Distr.
GENERAL

WP.28
30 March 2010

ENGLISH ONLY

UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE (UNECE)
CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE EUROPEAN UNION (EUROSTAT)

ORGANISATION FOR ECONOMIC COOPERATION AND DEVELOPMENT (OECD)
STATISTICS DIRECTORATE

Meeting on the Management of Statistical Information Systems (MSIS 2010)
(Daejeon, Republic of Korea, 26-29 April 2010)

Topic (iv): Innovation and related issues including census systems

An SDMX-based unified data catalogue (UDC)

Prepared by G. Becker & M. Bruschi, Bank for International Settlements (BIS)

I. Introduction

1. The UDC Prototype proves that the "SDMX Vision" is real: users can – via a central data portal, the Unified Data Catalogue – search and retrieve data from different types of internal and external data sources: SDMX-ML files on websites, SDMX conformant databases with a web service and databases that can be "mapped to SDMX". The prototype uses an "off the shelf" SDMX registry implementation, a custom-made GUI (the UDC itself) and custom-built web and data query services to the different connected databases. The paper reports on the technical implementation of the prototype, the conclusions of the feasibility study and also points out issues that need to be addressed to move the UDC concept forward.

II. Vision and concept

A. The SDMX vision

2. Users of statistical data want to work with the most up-to-date numbers, want documentation about the data they are using and want to able to find the "good data" to work with. Currently the same data may be provided by different organisations, e.g. the National Accounts data of a country may be provided on the websites (databases) of the National Statistical Office, the central bank and also on the websites (databases) of international organisations to which this data is reported. The user thus has the problem to decide which data source to use, the original data provider has to send the data to multiple organisations and, in the end, we have the duplication of the same data, possibly with slight differences with respect to their "freshness", in different databases around the world.

3. Let's assume, the original data provider of the published National Accounts were to make this data available in SDMX formats on a website (and/or via a web service), together with the information about the data structure used. All potential users, including other organisations, could pick up the information and either use it directly in their work or download it and stored it in their own local (SDMX conformant) environment. Ideally, in such a set-up all data would only be stored once, by their original provider … and

"some magic" would allow the users to find this data at this single storage place. The conceptual ideas behind SDMX say that this should be feasible, with the use of the SDMX technical standards, SDMX web services and, in particular with the use of one or more SDMX registries.

B. The Unified Data Catalogue (UDC) concept

4. The Unified Data Catalogue (UDC) tries to implement the SDMX vision: a single data catalogue allows the user to discover, select and retrieve statistical data from all registered data sources. For the user it should not make a difference whether that data is stored locally in a database, in SDMX-ML files or whether it comes from an external source (eg a web service or SDMX files provided on a website).

5. In order to "discover" data the user needs to have available the metadata for this data, eg the information provided in an SDMX data structure definition (DSD). In the DSD we have the key structure (ie how the dimensions for the time series identifiers are put together) as well as the code lists for these dimensions. A key part of the UDC therefore is the availability of the DSDs for all internal and external registered sources. The SDMX Registry is the natural place to store this information. It can also hold additional information that can facilitate the data discovery, such as a Category scheme, into which all registered data sources can be categorised. This provides the user with a high-level overview of the available data sources into which s/he can then drill down. The registry can also hold the information on the "connected" (= registered) data sources, eg web services ("queryable databases") or static SDMX-ML files.

6. In its simplest form the UDC basically provides a graphical user interface (GUI) to the registry information. It allows to view the categorised data sources, to select a data source for closer inspection. In practice this means that the DSD is loaded and provided in the GUI so that the user can then make selections based on the code lists for the key family dimensions and attributes. Once the selection is done, the UDC sends the query to the registered source and also handles the response, ie displays the data retrieved.

7. It is important to note that the basic "lingua franca" between all components of the UDC (the GUI, the registry and the data sources) is SDMX. For databases that are already natively SDMX conformant, this is not an issue. However, if we want to connect a data source that is not natively SDMX conformant, we need to set up a mapping between that database's structure and an derived SDMX DSD, which can then be loaded into the registry. We also need an adapter that can translate the SDMX query produced by the UDC GUI to the possibly different query language for the database and also a transformation from the database's output format to the SDMX DSD. With this additional step we are in a position to make "SDMX mappable" legacy data bases "look like SDMX".

II. Prototype and issues found

A. The UDC Prototype architecture

8. We started the feasibility study project putting together a set of user stories. These were built around the key idea that we wanted to learn using an SDMX registry as the base component for a system designed to help a user find statistical data in an heterogeneous and distributed environment. We put together some 40+ stories that spread on several functional and usage areas: from registration of metadata in an SDMX registry to the GUI features of the user interface of the UDC, covering navigation and search of the metadata, query and retrieval of data from the source repositories, usage of constraints, output handling, automation and security aspects like authentication and authorisation.

