Distr.
GENERAL

Working Paper No.21
28 April 2009

ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION STATISTICAL OFFICE OF THE EUROPEAN COMMUNITIES (EUROSTAT)

ORGANISATION FOR ECONOMIC COOPERATION AND DEVELOPMENT (OECD) STATISTICS DIRECTORATE

Meeting on the Management of Statistical Information Systems (MSIS 2009)
(Oslo, Norway, 18-20 May 2009)

Topic (iii): Architecture

A DISTRIBUTED ARCHITECTURE FOR STATISTICAL DATA EDITING, PROCESSING AND DISSEMINATION

Invited Paper

Prepared by Georges Pongas and Adam Wroński, Eurostat

Abstract

Nowadays the growing cooperation and exchange of information among statistics producing or disseminating organisations, the new statistical requirements and the pressure to diminish the burden of all the actors, forces the promotion of mutualisation (federation) of technical, human and financial resources.

The purpose of this paper is to present the basic technical and architectural requirements, and indirectly organisational implications, stemming from a distribution of data and statistical software in the various informatics environments of the statistical agencies network.

The proposed architecture is based on principles such as transparent distribution of data and metadata, web services orientations and minimal core functionalities required.

The proposal tries to present the possibility of various organisational alternatives leading to various degrees of inter-agency collaboration.

I. INTRODUCTION

1. This document attempts to present the basic components and ideas of a distributed statistical system that could be put in place for processing and dissemination of statistical data within the statistical agencies' network.

2. Concentrating on technical issues, what would be the features of a distributed statistical system that could be deployed in the statistical agencies' network to promote increased collaboration among all actors involved, decrease the production burden, provide better quality of statistical data to the end users and all this in a more timely fashion? The following discussion is based on our experience with the EU system of statistical agencies (national statistical institutes, other national agencies and Eurostat), the statistical data production procedures and the way Statistical Agencies (SA) communicate between them.

3. A realistic answer to this question should take under consideration the following considerations:
- A distributed system should integrate (evolve from) existing systems.
- It should provide flexibility of collaboration permitting several distribution levels.
- It should allow the SAs' participation in the network without enforcing either a unique system or a unique approach.

4. The implementation of a distributed collaborative system should result in:
- Better quality of the statistical data.
- More efficient production and dissemination processes.
- Improvements in the provision of data to the public in a more timely fashion.
- Allowing the SA to exchange data and metadata between themselves more efficiently

5. The necessity for varying levels of collaboration between SAs introduces different, alternative data flows based on the desired collaboration pattern. According to the level of collaboration, we propose here four levels (or schemas) of collaboration:

(i) Exchange data using common data models such as SDMX (dispatch data collaboration).

Exchange of validated and/or transformed data using a common data model and a common reference data and metadata:

(ii) Pull variant collaboration;

(iii) Push variant collaboration;

(iv) Work as part of a common system (advanced collaboration) by maintaining a local software instance in the interested SA or by accessing to another data processing node.

6. In order to establish a distributed environment, a common metadata and *logical data model*[1] is necessary for processing, storing and disseminating the distributed data. The management of a distributed metadata repository containing information for the data and metadata of the entire system is the cornerstone of the distributed system.

II. ULTIMATE DISTRIBUTED SYSTEM

7. The functional requirements and the existence of modern technology capabilities increase expectations for any new statistical data processing system. The attempt to define such requirements is presented below. The ultimate distributed system needs to match these requirements and distributed systems incorporating only some of the requirements are also plausible.

- **Fast, on-line data access for querying and loading purposes.**

8. The centralized systems have to provide:

(a) direct access to the local system by simplifying (or entirely removing) the existing off-line (outside the system) iterative process (e.g., transmission, editing, report transmission, new transmission, etc.)

[1] Definition based on Wikipedia: http://en.wikipedia.org/wiki/Logical_data_model

(b) locally processed validated data, ready to be loaded to the other systems;

(c) faster access to the final processed data.

The goal is that statistical data, described by a common logical model, remain close to the data source and everybody has easy and fast access to them.

- **Distributed process automation**

9. Because the data often contain errors, the loading of data to the existing systems is an iterative process. During each iteration the data administrators have to perform certain steps in order to edit and correct the data. In a distributed system, this task has to be executed near the data source, in an automated way, according to common rules for all SAs, to reduce the required time and human effort.

