Distr.
GENERAL

CES/AC.71/2005/20
17 March 2005

ENGLISH ONLY

**UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE (ECE)
CONFERENCE OF EUROPEAN STATISTICIANS**

**EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE
EUROPEAN COMMUNITIES (EUROSTAT)**

**ORGANISATION FOR ECONOMIC
COOPERATION AND DEVELOPMENT (OECD)
STATISTICS DIRECTORATE**

Joint ECE/Eurostat/OECD Meeting on the Management of Statistical Information Systems (MSIS)
(Bratislava, Slovakia, 18-20 April 2005)

Topic (iii): XML and web services

AUTOMATED ACCESS TO 100,000,000 STATISTICAL FACTS VIA STATLINE4 WEB SERVICES

Invited Paper

Submitted by Statistics Netherlands[1]

Abstract: The Internet has been evolving increasingly from a relatively simple hyperlink based web into a huge set of loosely coupled data repositories. Statistical databases, as a valuable information source for statistics about society, economy, environment and finance, should become an integral part of this loosely coupled data web. A statistical database can accomplish this by providing *automated access* to its contents. Content providers may use this facility to add the appropriate statistical background information to their services. Others may use it to be informed of certain specific updates in a specific statistical area. In this paper we present some of the new features for automated access to the statistical output database of Statistics Netherlands (StatLine4). We briefly describe the design and functionality of a set of web services, both from a technical perspective as well as from the perspective of an end-user. In addition, we compare this design with a related approach, the approach taken by the International SDMX Open Data Interchange (SODI) working group in which Statistics Netherlands participates and we highlight the main differences.

Key words: Statistical Information Systems, Statistical output databases, XML Web services, Statistical Metadata, StatLine, SODI, SDMX

I. INTRODUCTION

1. Statistics Netherlands publishes statistical facts in a statistical output database, called StatLine. This statistical output database evolved, in about 10 years, from a quite simple web-based application for browsing statistical tables, into a multidimensional database, which can be accessed via the Internet to browse its contents. Currently, StatLine contains *all* the statistical results published by Statistics Netherlands, which is a total of approximately 100 million statistical facts. The size of StatLine continues to grow, as the internal

[1] Prepared by Olav ten Bosch (obos@cbs.nl), Edwin de Jonge (ejne@cbs.nl) and Erik van Bracht (ebct@cbs.nl).

policy prescribes that nothing is to be published unless it is published in StatLine as well and that all data that has been published must remain available.

2. Due to the huge amount of rather heterogeneous information in this database, finding a specific statistical fact is not always straightforward. Although an intelligent search engine assists users by pre-selecting statistical tables and although the website of Statistics Netherlands directly points to the most important key figures in StatLine, end-users still need some knowledge of the navigation principles being used and the content structure to find more specific information. Therefore, better ways to access StatLine are still needed.

3. The upcoming version of StatLine, StatLine4, will offer a simpler and therefore more intuitive user interaction mechanism. But this is only one improvement. We noticed that many institutes are frequent users of very specific parts of the information in StatLine. Clearly, these StatLine users are most interested in those parts of the database that conform to their specific business needs and they consult these parts more frequently. They even put bookmarks on their web pages that point directly into a specific part of the statistical database (deeplinking). This poses the question how a statistical database on the web could better serve this kind of frequent and *partly automated access*.

4. In order to support this kind of access, StatLine4 will offer a number of XML web services. Customers may use these services to automatically check certain statistical indicators at regular intervals or to automatically search the content of the statistical database and interpret the results in a way they want. One of the clients of this feature is the main website of Statistics Netherlands itself.

5. In this paper we address the practical design of automated access to StatLine. We describe which design principles were used and what services are currently being developed. Since StatLine4 is not in production yet, we cannot report on the actual use of these services until next year. In the first section of this paper, we briefly describe base functionality of StatLine. In section III, we motivate the increasing demand for automated access to statistical databases in general. In section IV we describe the main access methods to StatLine4. In section V we compare this approach with the approach taken by the SODI project, in which Statistics Netherlands participates and in section VI we present some conclusions.

II. STATLINE IN A NUTSHELL

6. These days, national statistical institutes (NSIs) cannot do without a proper way of disseminating statistics via the Internet. Most NSIs have an advanced content driven website which provides access to a statistical database in the backend. Users may browse these databases by means of their metadata. Statistics Netherlands is not an exception to this approach. Its statistical output database, StatLine, has been around for about a decade and is still growing both in size as well as in functionality.

