STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS **WORKING PAPER No. 16**

Work Session on Statistical Output for Dissemination to Information Media
(Voorburg, Netherlands, 25-27 September 2000)

ENGLISH ONLY

Sharing Statistical Information over the Internet

Paper submitted by United Kingdom[1]

Introduction

The World Wide Web has already had a profound impact on the way NSIs publish data. By the end of the Fifth Framework Programme, in 2003, Europe will truly be working in a global market and providers of official statistics will have to work in this context. This has a number of implications:

- Users, governmental organisations, businesses, education and private citizens will have access to a wide range of statistical datasets from a diverse, and largely unregulated, group of providers.
- As access to statistics becomes easier, users will wish to compare the statistical information produced from these datasets and to perform comparative analyses on datasets that have not been harmonised by an authoritative body.
- The demand for statistics, the ease of access, and the rate of change in economic and social conditions will require a dynamic sharing of expertise by domain experts and statisticians in order to address the social, economic and policy issues which will arise.

This paper describes MISSION (Multi-agent Integration of Shared Statistical Information over the [inter] Net), an R&D project funded under the EU Fifth Framework Research Programme. The vision of MISSION is a number of independent organisations publishing their data within a framework that makes comparisons and harmonisation possible. The project aims to provide a modular system of software which will enable providers of official statistics to publish their data in a unified, and unifying, framework, and will allow consumers of statistics to access these data in an informed manner with minimum effort.

We intend to utilise advances in statistical techniques for data harmonisation, agent technology, the availability of standards for exchanging metadata and the power of Internet information retrieval tools, to build a modular software suite. This software will:

- Allow suppliers of statistics to subscribe to an integrated network of datastores via an interface to their existing data while retaining control over all aspects of access to their data. This includes their level of involvement; the data they supply; the users who can access it; and the level of resources to commit.
- Allow users to make declarative requests, with a minimum of understanding of statistics, or the domain area, and still retrieve meaningful results from our internal routines or through an interface with external statistical packages.
- Give the user a range of options for automatic harmonisation of statistical data, with clear indication on the interpretation of the results.
- Provide audit trails of data manipulation and analysis, so that methods can be retained, re-used and published.

[1] Paper prepared by Ms. J.M. Lamb, Centre for Educational Sociology, University of Edinburgh, St Johns Land, Holyrood Road, Edinburgh EH8 8AQ, Scotland, UK, Tel: 44-131-651-6276

- Maintain libraries of metadata that can be made available to other users.
- Provide a flexible architecture that allows third parties to act as Independent Metadata Providers, thus encouraging the free exchange of knowledge.
- Allow users to build up individual profiles, accessing data and methods most relevant to their needs.
- Offer a number of independent, interoperable systems that can run on different hardware platforms and access heterogeneous data storage systems.

Background

Research in the area of official statistics has developed significantly in Europe during the last five years [1], [3], [4], [5], [8]. The area reflects the needs of providers to address a number of significant questions concerning the quality of the data they provide, harmonisation across Europe, and the use of technology to improve the service they give. The Internet has proved to be a major factor in the way that National Statistics Institutes (NSIs) operate. Currently, most NSIs are developing, or adopting statistical databases such as Statline [7] of the Netherlands, Infoline [2] of Portugal and Statbase [6] of the United Kingdom. However, most databases are primarily aimed at national consumers, or are oriented to describing only the products of the NSI. The current proposal addresses a wider question, aimed at helping users to operate in a global context. That question is:

> *"How can we provide an environment where users, novice and expert alike, can publish, on the Web and elsewhere, statistics of a demonstrably high quality and transparency to permit their use in a global context?"*

To answer this, we must consider a number of issues. Firstly, NSIs are already using the Web and investing heavily in supplying data in this arena. This investment is much greater than the money that can be invested in a research project, so we must investigate what added value can be supplied. Secondly, the NSI is now very remote logically and physically from the user, who therefore needs much more automatic support from the systems he uses. It is worth remembering that different member states have different degrees of centralisation of data collection and dissemination, and that data is increasingly available from national government departments and agencies, as well as from a host of governmental and non-governmental agencies at all levels from local to international. Thirdly, metadata is required to aid this support, but it is notoriously difficult to capture. Also, users need help in the interpretation of the information, both context and statistics. Lastly, the Information Technology context in which we find ourselves by the end of the project will be very different from the current context; we will need to cope with change.

