Are all our users human?

Björgvin Sigurðsson, Statistics Iceland
Bjorgvin.Sigurdsson@statice.is

Introduction

In the past 10 years Statistical offices have been through enormous changes in dissemination of statistics. They have, to large extent abandoned paper publishing and relied on the internet for dissemination of both digital publications and statistical datasets. Current availability of statistical data and, time series was unthinkable 20 years ago.

Until now NSIs have allowed people to gain access but what about non-humans? In Statistics Iceland there is an increasing interest in data for systems. Our users want to be able to get latest data directly into their systems.

Roles of NSIs

The outputs of a statistical office are twofold. They produce analysis of statistical data and description of its meaning and interests. This is normally done in a report and press releases. The other output is the statistical observations as they have been calculated and their metadata. This paper focuses on the latter.

In the coming years NSIs will be requested to deliver statistical data in a more standardized way, a way more suitable for computer systems. Many users are already requesting such an access to data and such an approach will create new possibilities in using statistical data in the society.

Statistics Iceland will try to make this approach as a core components in its new dissemination database.

Needs

There are numerous functions in society which have great interest in getting data in an automatic way. It has come as a surprise that many user requests are for small amount of data, usually the most volatile data, such as monthly balance of trade, and monthly indices.

The financial sector in Iceland uses the CPI and other indices for indexation of loans. That creates a lot of demand from various business sectors. Obviously the financial sector is interested in getting the data automatically, so do many IT companies servicing financial software. There is also interest from the transport sector as well as others. These requests are typically content specific, the amount of data is little and the emphasis is on the statistical observations, not the metadata.

Other bodies requesting system to system data are more familiar power users. Those users require more detailed access to data and metadata. They are likely to want large datasets and be more interested in time-series rather than preliminary and volatile data.

The task is to have methods that can provide solutions for different needs, it has to as simple as possible and at the same time efficient to maintain and manage. Perhaps it is time that we serve these different user groups on different terms.

Web 2.0 and metadata

The web is in itself a system to system network. Links are essentially ways for system to travel from system (or website) to another system (website). We have many examples of computer application using them without humans, search engines are probably the most known and complex

systems using hyperlinks.

Recent trends and standards such are RSS, Atom and iCalendar, have allowed systems and humans alike to get access to links in a more sophisticated ways than the traditional hyperlinks. RSS and Atom allow users to subscribe to links getting the benefits of automatic updates, the system providing the service can remove deleted links and the subscription has time awareness allowing subscribers to monitor changes and new material much more effectively.

iCalender is a standard which defines calendar messages, such as meeting requests, appointments and similar. It should be of special interest of National Statistical Institutes. Advanced Release Calendars (ARC) could be provided in iCalender format, thus allowing users to add appointments or reminders to their own Calendars and schedules. The subscribers can be both persons and systems.

The recent standards discussed have also the added qualities of being widely used and well known and what is perhaps most important, they are simple and easy to use. It means that NSIs don´t have to spend time and money in implementing them.

These standards can be useful for increasing systematic access to statistical data. In particular they have the ability to provide access to metadata information. For direct access to statistical data, other methods are more suitable.

Systematic access to statistical data

When NSIs have discussed systematic access to statistical data traditionally they think of super-users who need detailed data in large quantities. As stated earlier the picture is a bit more complex. Businesses have great interest in automatic systematic access to statistical data. As their information systems become more advanced they are seeking outside information to support their own systems. Most queries and requests concern business statistics, such as price indices and external trade. The biggest need is for the most volatile data, the data changing most frequently. These users must be taken into account when NSIs design services for automated systematic access.

Statistics Iceland has identified three different technical approaches available. First option is to develop a simple xml format to describe individual observations. The second option is Web Services, using SOAP and WSDL protocols. Third option is to use a registry. A specialized registry, for statistical data and metadata has been standardized and an open source application exists.[1]

API

Creating xml files for statistical data is user friendly way of providing systematic access to data. It has the benefits of being simple interface to interact with. Most modern Content Management Systems (CMS) can easily interact with xml files and techniques to process xml is widely known and well developed. In addition the solution can be easily developed.

It is however difficult to create simple xml structure which can be used in generic ways. It is therefore impossible to create simple intuitive structure that can be used for all datasets. Users and developers working with many datasets must therefore work with many different types of formats.

