SPACE TIME RESEARCH

White Paper

Maybe the World can be Flat
An open standard for sharing statistics across the globe

Summary

Government organisations are responsible and accountable for collating large volumes of statistical data, and making them coherent. This data often forms the basis of important proposals and decisions. There are significant costs and exposure to risk if data is misinterpreted - both for data providers, who can lose credibility, and aggregators, who can suffer significant losses if they are basing decisions on poor quality information.

Well-structured reliable data and metadata is critical to reducing the costs and the risks for many organisations involved in the exchange of government data.

Most shared government data is statistics

Within federal and state government agencies, there is an increasing volume of shared data. Much of this data is in the form of authoritative statistics, used to measure and track a wide range of activities that are captured in operational government systems or deliberately measured through systematic surveys and censuses.

The challenge for agencies is to efficiently and effectively provide and/or collect these statistics and share them with other organisations, and increasingly with the public as well.

SDMX can revolutionise statistical exchange

Open standards and technologies offer an opportunity for organisations to reduce costs and compliance issues and be more proactive and efficient in the collection, analysis and dissemination of data. This paper will describe the challenges to doing so, and how a standard for statistics, SDMX, can help.

SDMX Highlights

- More value from investment in shared data
- Reduced risk of misinterpretation
- Dissemination not batch exchange
- Granular data queries on websites
- Web Services not static pages or bulk downloads (pull not push)
- Data not presentation
- Widely implemented in European Union

Data Challenges

Currently, many organisations are required to share statistical data on a daily basis, both as part of regular operations and in response to ad hoc information requests. For integrators and providers of statistics, what is important is exchanging high quality - or at least, known quality - statistics in a timely fashion, and being able to feed them into important models and reports, which power both businesses and government alike.

Known quality data is what counts

For many organisations that exchange statistics, “known quality” means having sufficient metadata to allow their users to accurately comprehend the meaning, scope and accuracy of the data. This allows users to make informed decisions about how the data is used, as well as understanding what kind of usage is inappropriate.

As statistical data exchange takes place frequently and is growing in importance, the gains to be realized from adopting common approaches are considerable both for data providers and data users.

Australian Bureau of Statistics
on Data Metadata Strategy, November 2009

Despite the critical importance of known quality statistics, many organisations still employ very basic methods for data exchange, such as sharing files filled with raw numbers and very few explanations. This can result in all kinds of costly errors, such as mixing of incompatible versions of data, ignoring definitional changes over time, and incorrectly joining disparate datasets.

Static metadata is easy to ignore

Flawed data exchange processes degrade data quality. For example, even when ample metadata is provided it is often presented in a static, unstructured form such as a PDF file. This places a strong and dangerous reliance on knowledge workers to manually synthesize metadata into data processing routines. A misspelling or a missed annotation regarding a definition can greatly impact the validity of any processing and, as data complexity and volumes rise, the risks multiply.

If both data and metadata are packaged together in a structured, machine-actionable format then system processes can be given a more complete view of the data. This enables automated identification and accounting for many kinds of changes that occur such as updated code lists, changed definitions, and reliability indicators. Even where an automated system might fail because of a significant change, an application can be programmed to at least notify the user so that appropriate modifications or annotations can be made.