A number of projects of the Fourth Framework addressed some of these issues. However, these systems were developed in a research environment outside of the NSIs normal procedures. The resulting systems were to some extent closed, and required a significant investment, or re-organisation, on the part of the providers in order to adopt or develop such a system.

The innovation of MISSION is twofold. First, it adopts emerging technologies and combines them together in an imaginative way, and second, it proposes an innovative organisational structure, which will overcome some of the difficulties NSIs have in integrating new technologies into their current practice.

Four innovative strands combine in the MISSION software: agent technology, XML based descriptions of metadata, harmonisation and data merging techniques, and information self-organisation.

Agents are used to provide an intelligent interface for the user. The user supplies the agent with a rough description of the information that he/she is interested in. Using XML descriptions of statistical data, the agent tries to locate the relevant data that satisfy the user's interests. Once this is located, a query can be formed with (or without) interaction with the user. These query agents can interact with the interface agent for a query answering mechanism to supply the answer. Query agents take a declarative approach to statistical queries. A user can state his needs in terms of a *goal*, that is a statistical (macro) table, plus its related metadata (footnotes). This goal can be described in terms of the rows and columns of the tables and the expected values in the cells. The agent turns the query into a series of operators which, working on the data and metadata will result in the required table. This is called the query plan. A query plan does not always uniquely map onto a goal; a number of query plans might result in the same goal. However, some of these will be more efficient than others, so query optimiser agents can be invoked. Agents can also identify whether a goal table has been previously computed, and if so retrieve it.

Mediation agents provide the mechanism for harmonising metadata without the need for a predefined global ontology (a set of shared concepts to which all data definitions can be mapped). Instead of a universally shared ontology, each user will carry his own personalised ontology that can be used for the mapping. Alternatively, third parties can act as repositories for shared ontologies in a particular domain or country. When a new data provider is accessed, there is a need to negotiate a modified shared ontology - the combination of the old shared ontology with the new provider's ontology. This means modifying the shared ontology by taking the product of the old and new classification schemes, a task carried out by the Negotiation agent. These agents provide the possibility of less directive and less centralised solutions, giving the user the ability to utilise dynamic classification systems tailored to the question in hand.

XML based information systems offer a number of opportunities for innovation. There are a number of reasons for developing XML based systems. First, an XML document is held in text format and therefore is easily transferred between platforms. Second, the well-defined structure (defined by the Document Type Definition) makes it possible to develop programs which understand the semantic content of the document. Third, this well defined structure also makes it is easy to generate from existing systems. Fourth, the rise in popularity of XML based data provides tools and incentives for data suppliers to make metadata available in this format. Among the uses we see are the interrogation by agents already referred to, indexing and searching of XML based documents, the generation of XML based results which are incorporated into the body of knowledge, and the ability to (re)structure libraries of accompanying information to improve efficiency of access.

This organisational innovation provides an infrastructure that gives different users with different needs access to tailored comparative statistical data. Users have different needs, depending on their background and the time they can allocate to a request for information. These requests range from in-depth studies, for example by statisticians and domain experts, interpretation of tables by policy makers and government officials, and simple queries perhaps by journalists and members of the public. The structure is an open system that allows the sharing of expertise and knowledge, and also minimises the effort required by the data provider.

The Architecture

The architecture of the system comprises four basic logical, or conceptual, units or building blocks, which can be deployed in different scenarios. The components are:

- The Client, which provides the interface to the system.
- The Library, which is a repository for statistical metadata.
- The Compute server, which is a statistical processing engine and stores no information of its own.
- The Data server, which is the unit that gives access to the data.

The Client

The Client component is a Web based user interface that connects a user to all sites participating in the architecture. It can be a *thin* client running, for example, a standard web browser, or a *fat* client that would provide more features. It obtains a request from the user, and sends an agent to search for a Library that can satisfy the request.

