từ cuộc Điều tra Thu nhập và chi tiêu hộ gia đình từ tháng 01/2007 đến tháng 5/2013. Trục hoành cho thấy chi tiêu trung bình cho việc đi ăn ngoài của các hộ gia đình tại các thành phố tương ứng và trục tung cho thấy chi tiêu đối với các quán cà phê (một phần của việc đi ăn ngoài).

Tính năng động của chi tiêu hàng tháng được hiển thị với chuyển động của mỗi bong bóng bằng cách áp dụng API (biểu đồ di động Google). Ở các thành phố lớn như Tokyo cho rằng nhu cầu trong ngành Công nghiệp nhà hàng đã tăng, Nagoya có ưu đãi duy nhất cho các quán cà phê, được biết đến như là một lựa chọn của ý tưởng "bữa ăn sáng đặc biệt".

Chuyển động - biểu đồ phân tán bong bóng cho thấy không chỉ những cách thức mới để thể hiện số liệu thống kê mà còn chỉ ra cách dễ dàng nhận ra các biểu thức mới bởi sự "kết hợp" (phối hợp với các số liệu thống kê API và chương trình khác). Đây là một trong những điểm hấp dẫn nhất của các hàm API.

Các chức năng của API được đưa vào sử dụng thử từ 10/6/2013, trong đó cung cấp quyền truy cập vào các thống kê chính thức được tiến hành bởi SBJ (điều tra dân số, thu nhập gia đình và chi tiêu với 22 loại số liệu thống kê, 32.000 bảng).

Sau khi chạy thử nghiệm, dựa trên phân tích sử dụng, hiểu biết nhu cầu và xác nhận hệ thống, hàm API đã được chính thức đưa vào e-Stat kể từ ngày 31/10/2014 để làm dữ liệu thống kê tất cả các Bộ (Điều tra cơ bản trường học, khảo sát bệnh nhân, 57 loại số liệu thống kê, 70.000 bảng) có sẵn trong định dạng máy tính có thể đọc được.

Kể từ đó, các hàm API đã thu hút được rất nhiều người sử dụng dữ liệu, từ phương tiện truyền thông đại chúng đến người dùng cá nhân, người đã phát triển hàng trăm chương trình ứng dụng và tung lên trên mạng Internet, chẳng hạn như "So sánh các dữ liệu khác nhau ở mỗi huyện trên bản đồ", "trang web về việc xếp thứ hạng các tỉnh", "Trang web trong đó sử dụng đồ thị để hình dung xu hướng của người dân và doanh nghiệp dựa trên dữ liệu chuỗi thời gian", và "Ứng dụng di động về số liệu thống kê chính thức". Mỗi trường hợp chứa đựng sự khéo léo của nhà phát triển để sử dụng số liệu thống kê từ các quan điểm khác nhau của người sử dụng.

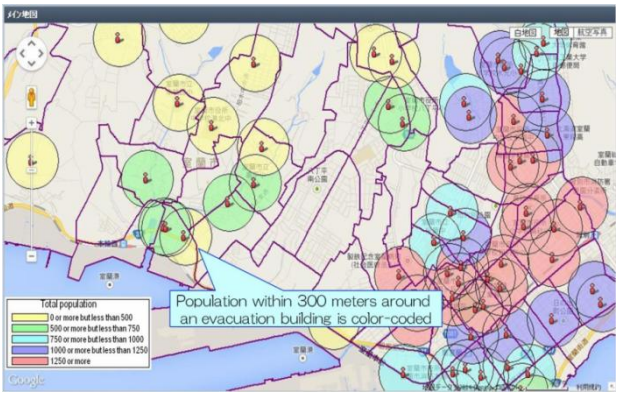
Bằng cách này, các hàm API cho phép phần lớn các dữ liệu thống kê mà máy có thể đọc được và có thể đóng góp vào ảnh hưởng nổi bật về hoạt động kinh tế và xã hội.