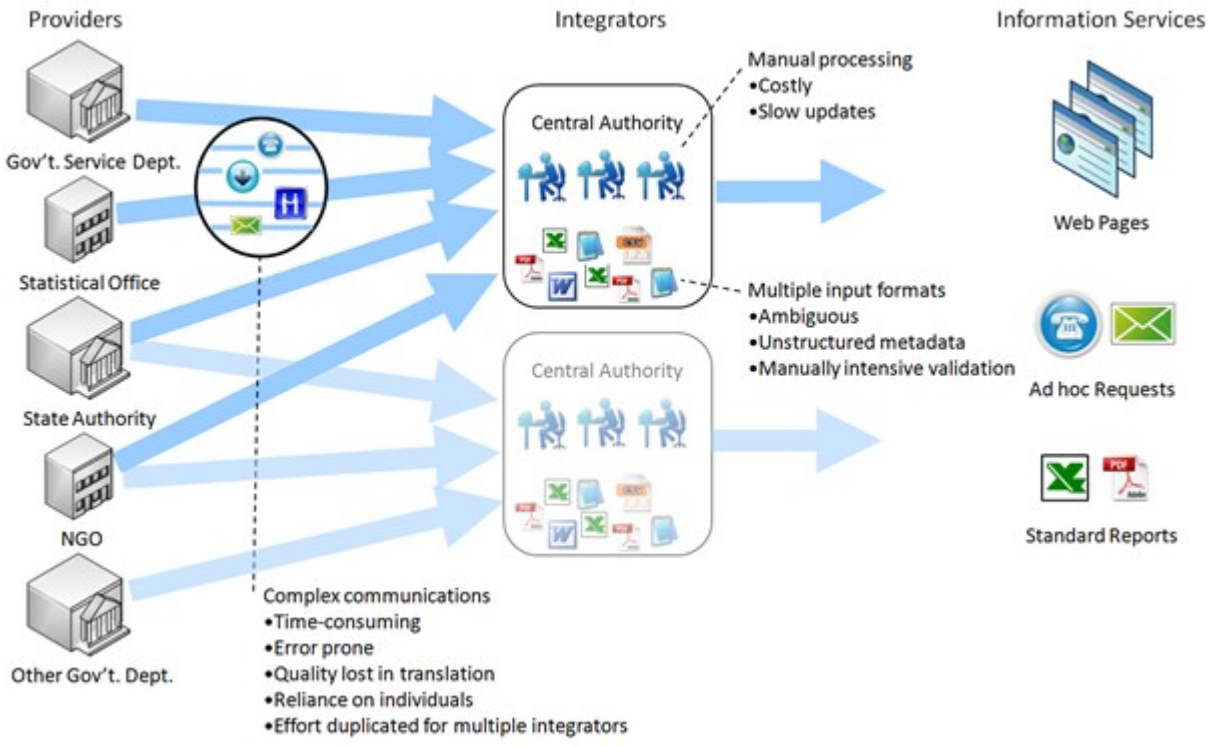
"Even when providers do add useful metadata like notes about series breaks, and reasons for blips in the data, this is often buried in a PDF file and so it's a manual and error prone process to review sets of data from a dozen organisations and try to pull it into a coherent whole."

Dr Stuart Muir
(Chief Analyst, Symbolix Pty Ltd, Associate of the Institute of Analytics Professionals of Australia Limited)

Transportation Example

Consider a national transportation department that needs statistics on bus usage to determine the allocation of funds from ticketing revenue to the various transport providers, as well as for planning improvements to public transport infrastructure. There are many issues in obtaining a set of consistent travel statistics across a nation. Some examples:

- One smartcard system might require passengers to "tag on" when they board and "tag off" at the end of a trip so the trip distance can be accurately measured, but another system might work using travel zones and have no data about precise trip lengths.
- Some data may be obtained through sampling, whereas other records might be derived directly from ticket data.



- Data can be skewed by one-off events, and the effect of this can easily be overlooked unless appropriate metadata is available during analysis.

The problem the transport department is faced with is how to ensure that

- differences, due to issues such as quality, level of detail, and coverage, between different transport providers' reports are avoided where possible
- where the differences are unavoidable, there are clear and accessible explanations in the metadata.

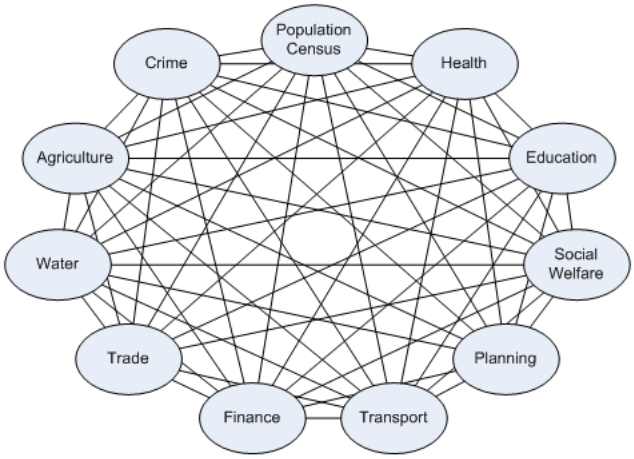
Given the complexity of scenarios such as this, and the problems of existing exchange methods, how can organisations improve the exchange process?