- **Quality of the overall data**

10. The data are frequently heterogeneous because of the diverse origin and use of different systems (logical models). The data are in general processed (transformed, edited and aggregated) in a Coordinating Statistical Agency (CSA) and eventually loaded to a reference system. A distributed system should allow all SAs do the processing locally, under the constraint that the processing is done according to the common standards. The common standard existence is necessary to guarantee the ability of all parties to share the processed distributed data. On the other hand, the rules must be adaptable locally in order to be processed according to the local requirements.

- **Common logical data model**

11. A common logical data model is mandatory for the operation of a distributed system because it enables it to specify the information provided at individual SAs in a common way and allows exchange of data with the SAs. The SAs should provide statistical data according to the common logical data model. The distributed system should exploit the existing logical data models, providing the mechanisms and/or interfaces for mapping of the data to a common logical model.

- **Interoperability at SA level**

12. In a distributed system, each participating SA provides access (view) to a subset of their data according to the particular SA data access rules. The data accessed must be according to the common logical data model. To achieve this, either the local system follows a logical data model or there is a system component (Distributed Agent provided by SA or CSA) translating the local logical data model to the common one. The Distributed Agent is a part of the overall distributed system, strongly connected to the local SA systems. The Distributed Agent must fulfil all the security and administration constraints. To make such systems interoperable the following services must exist:

(a) A Distributed Agent (DA) can retrieve data from another Distributed Agent. As a result distributed query processing will increase the total throughput of the entire system.
(b) Distributed Agents (DAs) can exchange messages.

- **Security**

13. System and data security is one of the decisive factors in the system success. The interconnection among SAs should not pose any risk for the security of any party involved. The communication between the DAs and other parts of the distributed system and with the local systems must satisfy all the particular access and security rules for each participating SA

- **Distributed site autonomy**

14. The SAs should be able to maintain and develop their systems autonomously. This practically implies that the existing systems would be retained without major modifications. Logically a DA can be seen as separate and external part of the existing system that gives access to authorized parts of the data of the local system using the common logical model. The installation and / or operation of the DAs in SAs must not affect the operation of the existing statistical data management systems.

III. THE DISTRIBUTED SYSTEM COMPONENTS

15. It is proposed that the distributed system consists of the following major types of components and corresponding interoperability connections:

- **Central Agents** hosted at Coordinating Agency(ies) (CSA).
- **DAs** installed and operating at non-coordinating SA.
- **SA local systems** that provide the statistical data to the DAs.
- **Web clients** connected to SAs or CSA of the distributed system. The functionality of these clients can vary from dissemination (extraction oriented) to production oriented clients (for example data editing oriented clients, submitting to CSA data for editing).
- **Firewalls** that guarantee the security of the entire system by controlling the information flow within distributed system (among SAs and CSA).

Central Agent

16. A Central Agent (CA) is a software system connected with the DAs. The CA should allow distributed processing functionalities:

- The CA requests that the DAs make available a subset (defined by a set of parameters) of local data.
- The CA provides to all users the coherent logical structure of the accessible data available through the DAs. A Central Agent also provides and controls all the metadata needed for all the operations within the Central and DAs. For this purpose it includes the SAs metadata residing in the DAs.

Distributed Agents (DAs)

17. In the distributed system, the DAs (or its local equivalent in functionality) have to exist at each SA. Each Distributed Agent must be able to recognise and perform agreed processing on the statistical data that the corresponding SA provides to other SAs. The reason for the Distributed Agent is to provide uniform access (using the common logical data model) to the statistical data produced by the corresponding SA. Additionally, the DAs can process the data (transformation, derivation, editing, etc.). The common logical model allows common data processing schemes that would in turn permit more efficient (distributed) data processing. E.g., one can use editing rules from another SA or one can edit data of other SA using its own editing rules.

18. It is important to emphasise that the DAs are the key to the success of the entire distributed system and they have to be developed, installed and operated according to the cooperation principles that will be established among all SAs. Each SA will be a user of its Distributed Agent. To make the development easier DAs have to rely on, as much as possible, the technology the corresponding SAs is familiar with.