Dự án 2: Cải thiện Thống kê GIS

Thống kê GIS (Hệ thống thông tin địa lý) là trang web cung cấp đầy đủ thông tin thống kê và thông tin địa lý để hiểu về dữ liệu thống kê trên bản đồ. Bây giờ, e-Stat đã cung cấp thống kê GIS như là một "Bản đồ chuyên đề". Dự án GAUSS đã phát triển chức năng mới để cải tiến số liệu thống kê GIS bằng cách thêm tính năng nhập bất kỳ dữ liệu thuộc sở hữu của người sử dụng trong một khu vực tùy ý trên các trang web. Ngày 20/01/2015, thống kê GIS mới "jSTAT MAP - Phân tích khu vực nhỏ trên bản đồ" bắt đầu cung cấp các chức năng trên.

Dưới đây là một ví dụ về cách sử dụng phân tích jSTATMAP.

Hình 3: Ví dụ về việc sử dụng phân tích jSTATMAP



Trong phân tích, thông tin về sơ tán tòa nhà hoặc nơi trú ẩn trong thành phố trong trường hợp có thảm họa, được chính quyền địa phương công bố như một phần của chính sách Open data, được kết hợp chặt chẽ với thống kê GIS (kí hiệu hình người đại diện cho tòa nhà sơ tán). Nếu bạn có thông tin địa chỉ mỗi tòa nhà, bạn có thể nhập dữ liệu vào bản đồ bằng mã địa lý, trong đó chuyển đổi địa chỉ vào các tọa độ kinh độ và vĩ độ.

Mỗi vòng tròn mã màu cho thấy dân số trong vòng 300 mét xung quanh một tòa nhà sơ tán. Dân số chi tiết của từng khu vực được tính dựa vào kết quả ước lượng khu vực nhỏ của cuộc Tổng điều tra Dân số và được hiển thị trên mỗi vòng tròn.

Nếu chính quyền địa phương công bố sức chứa của mỗi tòa nhà, nó có thể so sánh được dân số ở đây. Hơn nữa, khẳng định sức chứa của một khu phố có thể cho phép xác định các tuyến đường sơ tán thích hợp nhất trong khu vực. Hơn nữa, ước lượng dân số là có sẵn không chỉ trong khu vực được khoanh tròn mà còn ở các khu vực bất kỳ được chỉ rõ.

Hình dung thế này giúp dễ dàng cân nhắc sự khác nhau, bao gồm xây dựng những người dễ bị tổn thương như người già, trẻ sơ sinh và

trẻ mới biết đi, nên sơ tán đến đó và chấp nhận yêu cầu hỗ trợ từ chính quyền địa phương. Làm như vậy giúp chính phủ có thể phát triển một chính sách chi tiết về giảm nhẹ thiên tai hơn hiện tại, thông qua khai thác jSTAT.

Ngoài việc làm cho việc sử dụng số liệu thống kê trở nên dễ dàng hơn, số liệu thống kê GIS cải tiến sẽ cung cấp một lợi thế trong Open data được công bố bằng các Bộ khác trong chính quyền trung ương hoặc chính quyền địa phương và các số liệu thống kê sẽ có hệ thống liên kết. Dự kiến thống kê GIS sẽ có vai trò khuyến khích và kích hoạt các sáng kiến Open data trong lĩnh vực hành chính khác.