9. Behind the UDC concept is a very simplistic approach: to search and retrieve data from a data source all what we need to know are the source data structure and its query language. If a data source has a structure following the SDMX information model (IM) the only other thing needed is a (web) service connected to this source able to respond to the SDMX query language. If a source data structure can be "mapped" to an SDMX IM then the query (web) service is also requested to implement this mapping, ie it also works as an adapter.

10. Following these simple ideas we can categorize data sources and systems: they are SDMX-enabled if they are natively conformant to the SDMX-IM and support a SDMX-ML query interface, or if adapters can be built so to make them exposing and SDMX interface. At the simple level even a data file in SDMX-ML (supported by an SDMX DSD) is a data source, and a generic "file-query-handler" can implement this SDMX-enabling interface.

11. The schematic architecture for our UDC can be seen in fig.1.
The key components of our prototype are:

(a) An SDMX registry. We used an "off the shelf" implementation from the SDMX.org website.

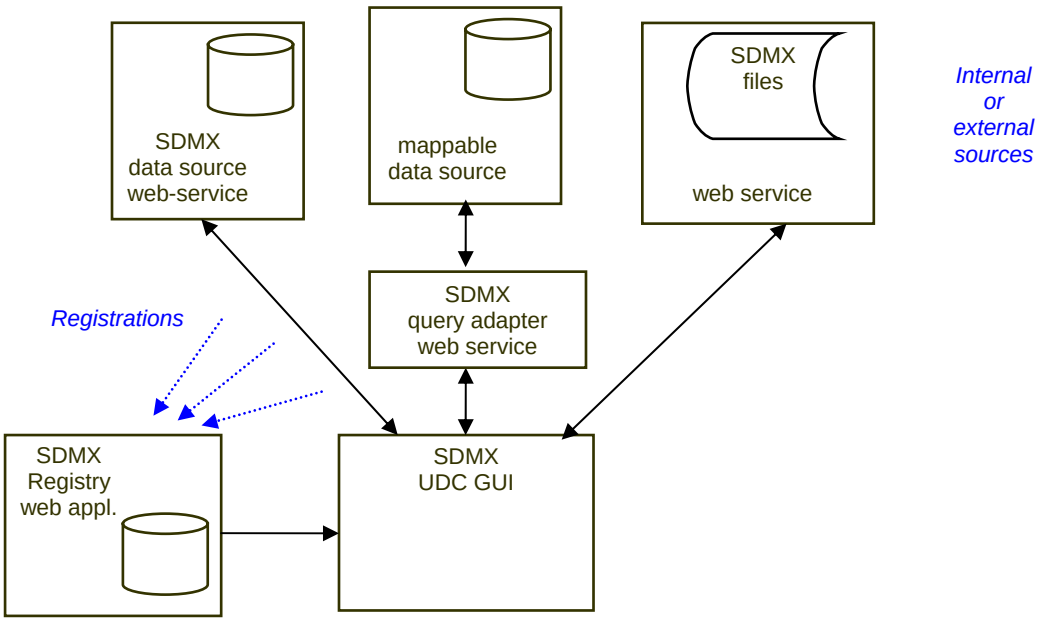

Fig. 1 – UDC schematic architecture

We wanted to build experience in using the registry by making it the central repository for the data structure definitions (DSD) of all the "connected" data sources (internal or external) and for the registrations of all the related data flows. In particular these materialise as URLs of the SDMX-ML files and query web services that can respond to data requests. The registry itself is a service that can be updated either via SDMX-ML messages or via an administration web based interface.

(b) The UDC GUI, developed for the feasibility study. Our web-based user interface allows the user to navigate the information stored in the registry, to interactively build data queries based on the DSDs and to submit them to the registered query-able data sources and, finally, to retrieve the matching data from the sources and present them to the user.

(c) A set of SDMX query and adapters web services, developed for the study, for different types of data sources. Responsibility of these query handlers is the mapping from the SDMX-ML representation of the query to the source specific model and query language and to format in SDMX-ML the data matching the query.

B. The UDC Prototype data sources and functionality

12. With our prototype we succeeded to provide central access to a variety of data source "types". In particular we developed SDMX query adapter that allow us to search and retrieve data from:

(a) The BIS Data Bank, a time series repository implementing a "generic" SDMX-EDI base model, i.e. the system is able to self configure itself to host data for a statistical key-family after having

received its DSD. Observation data is stored in FAME while the other data is in an RDBMS. This system offers an own query interface (SDMX-EDI doesn't define a query language) so our adapter had to implement a map across the two query languages. This has highlighted issues in the interpretation and expression ability of the SDMX-ML 2.0 query schemas.