A Common Statistical Language

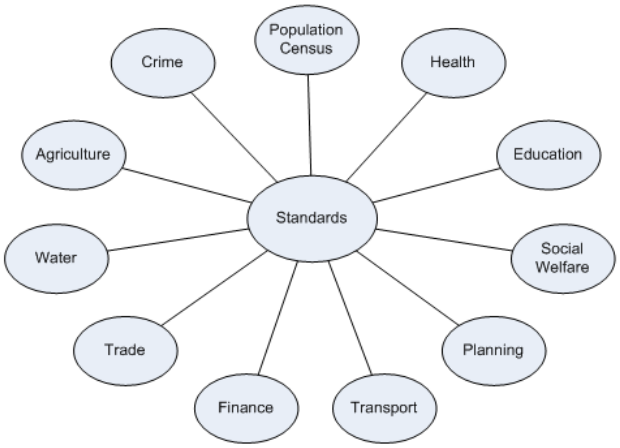
An effective solution to the problem of statistical exchange requires all organisations involved to agree on a process for doing so, and for them to be able to execute consistently against that agreement. The process needs to be as easy as possible without compromising its goals and this is where a common statistical language can help.

Eliminating bilateral exchange agreements

A language designed specifically to deal with the general issue of exchanging statistics can simplify the data exchange process enormously. It means that, in many cases, there may not be a need for a specific bilateral agreement between two organisations because the language is good enough to get a match without any negotiations on technical data formats.



Bilateral agreements



Data Sharing through standards

The right language must be able to describe consistent, comparable statistics, with detailed information about the accuracy and meaning of the data. Additionally, it should allow processing tools to examine statistics holistically rather than just working on the raw data. That means supporting combined, actionable data and metadata. Doing so in a way that leverages existing, well-accepted technology standards such as XML helps to accelerate and simplify adoption.

A language to increase automation and efficiency

The use of a comprehensive language for statistical exchange enables automated processes that ensure both data *and metadata* comply with organisational business rules, which reduces the chance of manual errors and improves efficiency. In the transport department's case, critical metadata could be examined by validation processes, which can in turn feed into models and reports, all potentially without human intervention.

A comprehensive language helps to maximize consistency across the different data sources, while still informing the knowledge worker or external stakeholder where inconsistencies remain and the reasons behind them.

Common standards and guidelines followed by all players not only help to give easy access to statistical data, wherever these data may be and without demanding prior agreement between two partners, but they also facilitate access to metadata that make the data more comparable, more meaningful and generally more usable.

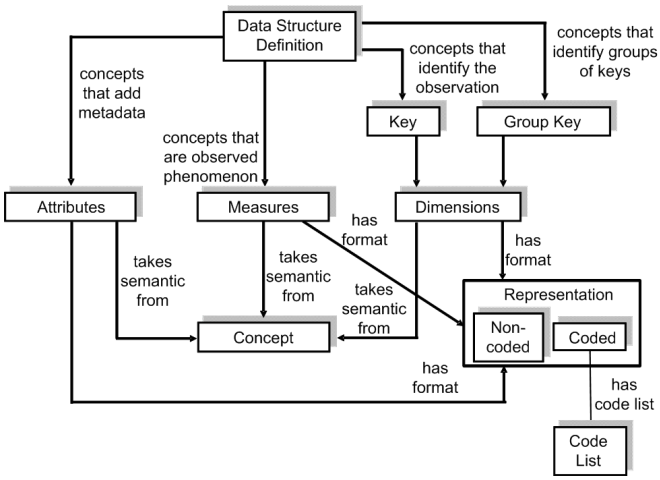
www.sdmx.org

Statistical Data and Metadata eXchange (SDMX)

The SDMX standard was created for the very purpose of allowing organisations to efficiently and automatically share statistical data and metadata and, *unlike any other existing standard*, it addresses all of the requirements of a common statistical language. Because it is not specific to any one business domain, it is capable of mixing data from many different sources , which is what many organisations commonly need to do.

A very common and basic requirement is to put statistics in the context of the whole population. An aggregator could do this by pulling SDMX from census statistics, combining it with SDMX made available from a criminal justice department, and then reporting on crime figures per 1000 head of population.

At the core of SDMX is a model of a statistical dataset, called a data structure definition, which is broken down into reusable structures such as dimensions, concepts and attributes. Collectively, the structures defined under this allow for any kind of statistical output to be described, complete with data types, relevant annotations and code lists.



SDMX Data Structure Definition

SDMX includes a complete model of a data cube

In general, this means any kind of table or data cube can be described unambiguously and without any customized extensions on top of the model, ensuring that data defined by different organisations can remain compatible without any further transformations.

