Socio-Economic Database Online

Thomas TAMISIER, Fernand FELTZ, Stéphane RIEGEL

Cellule de Recherche, d'Etude et de Développement en Informatique
Centre de Recherche Public - Gabriel Lippmann
162a, av. de la Faiencerie, L-1511 Luxembourg, Grand-Duchy of Luxembourg
tamisier@crpgl.lu

Abstract: The SEDO project develops a platform combining fast access, user freedom, and coherence of the results for presenting complex socio-economic data. As an initial yield, this joint project involving two research actors of Luxembourg, the CEPS/INSTEAD for the content and the CRP-Gabriel Lippman for the technical realization, will deliver on the Net the results of longitudinal surveys about the life in Luxemburg in a generic tool comparable to those of the specialized community, characterized by flexibility and reutilization. Several search methods are available: a hierarchical browsing, a query engine with Boolean operators, and a top down navigation with minimal clicks for convenient and quick access to the main trends. Without the use of statistical tools or expertise in the domain the user can perform on the platform advanced statistical calculations. A new scheme for relational database has been designed to assure the portability of the application over different architectures. Last, the service manages the miscellaneous aspects of user identification, variable baskets, and the preparation of data for coherent analysis.

1 Networked access to socio-economic data

The use of the Net for querying socio-economic databases has lately drawn the attention of both experts and the public at large, with a number of online services provided by statistical organizations of various countries. However, the volume of the data as well as the technical specialization of the domain curbs the customization and exploitation of the information [Br98].

Not only the networked publication of statistics gives the possibility of instantaneous update and correction, but it is also, more than the traditional paper communication, suited for building relationship with the data provider and the data consumer. The consumer can deliver his feedback and formalize his requirements, and the provider has means to control the usage of the data, so that the figures are better and more properly used in the practice. Other advantages are to ease the retrieval of large data sets, to allow their customization, and to enable end-users to execute on the server statistical functions for processing the data [Su97].

However, facilitating access to the information lead to specific problems such as selecting in the volume of the information, the pertinent data and operations made available to the public, and protecting the confidentiality of the data in whole. Infrastructure has to follow with the development of the functionalities and the evolution of the demand. When performing useful statistical analysis becomes easy, and leads to an increased demand, supporting exhaustive calculations simultaneously performed by an undefined number of clients forms a real technical challenge.

2 Related techniques

The *metadata*, or data about data, are the specific information for describing the data considered as raw material for study or discussion. In the last decade, statistic frameworks have emerged from the development and normalization of metadata, so as to unify the collections of similar or correlated data [Me04]. As a result, various classifications, or views, of the raw data can be proposed according to the criteria of these metadata in such a way that static presentation of figure charts is no longer sufficient to fully render the information contained in the collection of statistics.

Based on the metadata, two kinds of strategies for exploiting the data are proposed. The hierarchical approach proceeds by the exploration of an arborescent structuration of the data. The structure is represented in a *Thesaurus*: a conceptual lexicon organising abstract keywords in a graph of semantic dependences without loops. Searching for the information consists in walking the graph until the leaves where the effective data are stored. The engine-supported approach implements multi-criteria search, and proposes to the user a direct access to the data using his own keywords.

Most services available on the Net implement both a hierarchical and a search-engine method. However, they often differ on the information area on which they are applied. For bibliographic search, engines are restricted to a collection of search criteria: period, localization, keywords, domains, author, etc… For access to statistical data, engine query leads to a list of survey variables with a brief description on the terms of the metadata, while the real results are kept in charts accessed by the Thesaurus.

3 The SEDO project

The CEPS/INSTEAD in Luxembourg is a public institute for statistic and socio-economic research dedicated to Population, Poverty, and Public Policy Studies. One of its current missions is a to maintain a voluminous database about the national population through the Socio-Economic Panel "Living in Luxembourg" (PSELL). The PSELL is a *longitudinal* survey and it collects *longitudinal* data, which means that the same questions are periodically checked for the same population. It is performed yearly since 1984.

The PSELL manages a huge quantity of socio-economic information pertaining to Luxembourg residents and suited to implicate them further in the public life. The survey takes notably into account the following domains: demography, education, post-natal child care, employment, right equality, resources, housing, equipment and consumption, and the opinions about the life quality and its evolution.

Thanks to its precision and its constant corrections, the socio-economic picture drawn by the combination of these miscellaneous themes is apt to significantly ease the communication between the population and the civil actors and to sustain the public choices made by the political authorities. In particular, the evolutions in the household situations and the opinions shed light on the effects as well as the efficiency of the administrative and political decisions.

