Final project report: ”Citizen-centric access to statistics”

2012-01-15 Bo Sundgren

Summary of the project

The project has investigated the possibilities to define an easy-to-use, powerful, and flexible standard interface for communication of statistical data and metadata between the dissemination databases of statistics producers (statistical agencies) and citizen-centric applications, developed by entrepreneurs or by citizens themselves.

A consortium consisting of members from Estonia, Iceland, and Sweden has run the project in the form of a network. In addition to the consortium members, interested entrepreneurs and other actors from government agencies and private business, even outside the Nordic-Baltic countries, have participated in the network. Three major seminars have been organised in Estonia, Iceland, and Sweden during 2011. The project has also participated in the conferences “Offentliga Rummet” in Trollhättan, Sweden, 24-26 May 2011, and ICEGOV in Tallinn, Estonia, 26-28 September 2011. In addition to producing its own conclusions concerning the problems, opportunities, and possible solutions of citizen-centric access to statistics, the project has prepared for development and implementation projects to be proposed for the next phase of the Nordic-Baltic initiative on Citizen-Centric eGovernment Services, and within the broader framework of the European R&D program.

Project background

There is a rapidly growing supply of statistics, and everyone is using statistics in all sectors of society, private and public: researchers, analysts, business decision makers, politicians, journalists, and ordinary citizens – all use statistics to try to understand an ever more complex world, to make decisions for themselves and others, to participate in democratic processes, in arguments and negotiations, etc.

National statistical offices have the role in the government (and hence eGovernment) of a modern, democratic society to provide all those mentioned above, including the ordinary citizens, with objective statistics of good quality, and to make those statistics available to everyone, often as a public good, free of charge, financed by the taxpayers. In addition, statistical offices in some countries may undertake special assignments for a fee. For example, they may do more advanced processing and analysis on the basis of the general-purpose, public data. They may even carry out special-purpose statistical surveys for clients who are ready to pay for them.

Traditionally national statistical agencies and other producers of statistics have made statistics available through printed publications, more or less well designed to correspond to the needs of users. Since the advent of computers, and especially since the general breakthrough of the Internet in the mid 1990’s, statistical agencies have made statistics available via websites. Nowadays the web-based dissemination of statistics is offered for free by most national and international statistical agencies – with OECD still as a notable exception. Many different forms of dissemination are offered: from electronic versions of printed publications to multidimensional databases, giving the user very flexible access to subsets of statistics, time series as well as cross-sectional data, selected by the user herself, and presented in a variety of forms, ranging from traditional statistical tables to graphs and animations. The statistical data are accompanied by metadata in order to enable the user to interpret and understand the presented statistics.
All this is excellent, but it is still the producer statisticians in the statistical offices who by and large determine which statistics should be made available, and how they should be made available. Of course many statistics producers have an ongoing, constructive communication with their users and make their best to fulfil their needs and demands within the budget limits.

However, more and more often it happens that users of statistics themselves, or innovative entrepreneurs acting in cooperation with them, come up with new ideas about how to make statistics more understandable and useful. For example, Hans Rosling, professor of international health at the Karolinska Institute and founder of Gapminder, www.gapminder.org, has very convincingly demonstrated how statistics can be presented in new ways, e.g. through animated, so-called bubble graphs, so as to give everyone, from the world’s top leaders to the ordinary citizens, completely new insights about our world and where it is going.

Innovate initiatives by entrepreneurs or by the citizens themselves are often hampered by the lack of simple, easy-to-use standards for downloading statistical data and metadata to the users’ own systems and to new citizen-centric services developed by entrepreneurs. The EU Public Sector Information (PSI) Directive deals with these problems and urges member states to implement legislation supporting initiatives by entrepreneurs to develop and market new, innovative services based on available public sector data. However, many governmental agencies, including national statistical offices, have still not properly understood the intentions of the PSI directive, and the legislations based on this directive. They often still prefer to develop and implement new services themselves, rather than leaving this to the entrepreneurs, working in free competition and on their own risk, without the budget limitations and other restrictions existing for government agencies.

project proposal

The consortium

There was a consortium behind the project proposal consisting of members from Iceland (Statistics Iceland), Estonia (Statistics Estonia and the e-Governance Academy), and Sweden (Stockholm University, Mid Sweden University, Statistics Sweden, and Gapminder Foundation).