19. DAs must not affect the operation of the local systems at the SAs. Each Distributed Agent should be installed at a possibly separate host, permanently connected to the corresponding SA local system. It will also be connected with the other parts of the distributed system, the Central Agent and other DAs via the network.

20. The DAs will allow access to and external processing of the SA statistical data accessible by the rest of the distributed system. Each Distributed Agent has to provide statistical data at least to a Central Agent.

21. There is no one-fits-all Distributed Agent. Different levels of cooperation among SAs require different types of DAs. Each type of Agent would perform different processes. In the case where high cooperation between Agencies is available, DAs will provide statistical processing functionality (editing, aggregation). In other cases the DAs will provide solely online access to statistical data. The major requirement of the distributed system is to produce maximum benefit from the participation for each of the SAs.

Existing local systems

22. Each SA normally has its own data processing system based on a particular technology. The different hardware platforms, database systems, data models, applications and software standards that are used among SAs render difficult any direct connection between them. Nevertheless we are constantly obliged to transmit statistical data from one SA to another because one SA produces data that is used by other SA. Consequently, a DA must be connected with the local system systems of the SA, in order to transmit or process data.

Firewalls

23. For security reasons, the connections between the DAs, which are parts of the distributed system, and the SA local systems must be totally controlled by firewalls to prohibit unauthorized data and service access. To further minimise security risk the DAs should allow only a limited set of interventions, such as the description of accessible data and user access rights. The DA management would be done by the respective SA. Firewalls totally deny any direct access to the SAs local systems from any point of the entire distributed system.

Web clients

24. The Web clients can connect to the distributed system and use its services. End users will use Web clients to access the data and process it. A client of a given type gives access to a certain number of services and SAs of the distributed system. There can be many types of specialised clients such as:

(a) Data dissemination client able to perform queries on the statistical data through the DA. In these cases the DA, depending on its configuration, coordinates the execution of the distributed query (connects to various DAs of SAs as functionality can be distributed as the data can be).
(b) Data editing client able to visualise editing rules, validation procedures and execution results and submit data for editing. In that case all the DAs can be distinct: the DA executing editing, the DA providing metadata, the DA providing the data.
(c) Computation client providing aggregation, seasonal adjustment and other statistical services. In that case several distinct DAs can be implicated as well.
(d) Various content maintenance oriented clients updating structural and process metadata, e.g., dataset definition, security rules, variables, etc.

IV. DISTRIBUTED AGENTS ARCHITECTURE

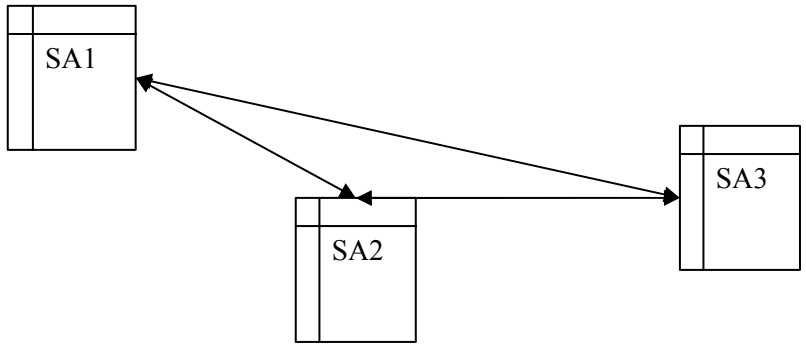
25. Different levels of cooperation among SAs can be entertained. The level of cooperation defines the processes that are executed by the DAs. One can theoretically imagine the full level of integration of SAs. However, today, the situation is far from that idea. To prompt a gradual increase in collaboration, a less powerful DA can be designed in order to exploit the SA's infrastructure and know how.

26. Below four different types of DA architectures with different level of integration and collaboration are described.

Type 1 DA: data dispatch

27. This type of DA represents approximately the current situation. The data is transmitted among SAs according to specified agreements. Usually data transmissions follow predefined data and/or metadata formats such as SDMX. To transmit, commercial software is used or specially designed software is installed in the SAs. An example is communication within ESS using eDamis. This type of collaboration does not require a common logical data model.