Users of RDBMS solutions may wonder what a star schema model is missing that SDMX can capture. The answer is compatibility and completeness for exchange. There is no common comprehensive exchange format that helps organisations exchange data if they are using different RDBMS products, or indeed statistical packages. Also, a star schema by itself is not a complete description of a statistical dataset.

Without custom extensions, a star schema cannot capture metadata such as descriptive names, additional descriptions of any item in the model, cell annotations, currency information, additivity, periodicity, and many other attributes. It also does not have built in multilingual support, which is an integral part of SDMX.

SDMX can be a bridge between organisations

Naturally, it is quite possible to define all this in an RDBMS but the power of SDMX is that it enables any group of organisations to exchange data without first agreeing on all the relevant extensions to whatever data format they may agree is the lowest common denominator. Internally, they are able to continue to use their existing systems, which may vary considerably, but they can use SDMX as the language for any kind of statistical exchange.

SDMX Web Services

Web Services have a big advantage over more static solutions for sharing statistics because they enable a “pull model”. This reduces network usage and, more importantly, makes it possible for web applications and processes to automatically retrieve data only when it is required and with no manual involvement.

While other Web Services solutions can make this possible, the fact that SDMX services are based on a statistical model ensures a much more complete fit that eliminates the need for large, unstructured documentation to complement and explain how the services can be used. They also eliminate the compatibility issues that exist when, for example, different relational schemas are used by different providers.

All institutions considered SDMX would be useful for their organisations, with close to 70 per cent indicating it as very or extremely useful .

SDMX Global Conference,
Paris 2009

The 2.1 version of the SDMX standard includes a REST API, meaning that developers will be able to interact with the SDMX model through simple HTTP requests, enabling greater testability, rapid development, and an inherently more scalable solution for providers.

Enabling mashups

Because the SDMX Web Services encapsulate metadata, it enables new kinds of tools, such as mash up engines, to identify and match against relevant output. For example, a third party application could interactively build queries in response to a user's actions and allow the user to browse across multiple datasets and join data that shares common dimensions.

SDMX Registries

More advanced users of SDMX can employ a registry to allow for a spoke and hub architecture, where organisations contribute to a central registry. The organisations retain ownership and full control over their data, but the registry allows for them to automatically publish the existence of any available statistics and make them accessible to other authorized parties. An example of such use is the European Census Hub project.