The proposal

The proposal was to investigate the possibilities to define an easy-to-use, powerful, and flexible standard interface for communication of statistical data and metadata between the dissemination databases of statistics producers (statistical agencies) and citizen-centric applications, developed by entrepreneurs or by citizens themselves. The interface should fulfil the requirements stated above, in particular the requirements implied by the EU Public Sector Information (PSI) Directive.

The investigations were based upon, *inter alia*:

- a preliminary investigation made by Björgvin Sigurdsson, Statistics Iceland, presented at a conference in Copenhagen in August 2010 (see reference)
- a preliminary investigation made by Statistics Sweden and Gapminder Foundation in cooperation
- the PSI directive and related legislation in EU member countries
- requirements by citizen-centric systems and applications, as expressed, for example, by Hans Rosling and other entrepreneurs and creators of citizen-centric services
- relevant existing and/or emerging standards, such as SDMX, www.sdmx.org, developed by a consortium of seven international organisations; however, these standards have not been developed with citizens and entrepreneurs in focus, and they have to be further developed for these purposes, especially as regards simplicity and richness of user-oriented metadata
Time frame and size of project
In accordance with the directives from Vinnova, the project was scheduled for 12 months, starting the 1st of January 2011, and was planned for a budget with financing from Vinnova in the order of 500,000 SEK plus in kind contributions from the participants in the consortium.

Ambition level
Given the time frame and the financing as stated above, a reasonable level of ambition was stated as follows:

- the members of the consortium will investigate and analyse the issues, as described above, aiming at defining a limited number of alternative standard procedures and technical solutions to be tried and tested in practice – to a limited extent within this project, and more extensively in subsequent projects to be defined
- the consortium will invite others interested to participate in the discussions, thus creating a broader network of interested persons and organisations inside and outside the Nordic-Baltic countries
- the consortium, extended with the network just mentioned, will lay the foundation for further projects, developing and testing more full-scale solutions and standards; these projects could take place within the Nordic-Baltic framework, but also within the EU research programs

Method of work
According to the plans, the members of the consortium and the emerging network, or reference group, described below were assumed communicate via e-mail, social media, etc. Three seminars would be organised: one in Estonia, one in Sweden, and one in Iceland. Limited software development (prototyping and testing) might take place during the project – and be planned for subsequent projects.

Reference group
As soon as it became more widely known that the project been defined, and a project consortium had been formed and was going to apply for funding from the Nordic-Baltic sponsors, other actors indicated their great interest in this project, and asked about the possibilities to participate in some form. The actors who showed their interest come both from the public sector (e.g. statistical agencies in other Nordic-Baltic countries) and from the private sector (entrepreneurs). We also assumed that software vendors and developers of open source solutions could be interested to follow and participate in our discussions. For all these reasons, plans were made for a broader reference group of people, who would be invited to participate in project seminars and other forms of communication (blogs etc) to be organised.

Project website and blog
As soon as the project was approved by Vinnova, a project website and blog were set up:

Project website: https://sites.google.com/site/accessstostatistics/

Project blog: http://stataccess.blogspot.com/

These initiatives turned out to be very successful. All documentation emanating from the project and its activities have been organised and presented on the website, and this has been – and still is – very much appreciated by a world-wide audience. Lively discussions have been started and are still in full swing on the blog.
Experiences and conclusions from the project

The consortium and reference group behind the project “Citizen-centric access to statistics”, led by Stockholm University, Department of Computer and Systems Sciences, consisted mainly of members from the statistical agencies in Estonia, Iceland, and Sweden, and from entrepreneurs developing tools for citizen-centric use of statistics, notably Gapminder (Hans and Ola Rosling), NCVA (Mikael Jern), and The factlab (Peter and Mike Andersson). In addition there have been participants from the Estonian eGovernment Academy and observers from some international research groups and think tanks with focus on citizens and their involvement in decision processes: Gov2u, Izwe, Praxis Center for Policy Studies, and the Institute of Baltic Studies.