(b) MSTAT OLAP, an internal system offering multi-dimensional access to the BIS International Banking and Financial Statistics. To support our internal analysis and publication processes we had already implemented for this system a time series access interface able to map an external SDMX DSD model to its internal OLAP / MDX structure. Our adapter can then leverage on this interface and implement a simple SDMX query access.

(c) A relational and OLAP environment for research data. In this case we were interested in the conceptual and technical challenges implied by defining a DSD for an "unstructured" dataset and in defining a relational data access interface (SQL based) that could be used when converting from an SDMX query.

(d) SDMX-ML formatted data files, created in-house or picked up from public websites. To this purpose we developed a generic web service that accesses these files and implements a query handler able to filter from the data in the file to match the query conditions.

13. Based on the SDMX-IM, the key features offered for our prototype GUI are:

- Browsing of categories, data-flows and provision registrations
- Browsing of a selected DSD: dimensions, attributes, code-lists
- Building of queries based on the DSD, via code selection
- Execution of the query and visualization of the results as a simple table
- Possibility to download the result table and the relevant DSD in SDMX-ML format
- Search by concept / code-list, a Google-like search that, starting from the matching text descriptions identifies which DSDs use the matching entries and proposes to search the related data-flows.

C. Some results and issues found

14. With this implementation we demonstrated that a UDC can provide centralised user access to an heterogeneous and distributed set of (SDMX enabled) statistical data sources. Our systems are based on different operating system platforms (Windows and Linux), database systems (FAME, SQL Server, Sybase), development environments (.Net, Java, Perl). We used our own SDMX-ML data files as well file made available from other institution on their web-sites. Where needed and possible we used the tool made publicly available to support the SDMX initiative. Furthermore we experienced some of the impacts the usage of an SDMX Registry system can have on the organisation and work of statistical data managers. This clearly supports our "SDMX vision". But we also fund practical issues…

15. At the time we implemented our prototype (i.e. with the version of the tools and data files available in August 2009) we faced the following issues:

- Registry implementation did not allow to register files in compact or utility format
- The SDMX 2.0 query message schemas do not allow to properly express some user queries
- Registry implementation did not support constraints processing
- Incompatibilities between the formats and versions supported by the tools we used and the files we gathered from some institutions' websites, as well as the unavailability of the underlying DSDs.
- Cross-platform communication with security not solved: authentication and authorization seems to be particularly challenging when the services involved are implemented on different base technological frameworks (Java vs. .Net). Interoperability is still complicated.
- In general: access authorisation to query-able data sources is unresolved. This is not an SDMX issue, but we all know that not everybody has, or should have, the same access to all the data. How this can be handled in an heterogeneous, multi-owner network needs still to be defined.

- In the prototype we updated the registry more or less manually (by loading SDMX-ML files). In the real world this would need to be done in an automated way.

III. Conclusions

16. Based on this short feasibility study we see potential for a UDC–type application as an integral part of a future statistical infrastructure for the economic research work performed at central banks and international organisations. It will take a number of years to get there … the issues that need to be resolved are not only technical (as listed above) but also relate to applying the SDMX technical standards in a coordinated and cooperative way.

17. The agreement "to use SDMX for publishing data" is a first and important step. However, SDMX is a rich standard and there are, for example, three different data formats to exchange time series information and related metadata (compact, generic and utility). A registry and UDC that need to cope with both format types will be more complex to set up and maintain than one that only needs to deal with one format. Adding the support for cross-sectional SDMX will further increase complexity. The same holds for different ways or styles of defining data structure definitions. The issue is not the fact that there are different DSDs but more that organisations may interpret the SDMX standards in a slightly different way, possibly driven also by the way their statistical IT systems are set up. This can lead to different "DSD dialects" which the UDC would need to cope with.

18. The SDMX registry holds all crucial information to make the UDC work. At this point in time it is not yet clear how many registries will be set up and by which organisations. Would the UDC need to deal with a single registry or a set of federated registries? How can we ensure that the UDC always has the most up-to-date information about the data sources and the DSDs used? While this is also a technical issue it is possibly more so an issue of agreeing within the SDMX community how to practically organise the SDMX registry landscape.

19. A key conclusion from our work is therefore that in order to pursue the idea of a UDC we need to sort out a set of technical issues but also need to cooperate with other SDMX data providers, possibly convince them of the UDC idea and cooperate with them so that we can arrive at a common way of publishing our data with SDMX. This will move us a step closer to realising a UDC.