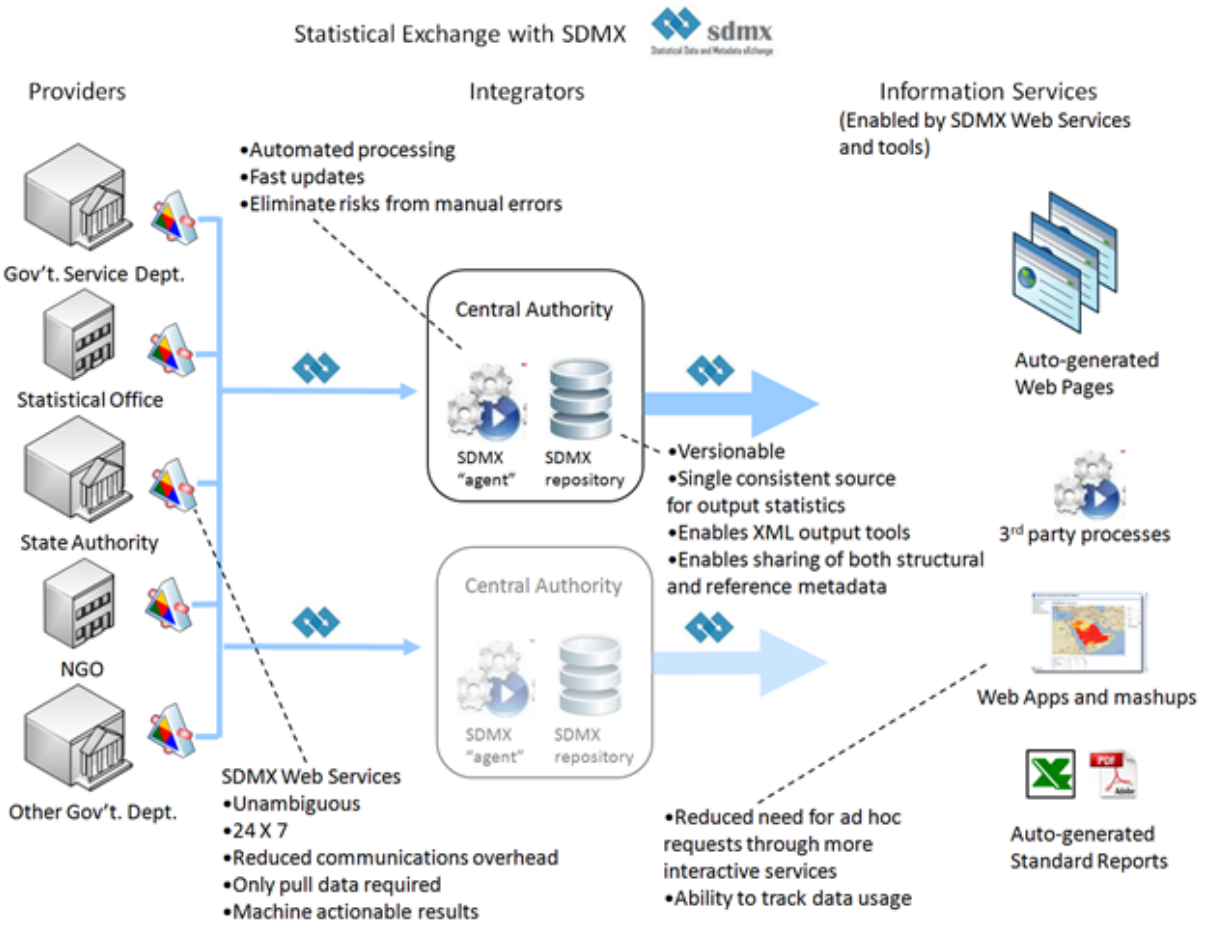
More on SDMX

While the data structure definition is at the core of SDMX, there are also constructs specifically designed to inform and drive the exchange process itself. For example, constructs such as versioning, which supports accurate interpretation of data that has varied in format and content over time, and the concept of a data flow, which can link many similar datasets from different sources into one coherent series. SDMX has been designed to allow effective use of subsets of the model such as the data structure definition, meaning that organisations can achieve a useful result from just a partial adoption of the standard.

SDMX supports other established standards

SDMX has also been designed with other adjacent or overlapping standards in mind. For example, SDMX

- supports Dublin Core metadata
- can be used to construct data elements based on the ISO/IEC 11179 metadata standard
- is being developed to complement the increasingly popular standard for the collection, analysis and dissemination of microdata, DDI.



Conclusion

SDMX can act as a complete, actionable format for the exchange of statistics. It provides a way for organisations sharing statistical information to deliver packages of statistical data and metadata in a modern and comprehensive format that is easy to consume. The use of Web Services enables exchanges to occur based on a "pull model", reducing network traffic, enabling automation, and helping to improve efficiency. It eliminates the degradation of data quality that results from traditional, poorly defined and non-standard exchange methods.

Ignoring or tolerating poor quality data can have a major detrimental impact on many organisations. SDMX is a reliable way to reduce the risks of this happening, and to take advantage of more efficient and effective practices for the exchange of statistics that underlie so many important functions of government.

Further information

- The Official SDMX website, official standards and guidelines for SDMX.
- SDMX User Guide – more detailed information that does not assume any prior knowledge of SDMX.
- SDMX and DDI comparison— note exploring relationship between these two standards. Open Data Foundation.
- Working document on data metadata strategy. Describes benefits of adopting SDMX and DDI. Australian Bureau of Statistics.
- SDMX, ISO 11179 and the CMR — explores mappings between SDMX and ISO 11179. Metadata Technology 2006.

About Space-Time Research

Space-Time Research is a leader in data transparency solutions for providers of official statistics. SuperSTAR analytics and visualisation improve information accessibility by ensuring that authoritative statistics are accurate and correctly understood. Space-Time Research solutions improve the operational productivity of statistics departments, with particular focus on statistical production and dissemination, information privacy protection, and interactive visualizations.

Space-Time Research Pty Ltd
Level 1, 386 Flinders Lane
Melbourne 3000
Victoria, Australia
Tel: +61 3 9615 5200
Fax: +61 3 9615 5299
Twitter: @spacetimeresrch
info@spacetimeresearch.com

www.spacetimeresearch.com

Copyright © Space-Time Research Pty Ltd