Summary of major experiences and conclusions

- Most statistical agencies and international organisations are quite good at making large parts of their statistical outputs (databases and electronic publications) available for access and free downloading from their websites in a citizen-friendly way.
- There are still negative exceptions to what was said in the previous point, despite many excellent initiatives by statistics producers, and despite heroic efforts by Hans Rosling and others to convince statistics producers to make their data available.
- The availability of necessary documentation and metadata, associated with the statistics, is still lagging behind, but some progress can be noted.
- Unfortunately it is not enough that each statistics producer makes its data and associated metadata available in a standardised way. Most users of statistics need data from many different producers of statistical data, and then it becomes important not only that each producer uses standard definitions, standard formats, and standard procedures, but that different producers use the same standards, and provide user interfaces with the same “look and feel”. Statistical data from different producers must be easy to compile and compare.
- Direct access and downloading of statistical data and metadata from the websites of statistics is only one type of use of statistics, let alone an important one, especially for satisfying basic, informative needs of citizens. However, there are many other important needs that are still not met in a satisfactory way by producers and providers of statistics, for example:
  - needs for bulk downloads of statistical data and associated metadata in an easy, inexpensive, and uniform way from different statistics producers
  - needs for automatic delivery and updates of data and metadata defined by users by means of requests tailored by the users themselves to fit their needs
  - availability of statistical data and metadata, from different statistics producers, through application program interfaces (API:s) in an easy, inexpensive, and uniform way
- Entrepreneurs, research institutes, and others have to spend unacceptable large efforts and resources just to get the statistical data and metadata they need from different producers and providers of statistics. They tend to silently accept this, because they see no alternatives, but it is really a waste of valuable time of very qualified people, and it prevents many potential, citizen-oriented products and services from being implemented, because data/metadata management operations, which should be simple, inexpensive, standardised routine operations, become deterrently complex and costly tasks.
- Many users of statistics need only a small subset of the statistics that statistical agencies and others produce. However, this does not mean that statistics producers should limit themselves to making only a subset of their data available. Different subsets, organised in different ways (time series, multidimensional, etc) are needed by different users for different usages. Statistics producers should have the ambition to make all their data, and associated metadata, available in a simple, standardised, and flexible way – restricted only by laws concerning confidentiality.
- A number of potential candidates for becoming standard formats and associated procedures for exchange of statistical data and metadata exist, for example:
o PC-AXIS, developed and managed by Statistics Sweden and used by many statistical agencies around the world
o SDMX, developed and managed by a consortium of seven international organisations
o The Dataset Publishing Language (DSPL) with the visualisation tool Google Public Data Explorer, developed and managed by Google
o The Data Documentation Initiative (DDI) and associated metadata tools, managed by the DDI Alliance
o Documentation and metadata standards developed and managed by national statistical agencies and international organisations like Eurostat, IMF, World Bank, and OECD

- None of the existing candidates for becoming standard formats and procedures for exchange of data and metadata fulfil the different requirements of users and usages of statistical data and metadata, as briefly described above.
- Most likely, no single tool (format and associated procedures) will ever be able to satisfy all requirements. More likely, a toolbox consisting of a number of compatible tools and standard solutions will meet the whole range of needs.
- It is not a feasible strategy to wait for a final standard solution. It is better to implement a few promising solutions and have them iteratively tested and incrementally improved by engaged users of statistics with different needs and profiles.

**How to continue the work for citizen-centric access to statistics**

The project consortium has concluded that it would be a natural continuation of the project to develop and test a workbench and a toolbox for citizen-centric access to statistics, first a prototype and then full-scale solutions. Like in the project, which is now finalised, the work should be done in close cooperation between researchers, statistics producers, and entrepreneurs. The solutions should be based on standard tools and interfaces.

To the extent that the presently existing tools and procedures need to be further developed, new efforts would have to rely on constructive cooperation with those agencies and persons who are responsible for these tools and procedures, e.g. Statistics Sweden (PC-AXIS), the SDMX consortium (SDMX), and Google (Dataset Publishing Language and Google Public Data Explorer).

The new work should start from a number of typical use cases. Such use cases have been identified by present project, “Citizen-centric access to statistics”, for example:

- Repetitive requests from a particular user for “the same” statistical data and metadata, tailored by the user to the needs of a particular presentation, analysis, or other application, properly updated and delivered to the user, or rather to a user application, either on demand, or automatically, as soon as new data/metadata are available.
- Bulk downloads of large subsets of statistical data and associated metadata, specified by the user. The bulk downloads may be requested as unique, singular events, or more or less regularly. In the latter case, the download may contain complete data/metadata according to the original request, or it may contain only data/metadata that have been added or updated.
- Requests may concern statistical data organised as multidimensional data or time series data, with their respective special requirements.