Figure 1. Data dispatch outline



Type 2 DA: pull data collaboration

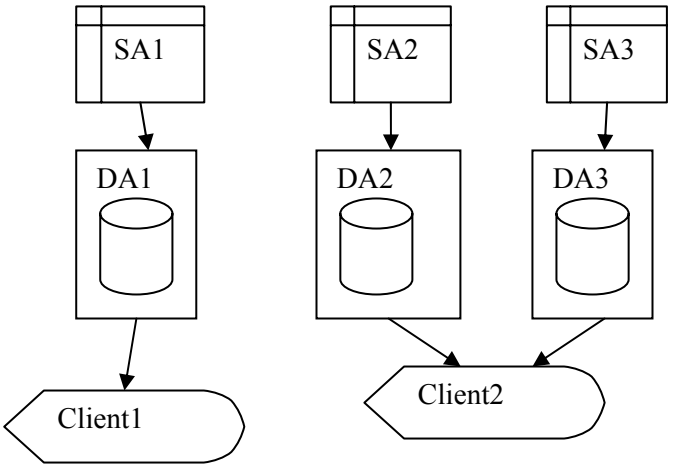
28. In a pull data collaboration, the DA provides access to the exported data via a DA database or a set of files put in predefined directories accessible from the outside of the SA,. Other SAs can retrieve data by connecting and querying directly the database or the file set. The DA of this type operates as follows:

- SA staff uploads/copies periodically to the DA database the files that contain statistical data to be accessed.
- Querying software accesses the DA database that the SA provides.

The query results may be compressed and decompressed during the transmission for network performance reasons. For security reasons, data access authorization and authentication is performed by the DA.

29. This type of DA is installed at the SAs who agree to create a DA database or set of files to store the data they provide to other SAs. The owner SA needs to manage a database or set of files to satisfy the needs of the other SA. An example of such an agent is implemented in the Eurostat dissemination site under the name of Bulk Download (files of predefined format ready for download). Another possibility is through the use of a set of Web Services which are data extraction oriented. A certain level of cooperation among SAs is required for the DA content administration and for the specification of the shape of the accessible statistical data. This type of collaboration does not require common a logical data model.

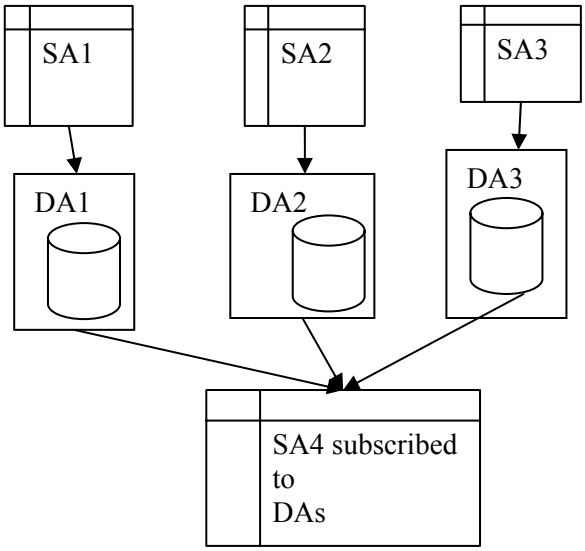
Figure 2. Pull data collaboration outline.



Type 3 DA: push data collaboration

30. In a push data collaboration DA provides to other SAs validated and processed statistical data. The statistical data are stored in a data and metadata database (warehouse). SAs can connect and query the warehouses. As in the second type, the third type of DA can be installed by SAs who agree to maintain a database with clean data. In addition there must be a functionality of automated data pushing functionality. Interested requesters (other SAs) subscribe to the data pushing service by formulating and storing their queries in the "push oriented" DA. If any data change occurs, the Distributed agent executes the relevant queries and pushes the data to the appropriate recipients. This type can be seen as a counterpart of a subscription. Again, as in the second type a decent level of cooperation among SAs is required for the DA content administration and for the specification of the shape of the accessible statistical data. This type of collaboration does not require a common logical data model. However, in this approach the common logical data model would simplify greatly the collaboration.