The Library

The Library software supports a repository for statistical metadata. Different Libraries can communicate with each other. A Library holds three different kinds of metadata. The most basic type of metadata is *access* metadata, which is the physical, and logical information required to access statistical data. The second kind is *methodological* metadata, which is the information required to process that data in order to satisfy requests for statistical analysis. The third kind of metadata is *contextual* metadata, which supplies background information and explanatory notes for the user. This kind of information includes, for example, the purpose of a survey, or an explanation of a break in series for a time series. This information can be attached as footnotes to a query result. The first two types of metadata are machine understandable. The last is machine readable and human understandable.

When a Library receives a request, it decomposes it, and, if necessary, it can send to other Libraries in the system for any metadata it requires. Once it has built up an operation, it submits it to a Compute server. On receiving the reply to the request, it returns the answer to the Client.

The Compute server

The Compute server is a statistical processing engine, which stores no information of its own. Based on the query it receives, it obtains the necessary data from various data servers, performs the request, and returns the result to the

Library that made the request. It may also make a request to third party statistical packages. A primary objective of the compute server unit is to integrate a distributed declarative querying facility and a distributed statistical aggregation system, using distributed database and web technology. The compute server architecture is designed to incorporate intelligent agent techniques. Query agents will facilitate interaction with library server units concerning positional and other operational metadata; query agents will also facilitate interaction with data server units concerning macrodata; mediation agents will enable the user-specified merger of heterogeneous macrodata and accompanying metadata. The principal task of the compute server unit is to receive and interpret queries from library units and to return macrodata and metadata results along with action plan metadata useful for future query optimisation.

The Data server

The Data server is the unit which gives access to the data. It holds the data itself, management tools for registering and maintaining the system and a gateway module. The gateways hold the minimum amount of metadata necessary for the safe use of the data. This includes registration details to allow the Provider to control access to the data and information about the physical structure of the datastore. Other metadata is made available to be uploaded to Libraries that request it.

Agents

The units described above form the static components of the system. These components operate using generic computations and access or store various data repositories. However, the user orientation of the system is carried out by *agents*. Agents perform intermediate processing and navigate the Internet to access the appropriate building blocks of the system. Once these are located and accessed, agents are responsible to invoke the appropriate computations on the engines or retrieve the appropriate data and metadata according to the user request/goal.

Actors

We have described the four building blocks. They can be deployed in different scenarios, which we will describe. First, however, we must introduce the Actors in the system. We identify three at this stage:

- The User: the user is anybody with access to a Client who wishes to make a query on statistical data. Users will have different levels of statistical expertise and domain knowledge, and the Client will cater for their differing needs.
- The Statistical Data Provider (SDP): the SDP is a site that has data which it would like to offer for statistical analysis.
- The Third Party Provider: This is a site which is neither user nor SDP, and which houses one or more of the basic units.

The different scenarios depend on three things: the policy of the SDP; the size of the user's needs; and whether or not a third party is involved.

Scenarios

In the first scenario, we assume that the Provider gives a minimum of support to the users. In this case, only the Data Server resides at the Provider Site, and everything else is with the Client. This is illustrated in Figure 1.