Users should take active part in defining and elaborating the test cases, while assisting in the design and testing of prototypes (mockups) of tools and interfaces acquired and developed by other participants in the consortium.
Figure 1. Outline of an architecture for a Statistics User/Provider Interactive Workbench&Toolbox.

Proposed architecture of a new solution

*Figure 1* illustrates a first outline of an architecture for the workbench and its subsystems and components to be first prototyped and then developed in full scale. It is a service-oriented architecture (SOA), where the different modules/services are relatively independent of each other and communicate via well-defined interfaces and messages – as prescribed by the SOA approach.

**Comments to the proposed architecture**

The system to be prototyped could be described as an interactive workbench&toolbox, where users and providers of statistics meet and interact. Wherever the workbench is implemented, it should provide the same “look and feel” for user/provider interactions.

The users focused on in this project are citizens or other actors who work for citizens and/or in close cooperation with citizens, such as entrepreneurs (developing methods and tools for making it easy and attractive for citizens to use and work with statistics), analysts (working with citizen-oriented problems and opportunities), and journalists and politicians (working for the citizens in their respective roles in a democratic society).
The interactive workbench should provide an overview of publicly available statistical data and metadata, wherever they are located, physically and organisationally. Thus the workbench will have, as one of its tasks, the function of being a portal to globally available statistical data and metadata. Links should be provided to the websites where the respective data and metadata are available, and short descriptions and explanations should be given.

Similarly, the portal should refer to available tools and toolboxes for accessing and processing statistical data and metadata, and provide short descriptions of them as well.

The workbench should also provide search tools that enable a user to find and navigate among statistical data that may be relevant for the user’s needs. The search tools should enable the user to work interactively, step by step narrowing down the contents of his or her request, by being able in each iteration step to ask for more precise metadata, e.g. more precise definitions of the statistical data that are available. Proposals for such search tools are available in the literature, e.g. “Statistics By Example” (Sundgren 2006, 2007), inspired by Moshe Zloof’s “Query By Example” (Zloof, 1975), and these proposals should be investigated and experimented with.

The workbench contains two kinds of couplings with statistical databases, (a) one very loose coupling (to the left and to the right of the workbench interface in figure 1), via links to other websites, which in turn interface statistical data and metadata; and (b) one tighter and more genuinely interactive coupling (below the workbench interface in figure 1) with a database, or a system of databases, under common control.

The latter databases, those under common control, may typically be organised in three different ways:

- **Local databases**, owned by a single organisation, e.g. a national statistical agency or an international organisation
- **Shared databases**, owned jointly in some way by a system of organisations, e.g. a national or international statistical system (the Swedish Statistical System, the European Statistical System, etc)
- **Public databases**, open for all providers and users of statistical data, typically residing “in the cloud”, that is, available via the Internet and not tied to any particular physical location, e.g. datasets made available via Dataset Publishing Language (DSPL) and Google Public Data Explorer

The databases (of all types) are interfaced by access and analysis modules. Each such module could be described as a set of well-defined standard services that access data (and associated metadata), specified by the user, from one (or even more) databases by means of services that have to be tailored by the statistics provider to the solutions prevailing for the respective databases. The selected and accessed data (and associated metadata) are delivered to the user in a standard format, chosen by the user, or in a user-defined format.

A user starts his or her interaction with the databases by specifying interactively

- **a request for statistics** (data&metadata contents and format); the formulation of the request can take place interactively and incrementally, using some kind of “statistics by example” approach as described above
- **a processing function**, e.g. a presentation/analysis function, if the user wants such a function applied on the selected data; alternatively the user may specify that the selected raw data (and associated metadata) should be presented as they are, without any further processing
- **a delivery mode**, e.g. delivery to the user’s terminal, file download, or a user application program via an application program interface (API)
- **an update regime**, e.g. data/metadata updated automatically, by subscription, or data/metadata updated on demand from the user only

Two major data structures will be provided: **multidimensional data** (and associated metadata) and **time series data** (and associated metadata); actually the latter structure may be regarded as a special case of the former, but users who mainly work with a relatively small set of time series data prefer convenient short-cuts for such data.