Đặc điểm Chính sách Open data trong lĩnh vực thống kê chính thức của Nhật Bản

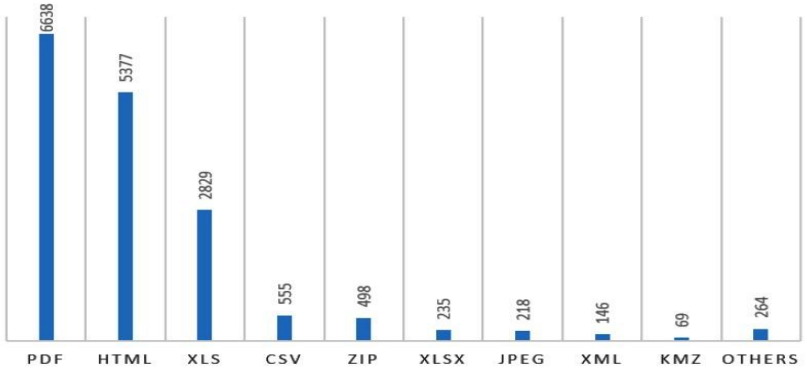
Như đã đề cập trong phần giới thiệu, lĩnh vực thống kê cần phải đóng vai trò trong tương lai và như người dẫn đầu về Open data trong Chính phủ, không chỉ trong phạm vi mỗi nước mà còn ở trên thế giới. Điều này có nghĩa hết sức đặc biệt dựa trên kinh nghiệm của Nhật Bản.

Đầu tiên, thực tế nhận thấy rằng các dịch vụ mở về số liệu thống kê Nhà nước đã được đặt ra nhiều năm trước khi các trang web Danh mục dữ liệu được phát triển gần đây tại một số nước. Tại Nhật Bản, trang web Danh mục dữ liệu "data.go.jp" đã hoạt động như một dịch vụ toàn diện từ ngày 01/10/2014. Tuy nhiên, "e-Stat" - trang thông tin thống kê Nhà nước tại Nhật Bản đã được đưa ra vào ngày 01/04/2008. Có thể lập luận rằng lĩnh vực thống kê Nhà nước đã có vai trò về Open data ngay cả trước khi thuật ngữ "Dữ liệu mở" được đặt ra.

Ngoài ra, đối với hàm API, SBJ và NSTAC đã hoàn thành việc xây dựng môi trường để sử dụng API tập trung vào tất cả các số liệu thống kê của chính phủ, hiệu quả sử dụng cơ sở dữ liệu e-Stat. Trong e-Stat, số liệu thống kê của chính phủ đang được lưu trữ không chỉ trong Excel (xls), các tập tin CSV mà còn trong các định dạng cơ sở dữ liệu. Các chức năng API giới thiệu lần này có lợi thế lớn nhất là chúng ta có thể truy cập thông tin cần thiết thông qua các cơ sở dữ liệu hiện có của e-Stat và lấy dữ liệu ở định dạng nhất định, chẳng hạn như XML và JSON. Điều này tránh khỏi việc cần phải phát triển một cơ sở dữ liệu riêng biệt cho việc cài đặt các hàm API và cho phép các hàm API để bao phủ số liệu thống kê được các Bộ khác tính toán.

Trong khi đó, hầu hết dữ liệu hành chính trong các lĩnh vực khác hơn so với số liệu thống kê Nhà nước được lưu trữ trong data.go.jp ở định dạng PDF, không phải là dạng thuật ngữ máy. So với tình hình này, công nhận rằng dữ liệu trong lĩnh vực thống kê Nhà nước là một trong những dữ liệu chuẩn nhất.

Hình 4: Bộ dữ liệu có sẵn trong data.go.jp (tính đến ngày 13/03/2015)



Các yếu tố dẫn đến tình trạng này có thể do hệ thống thống kê Nhà nước của Nhật Bản được phân cấp rất cao. Thay vì để cho từng Bộ tiến hành khảo sát và sản xuất số liệu thống kê cần thiết, tăng cường chức năng các hàm kiểm tra như một điều phối chính sách và sự phát triển của dịch vụ một cửa đã được phát huy trong một thời gian dài. Điều này có thể góp phần vào các chính sách Open data hiện nay của Nhật Bản.

Căn cứ vào hiện trạng này, tham chiếu đến một liên kết đến e-Stat được coi là đăng ký dữ liệu thống kê Nhà nước trong data.go.jp.