Web Services

Web Services using SOAP and WSDL protocols have in recent years gained popularity. They are relatively simple to use. Most major programming platforms, notably .NET and Java offer developers libraries to deal with web services in an efficient ways. Web Services focus on system access to data and as such they are efficient. They require more from users and developers than simple xml files but are much more efficient if the task is to access large datasets and store them within different systems.

---

[1] The registry has been developed within the sdmx standard, see http://sdmx.org for further details.

Registries

Perhaps the most sophisticated way of distributing the data is to use registry to store metadata and then use web services to feed the users of data based on standardized queries. This approach is used in the SDMX registry. The SDMX registry is part of the SDMX standard and will be used to develop the census hub for European data dissemination of the census in 2011. The registry itself stores metadata in RSS and knows how to query for data. The data is delivered using SOAP/WSDL type of Web Services. The output is however standardized as SDMX and queries of data are also SDMX queries.

The registry approach is perhaps the most advanced but demands also most of both developers and users. SDMX registry is based on the idea of standardizing queries and the registry will fetch the correct data. Statistics Iceland fears that the complexity of creating systems which will create to correct queries is not what our customers want. Users dealing with large datasets wanting to create custom queries are much more likely to be interested to download the data and use within other systems, instead of having to create tools to create queries to the registry.

Because of it complexities and high investment for users to utilize fully, Statistics Iceland will not try to setup registry, at least not in the immediate future.

Business Model

In all cases system to system delivery of data will be based on dissemination databases and metadata-repositories. There is considerable development cost involved and cost of maintenance and managing of IT system has to be taken into account. It is therefore appropriate to ask who will pay.

Possible customers

When NSIs started building up dissemination databases online many of them tried to sell access. Most failed in generating revenues from them, with notable exceptions though. What about System to System access, can NSIs make revenue of it? For many businesses it will make sound business case to buy into such an arrangement, provided they are efficient to setup and maintain. These customers already use statistics in their systems but can gain by automating their processes. System to system service can increase the number of such businesses making more attractive to import official statistics into systems for use within businesses. Statistics Iceland has identified some primary customers of this kind. They are various businesses in the financial sector, both banks and consultancies. Large companies especially in transports and tourism are known to be interested and due to indexation of loans software companies selling accounting systems have requested these type of services.

Service for free

System to system access to statistical data will increase usage and efficiency for those interested. By charging customers some users will decide against using it. Free services will maximize the efficiency gains in the society but the problem is that those gaining will not necessarily pay for increased costs, laying more financial burdens on NSIs without gains. By making the service free, the potential customer base can change. Universities and non-profit researchers would surely be more interested.
By making system to system service free NSIs are simply adding another dissemination channel. Thus users can simply choose the channel that best suits their needs. That channel can be paper, website or system to system service.

Future vision

Over the past decade a lot has been achieved in dissemination of statistical data. Next challenges are to allow users to integrate statistical data into their own system, whether they are websites, accounting systems or computer games.

Statistics Iceland strategy is to provide access to data and metadata by using widely used standards such as RSS and iCalendar. Ideally anyone interested should be able to find in its system domains what statistics are available and when they were and will be published. And users should be able to import the data into their systems in a standardized automated way.

The merits of such a vision are not only to create better services for users. Statistics Iceland plans to become the biggest customer of all. The strategy includes using the system to system services to serve the website and other dissemination needs of Statistics Iceland. Data for the organization's website will come from system to system services as will metadata and ARC.

Next steps

Statistics Iceland is experimenting with Web 2.0 technology such as iCalendar for ARC and RSS. Users will be given the possibility to get iCalendar messages for the ARC. Currently ARC, as well as press releases, is available as RSS. All interested users should be able to integrate the ARC into whichever system they like.

In the immediate future users will get access to all statistical tables via RSS. That will enable users to integrate our 4000 table collection into whichever context they prefer.

Over the next few months Statistics Iceland will experiment with using both xml-API approach and Web Services to feed interested users with statistical data.

There are many questions to be asked and the answers must depend on user's needs and experiences. The process of developing new services is a process of experiments, of trials and errors. It is a journey and we can be certain that the destination will change many times over in our quest for still better service to users of official statistics.