However, the exhaustive and disparate information produced by the PSELL since the beginning is difficultly exploitable due to the complexity and the volume of the data. Nowadays, only researchers with a solid background in longitudinal analysis can effectively benefit from this work whereas it is relevant to all aspect of the cultural, social and professional activities outside the area of the statistician experts.

The purpose of the SEDO (Socio-Economic Database Online) project is to offer a generalized and convenient access to statistical information through the Net. To this effect, it targets the development of a platform for publishing and browsing socio-economic databases. To begin with, the project concentrates on a segment of the PSELL, in order to publish online the results of the PSELL 2 survey, which covers 7 years from 1995 to 2001.

4 Specifications

Besides the delivery of socio-economic research results to the public at large, the SEDO project is also motivated by the willingness to integrate the CEPS/INSTEAD into the network of national statistical institutes and to help the cooperation inside the scientific community. Because of this diversified orientation, the SEDO Web platform offers navigation according to different levels of expertise: general public is provided with fast, precise and ergonomics indicators whereas most advanced users will be able to query the database for their needs and apply several statistical primitives to the results.

Compared to other information tool of the community, the SEDO project targets a large flexibility, a universal portability, and the possibility to adapt to different databases. The first benefit of the system is to obtain customized presentations of statistics with guaranteed accuracy without neither the use of specialized statistical tools nor the necessity of an advanced expertise in the statistic techniques. For the sake of the reusability and the generality of the design, it has been preferred to build dependent platform, rather than developing features on top of existing tools such as NESSTAR. First, our approach gives room to the integration of powerful statistical tools and let the user perform online pertinent calculation according to the data. Second, full latitude is allowed to the programmer for the integration of the thesaurus and the customization of the search-engine.

Last, the system is characterized by a total freedom in the choice and the management of the databases. A tool like NESSTAR generally requires storing all data in its own workspace. With SEDO, the database server can be distinct from the server where the client requests are processed and the statistical operations are performed. Thanks to this independence of the data storage and the data processing, SEDO is portable to any kind of database. This separation is also useful in order to protect the confidentiality of the data: we must here consider the different kind of users, who won't be entitled to the same access to the data. It moreover minimizes the volume of the data used during the calculations, because exactly the data relevant for the requests are extracted and passed on to the processing phase.

5 Architecture

The challenge represented by SEDO is to combine the approaches for the different user categories within a general architecture capable of supporting arbitrary number of simultaneous connections and delivering the relevant information on demand.

The information contained in the PSELL is split into 3 parts: (1) the *variables*, which are the question of the survey, (2) the *population* interviewed, which is decomposed into individuals and households, and (3) the *results*, or answers to these questions. The PSELL 2 survey manages 236 variables for a population of 6242 entries, during a period of 7 years, or *waves*, from 1995 to 2001.

The variables as well as the information about the population are stored in SQL tables. The variables have been structured based on the UNBIS Thesaurus of the United Nations [Un04]. Such an exhaustive standard was needed in order to take into account the very precise topics covered by the PSELL. As a corollary, the thesaurus has been pruned to fit to the specialization of the domain. The variables as well as the information about the population are stored in SQL tables. To navigate on the variables, the thesaurus, with keywords browser and search engines has been implemented in the Macromedia Cold Fusion MX technology.

Concerning the results of the PSELL, several software packages are available on the market for statistic processing, among which the SAS system offers a special implementation dedicated to Web interaction, SAS IntrNet [Tr02]. We have successfully integrated the SAS IntrNet statistical environment, and this makes possible to perform directly on the server advanced statistical analysis. As for the storage of the results data, SEDO is declined into two versions: in the first one, we use a (possibly separate) SQL Server that select the data according to the requests and send them to the SAS system. In the second, the data are directly stored inside the SAS tables used for processing.