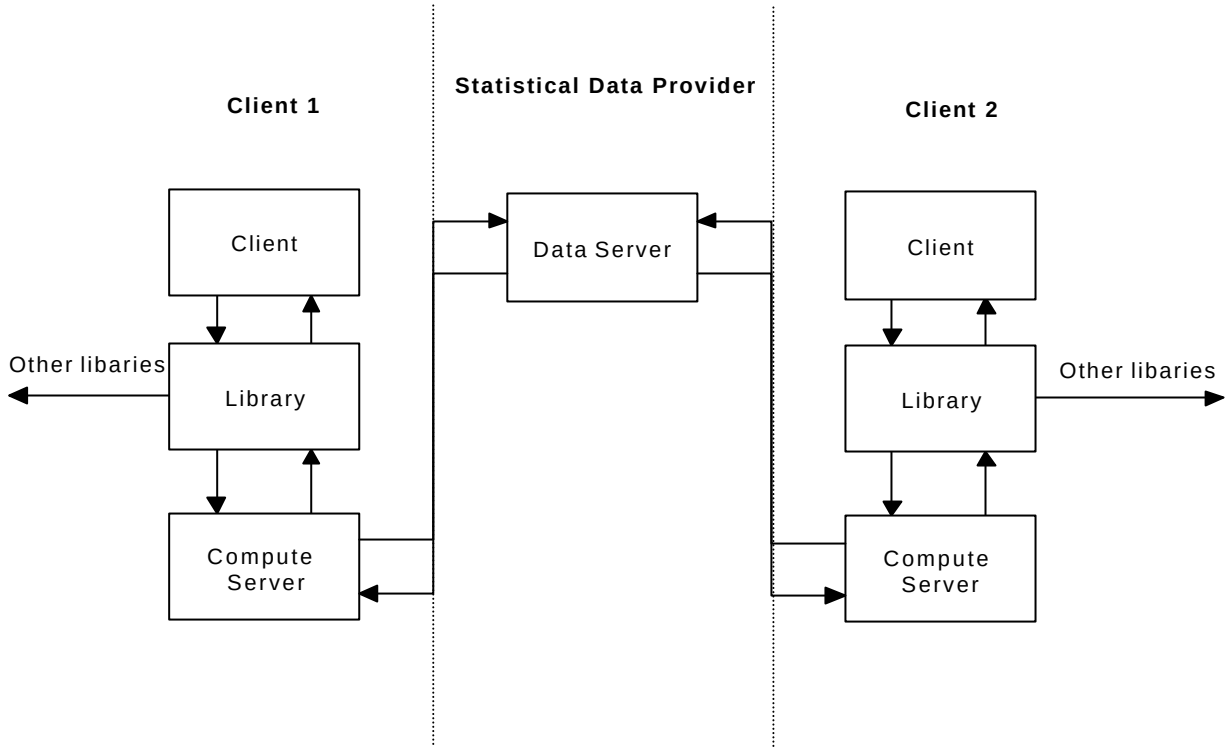

Figure 1: Parsimonious Data provider, Independent Clients

A second scenario is the case where the Statistical Data Provider provides all facilities for the users. The only unit the user needs is the Client, as shown in Figure 2.

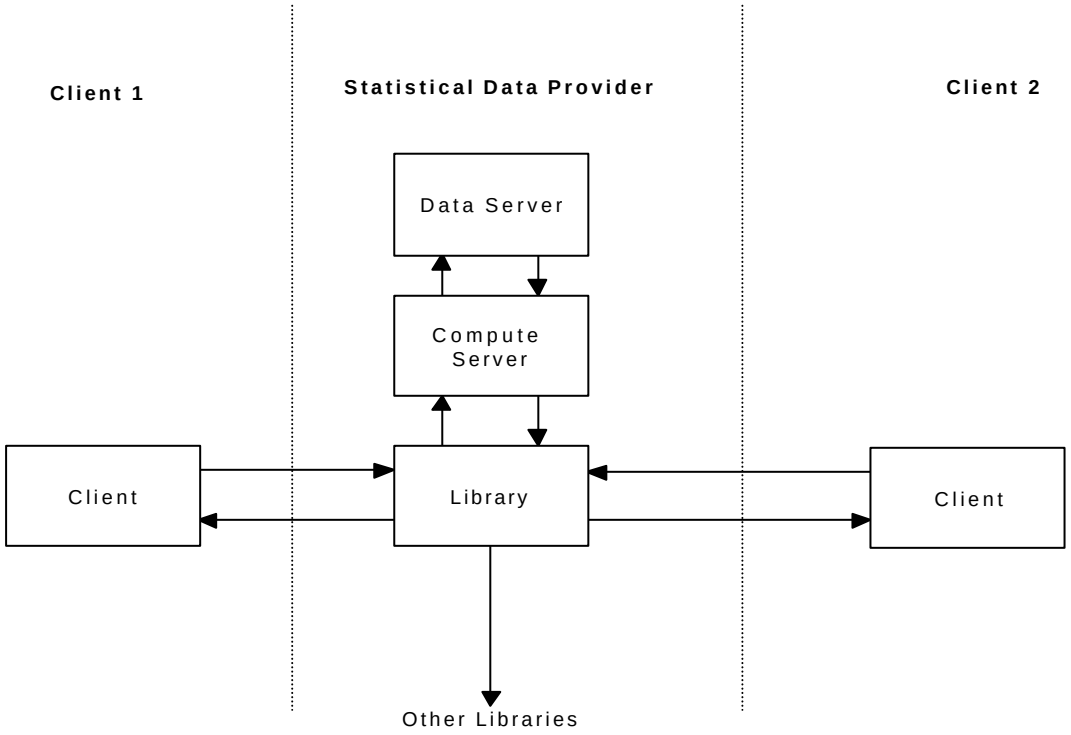

Figure 2: Generous Data Provider, Independent Clients

In the third scenario, a Third Party Provider, which is in effect a Statistical Metadata provider (SMP), hosts the intermediate modules, i.e. the Library and the Compute server. This is illustrated in Figure 3.