Ideathon, Hackathon[1]

Tại Nhật Bản, một cuộc thi ý tưởng và các chương trình khai thác Open data, được gọi là "LOD Challenge" đã được tổ chức dưới sự hợp tác của các chuyên gia ngành kinh tế từ năm 2011. SBJ và NSTAC đã tham gia cuộc thi với tư cách là "Các đối tác cung cấp dữ liệu" kể từ năm 2013 và đưa ra ví dụ nổi bật về sử dụng số liệu thống kê từ công chúng.

Ngoài ra, SBJ và NSTAC tổ chức nội bộ Ideathon, Hackathon trong năm 2013. Tác phẩm xuất sắc trong cuộc thi đã được thông qua cho một ứng dụng điện thoại thông minh, được gọi là "App On Statistics", bởi SBJ và ứng dụng này đã được cung cấp cho công chúng kể từ tháng 4/2014.

Hình 5: Giao diện của Ứng dụng Thống kê "App On Statistics"

Công việc sắp tới

Như đã mô tả ở trên, các sáng kiến khác nhau đã giúp cho chính sách Open data của thống kê chính thức tại Nhật Bản. Tuy nhiên, các vấn đề sau đây được công nhận là công việc quan trọng trong tương lai để nâng cấp các sáng kiến ở mức độ cao hơn.

Lập Cơ sở dữ liệu thống kê Nhà nước không được sản xuất bởi NSOs

Vì tất cả các số liệu thống kê của chính phủ lưu trong một định dạng cơ sở dữ liệu trong e-Stat là trong phạm vi chức năng API, công việc tiếp theo là biên dịch hơn nữa trong cơ sở dữ liệu.

Mặc dù có 57 loại số liệu thống kê trong một định dạng cơ sở dữ liệu, nhiều hơn 500 loại số liệu thống kê của chính phủ được bao phủ bởi e-Stat. Do đó, chỉ có khoảng 10% số liệu thống kê trong e-Stat được tập hợp trong một cơ sở dữ liệu. Tuy nhiên, không phải tất cả các Bộ không phải là NSO có được kỹ năng và bí quyết biên dịch các thống kê hiện có.

Trong tương lai, SBJ và NSTAC sẽ hỗ trợ các Bộ này để mở rộng phạm vi các số liệu thống kê có sẵn trong các chức năng API.

Thuận lợi của siêu dữ liệu

Khi làm cho dữ liệu thống kê có thể đọc được bằng máy, cũng làm cho siêu dữ liệu đọc được bằng máy sẽ góp phần tạo nên sự hiểu biết thích hợp về thống kê.

[1] Ideathon, Hackathon là một cuộc thi mà các lập trình viên, cùng những người liên quan trong ngành phát triển phần mềm như các nhà thiết kế đồ họa, thiết kế giao diện, quản lý dự án sẽ hợp tác với nhau trong thời gian ngắn để hoàn thành một dự án phần mềm. Ideathon thường kéo dài từ 1-5h còn Hackathon kéo dài tới 2 ngày. Từ năm 2014, Việt Nam cũng tổ chức cuộc thi này tại hai thành phố lớn là Hà Nội và TP Hồ Chí Minh.

Tuy nhiên, siêu dữ liệu của các số liệu thống kê chính thức tại Nhật Bản không phải luôn được chuẩn hóa. Hơn nữa, để làm cho siêu dữ liệu thực sự có sẵn, cần phải làm rõ tên cũng như định nghĩa về siêu dữ liệu đó. Đối với vấn đề này, mỗi quốc gia phải phối hợp, hợp tác với các sáng kiến quốc tế như SDMX (Số liệu thống kê và Trao đổi Siêu dữ liệu).

Cũng lưu ý rằng NSTAC hiện đang tiến hành chạy thử nghiệm cung cấp các số liệu thống kê như SDMX.