Figure 3. Push data collaboration outline



Type 4 DA: advanced collaboration

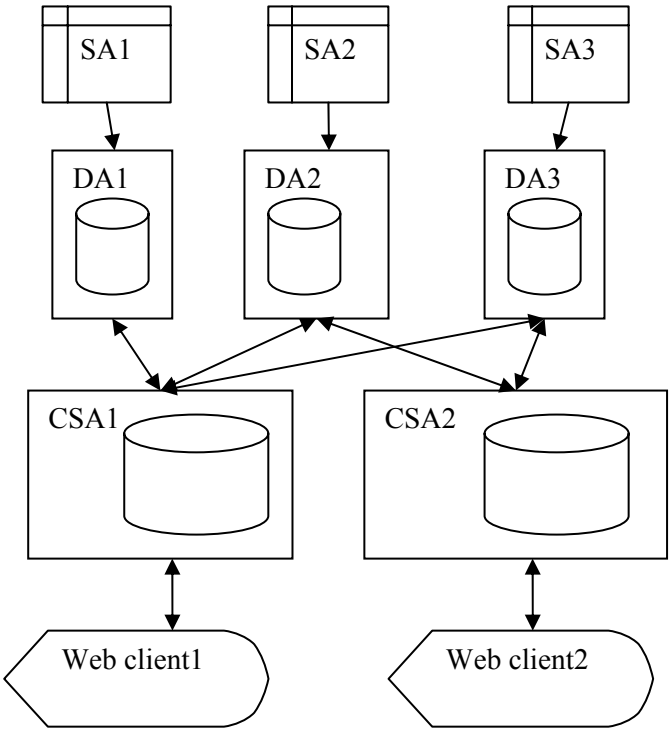
31. This type of collaboration requires an advanced level of cooperation among participating SAs. It requires establishing a CSA with an environment providing a large variety of statistical services permitting data manipulation (e.g., aggregation, editing, comparison, etc.). The description of such services is available to all internal and external (e.g., SAs) users accessing the CSA content. The CSA renders accessible the content and the description of data residing in the collaborating SAs through a set of queries or Web Services. As a result an extensive metadata describing statistical services must be present on top of the metadata describing the accesible data.

32. The DAs corresponding to this level of collaboration must have the functionality of the type 3 DAs based on a common logical data model.
This would allow distribution, not only of data, but also of functionality. As a result, an SA can submit data to another CSA to perform a data processing step (editing, seasonal adjustment, imputation, etc.) on the data belonging to any SA or on a union of data from many SAs. Again the results can be provided to any of the collaborating SAs.

33. A typical example concerns the statistical editing. Intelligent editing operations request not only current data but also historical and perhaps data belonging to other agencies. A CSA may contain not only the necessary data but also the editing rules and the editing software to perform the edits. In that case an appropriate Web client would be able submit to the editing software, data extracted from other SAs and receive back an error

report. This kind of "submission for processing" presents an advantage by allowing maintenance and execution of complex software and metadata in one place. A similar publicly available example is the Google Translate software. Instead of installing translation software in the PC, Google permits to submit your text for translation. Needless to say this kind of architecture is equally applicable to statistical functionality such as seasonal adjustment, outlier detection, etc. Moreover this architecture does force data replication between SAs. SAs data can be unified logically through the CSAs in a seamless way. There can be many CSAs each with specialised functionality and data content. There can be a CSA specialising in data editing or a CSA uniting specific data of a cluster of SAs.

Figure 4. Advanced collaboration outline



V. CONCLUSION

34. This paper provides a discussion on various degrees of integration of metadata, data and functionality among the statistical agencies. The approach chosen permits gradual integration and maximises reusability of the currently existing software components and databases. Key elements of success are the extensive use of metadata driven applications followed by harmonised common logical data models through the extensive use of metadata mapping the physical data onto a logical model.

35. To summarizes the basic functionalities of the four agent types:

Table. Summary of DAs' functionality

Functionality implemented	DA type			
	1	2	3	4
Dispatch data	Y	Y	Y	Y
Access for Download	N	Y	Y	Y
Web Services	N	Y	Y	Y
Customised requests execution	N	N	Y	Y
Analytical tools	N	N	N	Y
Logical data views (inter Agencies)	N	N	N	Y
Computational services (upload data, compute, download)	N	N	N	Y