All statistical data dealt with in this project are supposed to be aggregated statistics, so-called macrodata, without confidentiality concerns. Microdata could be managed in a similar way, but for such data confidentiality issues become important.

**A new project proposal**

During the latter part of the present project, “Citizen-centric access to statistics”, the project consortium developed a new project proposal in response to a new Call from Vinnova:


The work resulted in a new project proposal, which is described in a separate document:

Sundgren, Bo et al (2011): “Project proposal: Workbench and toolbox for citizen-centric access to statistics”.

This document, as well as the formal application to Vinnova, including budget proposal etc, are also available on the project website: [https://sites.google.com/site/accesstostatistics/](https://sites.google.com/site/accesstostatistics/)

**Some example of important tasks for the proposed new project**

- Overall design of the interactive workbench & toolbox (cf Figure 1 above).
- Design and testing of the portal part of the workbench, containing links to other websites and short descriptions and explanations.
- Adaptation of the multidimensional model of statistical data and metadata (including both cross-sectional data and time series data) to fit the needs of this project. Testing and evaluation.
- Adaptation of existing metadata models to fit the needs of this project. Testing and evaluation.
- Further development and testing of “Statistics By Example” and other proposals for user-friendly interactions with statistical data and metadata – helping users to find, access, and analyse relevant statistical data for them, making use of available metadata in convenient ways.
- Further elaboration of the use cases that were identified during the preceding project “Citizen-centric access to statistics” (bulk downloads, subscriptions, automatic updates, etc), and trying to implement solutions corresponding to these use cases in some statistical agencies (Iceland, Estonia, Sweden, maybe Denmark and elsewhere as well). Testing and evaluation.
- Further elaboration of the data/metadata formats and associated tools that were identified as important during the preceding project, e.g. PC-AXIS, SDMX, DSPL. Cooperation with the owners of these formats and tools, aiming at smooth use and co-existence of tools and formats for different purposes.
• Identification of suitable presentation and analysis tools to be associated with the workbench. Cooperation with the owners of these tools in order to find practical solutions. Testing and evaluation.
• Development and testing of one or more API:s enabling user application software to access and process statistical data and metadata from statistics producers in a seamless and efficient way.
• Outlining feasible routes for full-scale implementation and use of the results from this projects, including feasible business models for financing further development and maintenance.

References
All documents from the project “Citizen-centric access to statistics” are available on the project website: https://sites.google.com/site/accesstostatistics/ The website also contains a lot of other information about the project, including names, organisations, and email addresses of the members of the consortium and the open network (reference group).

The project is also managing a blog: http://stataccess.blogspot.com/

Other relevant references:


Blog about Stats: http://blogstats.wordpress.com/ A non-official blog, for dissemination professionals of statistical institutions. It is a network and meeting-point, where colleagues share their experiences, successes and failures, focus attention on new developments and stay informed in the vast domain of disseminating statistical information.


Krantz, Peter (2011): "Some links and thoughts" (in Swedish)  PDF


Nuka, Anne (2011): "Implementation of the PSI directive at Statistics Estonia"  PPT

Ott, Arvo (2011): "Estonian eGovernance - policies, frameworks, services"  PPT  PDF

http://www.oecd.org/dataoecd/47/33/40009271.pdf?contentId=40009272

http://www.oecd.org/dataoecd/47/35/40009289.pdf?contentId=40009290

http://www.dst.dk/upload/4.13_are_all_our_users_human_001.pdf (paper),  
http://www.dst.dk/upload/4.13_are_all_our_users_human_slides.pdf (slides)

Sigurdsson, Björgvin (2011a) What problem are we solving? Discussion paper.

Sigurdsson, Björgvin (2011b): “Different users, different user needs, and different solutions”  PPT  PDF

Sillajõe, Tuulikki (2011): "Trying to keep up with users' needs"  PPT  PDF


Sundgren, Bo et al (2011): “Project proposal: Workbench and toolbox for citizen-centric access to statistics”.


http://www.vinnova.se/PageFiles/2777232/Citizens_Services_programme100624.pdf

Services for Research and Innovation networks, 


Åström, Tobias (2011): "Consuming statistical data” PDF

Zloof, Moshe (1975): “Query by Example” AFIPS, 44.

Östberg, Olov (2011): "Some impressions from “Offentliga Rummet” in Trollhättan” PDF