7. Comparable to most statistical output databases, StatLine is conceptually a database of statistical facts organized in multidimensional *cubes*. These cubes typically have 3 to 8 dimensions, which describe the subject and the statistical classifications of the statistical facts. Examples are a regional classification or a classification of education types. Dimensions may in itself be hierarchical or even a graph like[2] structure. For example, this is the case for the regional classification of the Netherlands. StatLine cubes are based on the Cristal [1,2] information model for statistical data and metadata. This model also contains constructs to correctly handle changes in metadata such as classifications. In addition, StatLine supports different versions of statistical facts [4].

8. Basically end-users browse the database by selecting the appropriate elements in the dimensions. Based on these selections StatLine returns a dynamically generated table. If nothing is specified in one or more dimensions, a default selection is used, which was specified by the statistical expert that created the statistical

[2] In fact, in the Cristal model classifications are partial orderings of categories.

cube. The system has a search facility that generates meaningful selections in cubes based on a search string. Statistical data can also be presented in charts or cartographic maps. StatLine offers some more ways to browse its contents [3] one of which is a geographical entry that enables users to select certain statistical information using a map.

9. Cubes are organized into hierarchically organized statistical *themes*. A cube may be present in multiple themes. As with any multidimensional statistical cube, certain cells of cubes cannot be filled for logical reasons. Some multidimensional cubes therefore tend to be sparsely filled with data. StatLine4 provides a smart selection mechanism that tries to prevent end-users from making selections containing empty cells. The StatLine4 search engine makes meaningful selections as much as possible without empty cells.

10. The layout of tables presented in the browser can be manipulated by the end-user. He may rearrange dimension variables into rows or columns via a drag and drop mechanism. StatLine has a hyperlink mechanism, called dynamic user adaptable link (DUAL), which makes it possible to refer to any selection from the statistical database via one hyperlink. The main website of Statistics Netherlands uses this mechanism to point visitors directly to the main indicators available in the underlying statistical database. In addition, press releases and statistical articles are annotated with DUAL links, which direct readers to the relevant part of the statistical database.

III. AN INCREASING DEMAND FOR AUTOMATED ACCESS

11. Being a content provider, heavily covered by press and electronic media, statistics published by Statistics Netherlands are referenced a lot. More and more, we see consumers of these statistics deeplinking into the statistical database itself. Examples are certain portals of the Dutch public library, parts of wikipedia [7] and some portals for information about agriculture [8] or Dutch nature [9]. These portals use the already available DUAL mechanism to include the latest statistical information from Statistics Netherlands on their portal pages.

12. Although the current linking mechanism is powerful enough to deeplink any part of StatLine and although it has some nice flexible constructs for referencing content that will be updated periodically, it also has a number of shortcomings. For example, it doesn’t provide an easy way to detect the availability of new figures within a certain theme, which of course is common use in statistics. Also, it doesn’t provide a way to automatically call the StatLine search engine on certain subjects and automatically process the results in a way Google can be approached via the Google API. Finally, it is a proprietary syntax, which must be learned before being used.

13. The increased demand for automated access to StatLine, has led to the conclusion that the hyperlink mechanism must be improved. It needs to be extended in functionality and it must be based on industry standards as much as possible. Therefore, we decided to extend the DUAL mechanism with a more general set of web services, transforming it into a *statistical web service* (see also [5]).

IV. THE DESIGN OF STATISTICAL XML WEB SERVICES

14. When designing a set of statistical web services for a huge statistical database, we had to make a choice between completeness of functionality and simplicity. If we want to provide automated access to all metadata in the database we could easily end up with tens of different web service definitions. However, we want third parties to easily connect to the statistical database. Therefore it’s important to keep the number of services small and to keep them simple. When designing this facility for StatLine, we took the latter approach and opted for a set of web services that are *simple* and *effective*. Being complete by providing all possible functionality was less important. Nonetheless in StatLine 4 we use these webservices ourselves (“eat you own dog food”).

15. Conversations with some of our customers showed us that they would like to have a way to access the search engine of StatLine. This indicated the need for a *search* service. In addition, it appeared that there was a need for getting informed of certain changes in content, for example when a new statistical fact in a time series was added or a temporary result was published as final. This resulted in an *update* service. These services both have to do with access to statistical metadata. Logically speaking, clients also need to be able to access the statistical data itself. There was already a mechanism for this: the DUAL mechanism mentioned earlier. For simplicity, we chose to embed this facility into a new *data* web service. Below, we give a brief description of each type of web service.