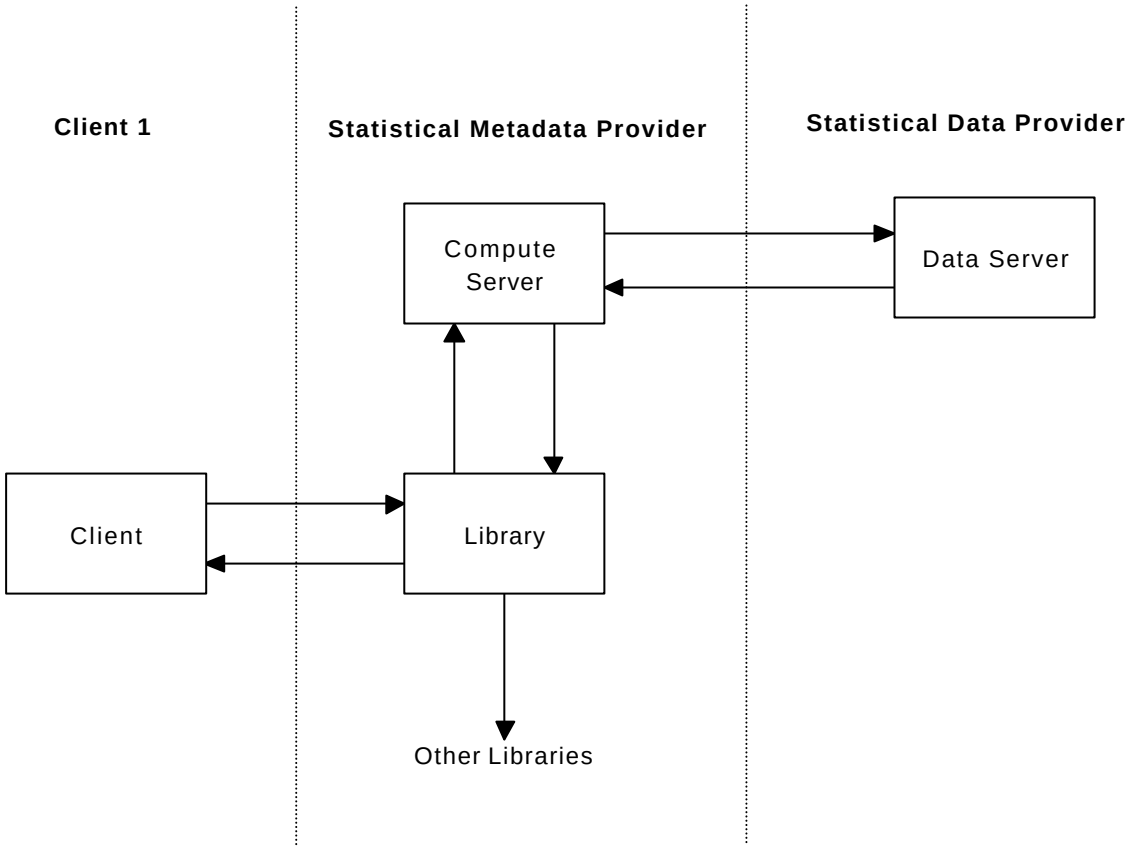

Figure 3: Third party Statistical Metadata Provider

These three scenarios show the basic models that can be replicated for different users and different data providers in numerous ways. Provided the actors subscribe to the system, it provides a flexible way of integrating disparate statistical data sources, and also provides an opportunity for developing specialised or local metadata repositories.