Multilingualization (đa ngôn ngữ)

Quan điểm của hợp tác quốc tế là điều cần thiết cho chính sách Open data. Các tổ chức quốc tế như OECD và IMF đang thúc đẩy việc cung cấp dữ liệu như SDMX và như vậy, về việc nâng cấp mở dữ liệu với, cần lưu ý rằng cần cung cấp sao cho có thể so sánh quốc tế được.

Tại Nhật Bản, nhiều số liệu thống kê dữ liệu và siêu dữ liệu được cung cấp ở dạng ngôn ngữ Nhật, không áp dụng cho dữ liệu quốc tế. Do đó, NSOs của Nhật Bản có một thách thức lớn của multilingualization, dịch sang ngôn ngữ tiếng Anh.

Đối phó với LOD (liên kết dữ liệu mở)

Khi Tim Berners-Lee (nhà phát minh Wide Web World) ủng hộ, có năm cấp độ của Open data (xem Bảng 1).

Chức năng API hiện tại đang ở cấp ba. Trong tương lai, chuyển đến cấp độ cao nhất sẽ được tạo điều kiện để cho phép tương tác với dữ liệu khác nhau trên web.

NSTAC tiến hành chạy thử nghiệm cung cấp mã dữ liệu đô thị như LOD từ ngày 06/12/2013. Đây là kế hoạch để bắt đầu xem xét làm thế nào để cung cấp số liệu thống kê như LOD.

Bảng 1: Bảng mô tả các mức độ của Open data (5sao)

Sao	Miêu tả	Các ví dụ
★	Có sẵn trực tuyến, được cấp phép công khai, trong bất kỳ định dạng điện tử	Số liệu thống kê trong bảng không đọc được bằng máy (GIF hoặc JPEG), HTML có cấu trúc, hoặc gắn vào trong các báo cáo PDF. Dữ liệu được công bố công khai, nhưng khó khăn để tìm kiếm và phải được nhập lại bằng tay.
★★	Có sẵn trực tuyến, công khai được cấp phép, trong các định dạng điện tử thông thường	File dữ liệu định dạng độc quyền như Excel, SPSS, SAS hoặc STATA yêu cầu phần mềm đặc biệt và đào tạo.
★★★	Có sẵn trực tuyến, công khai được cấp phép, trong các định dạng điện tử không độc quyền	File dữ liệu trong các định dạng mở như CSV, JSON, XML, DDI, SDMX, hoặc có cấu trúc ASCII
★★★★	Tất cả những điều trên, cộng với việc sử dụng các URIs duy nhất (định danh Internet duy nhất) để xác định và mô tả dữ liệu	File dữ liệu trong các định dạng liên kết dữ liệu như RDF, cho phép dữ liệu được liên kết với nhau với các tập tin dữ liệu khác một cách dễ dàng.
★★★★★	Tất cả những điều trên, và các liên kết đến các dữ liệu khác để cung cấp bối cảnh	Trang Dataset cung cấp siêu dữ liệu thông dụng có thể đọc, liên kết với các định nghĩa chuẩn và các thông tin liên quan.

Nguồn: Thách thức về Open data và Cơ hội cho Văn phòng thống kê quốc gia (18/01/2014, tại Sự kiện bên lề UNSC: Open data cho NSOs)

(Xem tiếp trang 43)

(Tiếp theo trang 49)

Hãy nhớ rằng số liệu thống kê Nhà nước là cơ sở hạ tầng xã hội quốc gia, NSOs phải nhận ra tiện lợi hơn và nâng cao sử dụng số liệu thống kê, cũng như góp phần phát triển kinh tế xã hội, giải quyết các chính sách Open data.

SBJ và NSTAC sẽ tiếp tục dẫn đầu các sáng kiến của Chính phủ như là người giỏi nhất trong lĩnh vực Open data.

Thái Học (dịch)

Nguồn: Hội thảo Thông tin thống kê của các nhà thống kê châu Âu, ngày 27-29/4/2015 tại Nhật Bản

https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.45/2015/OD21_Projects_of_OpenData_for_Offcicial_Stats_Japan-_Akatani.pdf