16. The *search* web service offers external users the possibility to search for Cubes or Classifications in StatLine. The user may specify the words that need to be found (query), the starting index of the hits returned and the maximum number of hits that need to be returned. The starting index can be used by clients to split up the result list in smaller chunks for further processing, comparable to the Google API service. The return value of the search web service is a hit list of found items, containing label, date published, key, DUAL etc. For most users this will do. It is possible though to do advanced search queries by specifying a complex query string, in the Lucene [10] syntax.

17. The *update* web service lets external users choose whether they want to be notified of updates and if so, on which level they want to be notified. Since the really simple syndication (RSS) protocol has become a de facto standard for this kind of update services, we adopted this format in StatLine as well. In this way, standard RSS aggregators can be used to detect StatLine updates. Three webservices were defined. The first one, called *GetRSS*, notifies users about any update in any cube of StatLine. Since most users are only interested in statistics about a certain theme, we also added a more detailed notification mechanism, which is called *GetRSSOfTheme*. This service notifies users of updates in cubes within a certain theme. Finally, for users with a very specific interest StatLine offers an update services called *GetRSSOfPublication,* which notifies users of updates in one specific cube. All results of these services are published in RSS 2.0 format.

18. The *data* web service gives access to the data in a cube. For this, one has to provide a selection in each of the dimensions of the cube. The web service user can easily obtain this by making a selection in StatLine manually and copy the resulting DUAL hyperlink. StatLine returns a simple table format containing the data requested. In addition, there is another web service that returns one bare statistical fact only without any metadata. This service is tuned for efficiency and may be of interest to clients that frequently retrieve one specific item from the database. The data web services provide backward compatibility with earlier versions of StatLine. Due to the intensive use of the DUAL mechanism, we don't want old references into the statistical database to malfunction. Using this new web service the old links remain valid. Because the data being retrieved may be processed automatically, the service opens up new possibilities: web portals can embed the latest figures of Statistics Netherlands on their homepage or companies can use StatLine directly from their proprietary systems.

V. THE SODI APPROACH

19. Recently, Eurostat started a pilot project, called SDMX Open Data Interchange (SODI) [11]. The aim of the project is to make certain short term national data (i.e. quarterly GDP data and monthly industrial production indices) quickly available in a common dissemination environment of Eurostat. For this purpose the pilot investigates two data communication schemes between European NSI's and Eurostat: *push* (GESMES oriented) and *pull* (web service oriented).

20. Together with two other participating members Statistics Netherlands decided to experiment with pull technology and web services [3]. The information interchanged with this technology is formatted in an emerging standard XML structure designed by the SDMX organisation [6], called SDMX-ML.

[3] The other member states participating in the SODI task force are Germany, France, Sweden and the UK.

21. At time of writing, the SODI pilot web service definitions are still under development. However, with our current experiences the main implications for this kind of information interchange becomes clear and a few cautious and early conclusions can be drawn:

- The original SDMX-ML schemas are quite generic.
- The consequence is in practice that for each data flow, for the quarterly GDP data as well as for the monthly industrial production indices, a more specialized subset of SDMX-ML must be designed first.
- Each specialized subset of SDMX-ML leads to the development of a corresponding specialized web service at the side of the NSI.
- Therefore the implication for a NSI in the current setup is the development of a separate web service for each data flow to Eurostat.

22. On the one hand the development of a special web service for each data flow to Eurostat does not seem efficient for NSIs, even if we disregard any additional burden due to future changes in the SDMX-ML subset designs.

23. On the other hand, as the specialized SDMX subset web services have many things in common their development can clearly be made more efficient by more generic web service designs at the side of NSIs.

24. To be efficient, generic software designs should facilitate a flexible matching of the output metadata structures of an NSI to the metadata structures referred to in the requests of Eurostat. In fact a generic solution to this metadata matching problem stays the most burning issue to tackle for efficient information exchange between Eurostat and NSIs, but this issue has nothing to do with web service technology in particular.

25. If we compare the web service approach of SODI with the approach of StatLine 4 the following difference comes about:

- The web services in StatLine 4 support only external information requests either by means of a looser key word search algorithm or by using *internal metadata structures imposed by StatLine*.
- The web services developed in the SODI pilot must support information requests of which *the metadata structures are imposed by Eurostat*.