Implications of using such a system

The architecture offers a variety of configuration models, so that private and public libraries of expertise can be built up. A private library may be local to a department and used to hold the processes needed for their day-to-day operation, while the public library would hold the processes needed to map the collected data into a format or aggregation needed by other actors in the business. The library will have registration formalities which ensure that only valid users have access to sensitive or controversial information.

The ethical and commercial issue of confidentiality of statistical data is well known. National Statistics Institutes have many years' experience of tackling this problem, and research is continuing in better ways of handling statistical disclosure. For this reason, the philosophy is that control of the data remains firmly in the hands of the data providers. Data sets are registered as public or are password protected. Every request for access to data is checked against a registration database before being carried out, and will be refused if the validity checks are not satisfied. Agents facilitate the registration process, but a provider has the option of externally validating users if he so wishes.

It is less obvious, but equally important, that the third party libraries are maintained to a quality standard. This is of equal importance in a system that is entirely internal to an organisation as the quality of the management information available is controlled by the quality of the metadata of the poorest library. It is important that the standards of the methods and interpretations supplied in libraries are subject to scrutiny. The addition of a method should go through the same screening used by the organisation for any other business indicator that they use. By

supplying quality checking mechanisms for the library, such as by validating the publisher and publishing his/her credentials along with the method, the system can help ensure the quality of published methods.

This verification of methods is very important. Statistical literacy among the general public, and even some professional bodies, is not high. There have been examples of misleading statistics being published in the press. With more and more critical decisions being made based upon the data extracted from online statistics, it is vital that the published figure be well defined and a true indicator of the current state.

Conclusions

Network resources are maturing rapidly. The rise of XML and the WWW provide many interesting and challenging problems to distributors of statistical data. In order to make the correct decisions and maintain the competitive edge in a market, organisations need to have access to timely, accurate and understandable data. This paper has outlined one system in which existing data and metadata can be combined and delivered to the desktop of any authorised person. Further, it has shown that this can be achieved without having to replace the existing data infrastructure, or alter the working practices of the departments or units of the organisation.

The ability to seamlessly include information from suppliers, government and other third parties in the data space helps to set the business information in context and unifies the data query/analysis process. This will enable the decision-makers within an organisation to operate on the best quality data that is available to them.

Acknowledgements

MISSION (Multi-agent Integration of Shared Statistical Information over the [inter]Net) is an R&D project funded by the EU Fifth framework Research programme. It is managed by Eurostat, on behalf of the IST (Information Society Technologies Programme), as part of EPROS (the European Plan for Research in Official Statistics).

MISSION is further described at http://epros.ed.ac.uk/mission

References

[1] D'Angiolini, G., Paolucci, M. and Signore, M. (1998) *Developing Tools for Managing, Exploiting and Disseminating Metainformation: the ISTAT Experience* in International Seminar on New Techniques and Technologies for Statistics, Specialised sessions 191 - 198, Sorrento, 1998.

[2] Infoline, Instituto Nacional de Estatistica Portugal: http://www.ine.pt/en_index.asp

[3] Lamb, J.M., Hewer, A., Karali, I., Kurki-Suonio, M., Murtagh, F., Scotney, B., Smart, C. and Pagrach, K. (1998) *The ADDSIA Project: Issues and Achievements*. DOSIS paper circulated at the 3rd Seminar on New Techniques and Technologies in Statistics, Sorrento, 1998.

[4] McClean, S., Grossmann, W. and Froeschl, K.A. (1998) *Towards Metadata-Guided Distributed Statistical Data Processing Libraries* in International Seminar on New Techniques and Technologies for Statistics, 327 – 332, Sorrento, 1998.

[5] Papageorgiou, H., Vardaki, M. and Pentaris, F. (1998) *Recent Advances in Metadata* in International Seminar on New Techniques and Technologies for Statistics, Specialised sessions 191 - 198, Sorrento, 1998.

[6] Statbase, National Statistics Office, UK: http://www.statistics.gov.uk/statbase/mainmenu.asp

[7] Statline, Statistics Netherlands: http://www.cbs.nl/nl/statline/index.htm

[8] Van Nypelseer, P. (1998) *Rainbow final report*, DOSIS paper circulated at the 3rd Seminar on New Techniques and Technologies in Statistics, Sorrento, 1998.