26. In the StatLine 4 web services there is no need to match metadata structures between the client and the server other than a simple text matching on key words entered by clients. However, in the SODI pilot a simple key word matching system is not sufficient for matching the complex metadata structures of Eurostat with the complex metadata structures in StatLine.

27. In a short time period it is not feasible to match all the complex metadata structures of Eurostat with corresponding metadata structures available in StatLine. As a consequence this match can be automated soon only for dedicated and fine tuned subsections of data and metadata in both organisations. This leads to the following difference:

- The web services in StatLine 4 are generic enough to support *all the output information* that Statistics Netherlands can offer on all possible subjects.
- Web services developed in the SODI pilot support only *very small subsets of the output information* that Statistics Netherlands can potentially offer, serving only very few subjects. Each subset of output information has its own dedicated and fine tuned web service.

VI. CONCLUSION

28. In this paper we described the automated access to statistical data from a number of perspectives. We described the design principles behind the new web services being added to StatLine4, the main output database of Statistics Netherlands. These services were developed with ease and simplicity in mind and offer end-user applications mechanisms to keep informed of updates at a certain granularity level. They enable

external applications to integrate the results of the StatLine search engine in their system. And finally, these services make it possible to retrieve any subset from any cube from the database with a single request.

29. We compared the approach taken in StatLine4 with the approach in the SODI pilot of Eurostat. The SODI pilot uses the standardized SDMX-ML format for communication. We feel the StatLine approach is more aimed at easy automated usage of online statistics where the SODI approach is aimed at complete and exact communication between statistical agencies. Our general conclusion is that opening up statistical databases using web services is a valuable approach for automated access to statistical facts via the web. We think there is a need for general access services such as the approach taken by StatLine4, as well as for more dedicated services for exact data interchange between NSIs such as the approach taken by SODI.

30. Ideally the ideal statistical output database will support both kinds of web service approaches. They should provide general and easy access to statistics for customers that don't have their own metadata systems as well as strict and dedicated access methods for partners that do have their own metadata models and require data to be provided in their own metadata terminology, such as Eurostat. We note that web services by itself do not solve the metadata matching problem between NSIs. Client and provider of a web-service still need to know the metadata they are communicating about. In fact, a generic solution to this metadata matching problem remains the most burning issue to tackle for efficient information exchange between statistical institutes and their customers. There may be multiple solutions to this problem. Standardization of metadata is only one way to go. Another promising direction to explore is to add more intelligence to our statistical output databases, such as Semantic WebServices. Experiments in this direction still have to be started.

References

[1] E. van Bracht, *CRISTAL, a Model for the Description of Statistics*, Seminar on the Exchange of Technology and Know-How and New Techniques and Technologies for Statistics (ETK-NTTS), Crete June 2001; http://webfarm.jrc.cec.eu.int/ETK-NTTS/Papers/final-papers/73.pdf
[2] E. van Bracht, *CRISTAL, a Model for Data and Metadata*, Working Paper No. 29, Work Session on Statistical Metadata (METIS), Geneva February 2004; http://www.unece.org/stats/documents/2004/02/metis/wp.29.e.pdf
[3] E. de Jonge, *New ways of disclosing StatLine, a large statistical database*, International Marketing and Statistical Output Database Conference, Annapolis Sep 2002.
[4] J.A. Reedijk, E. de Jonge, O. ten Bosch, *Handling Time Dependence of MetaData and Data in StatLine4*, International Marketing and Statistical Output Database Conference, Oslo Sep. 2004.
[5] O. ten Bosch, *New Developments in Information Processing at Statistics Netherlands*, Meeting on the Management of Statistical Information Technology (MSIS), Geneva February 2003; http://www.unece.org/stats/documents/ces/ac.71/2003/7.e.pdf
[6] Statistical Data and Metadata Exchange (SDMX) Initiative; http://www.sdmx.org
[7] Article on the Netherlands from Wikipedia; http://en.wikipedia.org/wiki/Netherlands
[8] Website Agriholland; http://www.agriholland.nl/cijfers/cijf_statline_akker.html
[9] Website "natuurPUNTnl"; http://www.natuur.nl/article1629.html
[10] The Apache Lucene project; http://lucene.apache.org
[11] G. Pongas, *SDMX Open Data Interchange (SODI)*, Meeting on the Management of Statistical Information Technology (MSIS), Bratislava Slovakia April 2005; http://www.unece.org/stats/documents/2005.04.msis.htm

-----