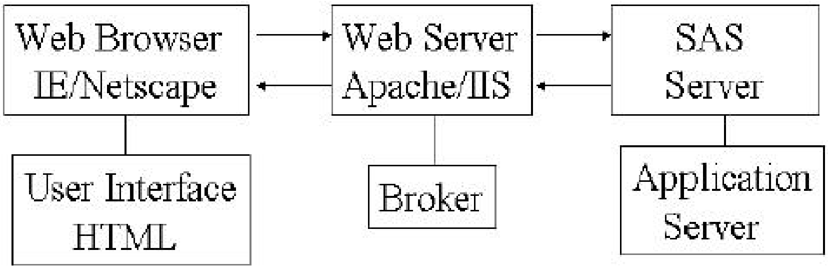


Figure 1: Integration of SAS IntrNet

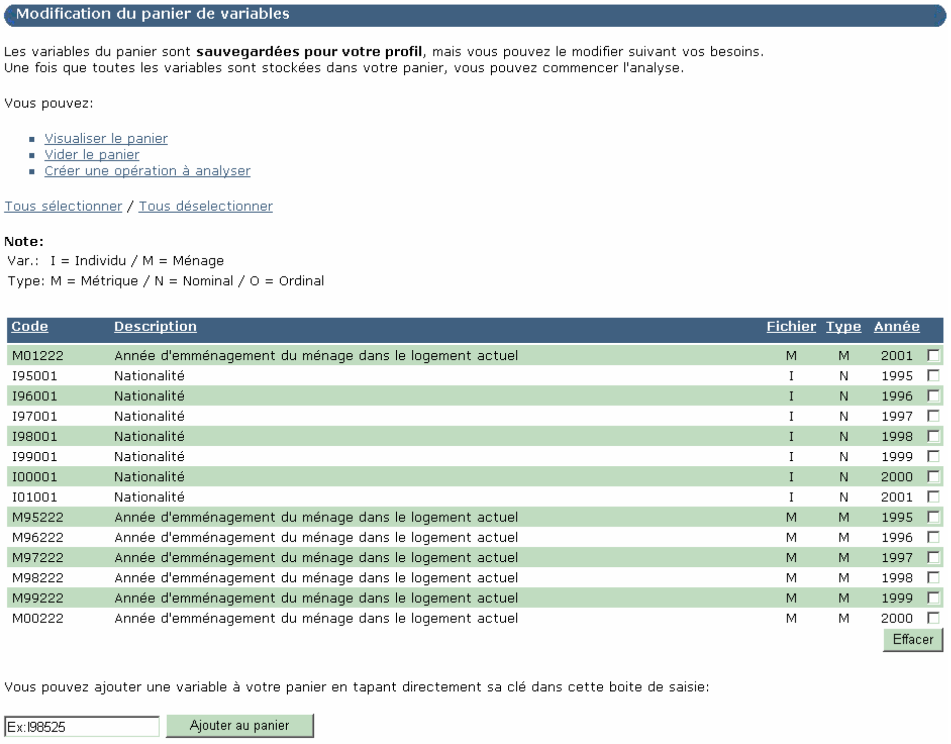
Due to the huge amount of information resulting of the combination of the variables, the statistical weaves, and the population in the survey panel, a special scheme of SQL tables has been designed to reduce the column numbers, and make Sedo portable on standard database management systems. A first table contains a longitudinal instantiation of the variable, which is the combination of the variable by the wave. A second table lists all the population individuals and household separately. A third table is made of the combination of these two tables, and contains the results.

With every single variable or result are furthermore associated information types such as the statistical measurement (metric, ordinal, nominal), the population concerned (individuals, households), or the expected results for the variable. Based on these types, the system define the statistical analysis that can be performed in order to ensure the coherence and soundness of the operations delivered to the user.

6. Practical results

On the current Sedo prototype, several accesses to the information are available. Summarized results are displayed with a minimal number of clicks through charts introducing the main tendencies or a direct browsing of characteristics tables. Complete databases are searched by 3 methods: a hierarchical approach implemented upon a self-documented thesaurus, a search engine with Boolean operators, and direct access by variable identification. The search engine performs searches on the thesaurus keywords or descriptors (in term of metadata) associated with the variable.

From the practical point of view, the user selects the variable by one of the methods and constitutes a basket that can be edited on demand. For every variable, the navigator shows the waves for which the variable is available, and gives a technical description based on the metadata. Using the variable in the basket, the user constitutes a list of statistical operations and sends it to the system. The system checks the correctness of the operation according to statistical types, then performs the correct computations and sends back the results on the fly.



Modification du panier de variables

Les variables du panier sont **sauvegardées pour votre profil**, mais vous pouvez le modifier suivant vos besoins.
Une fois que toutes les variables sont stockées dans votre panier, vous pouvez commencer l'analyse.

Vous pouvez:

- Visualiser le panier
- Vider le panier
- Créer une opération à analyser

Tous sélectionner / Tous déselectionner

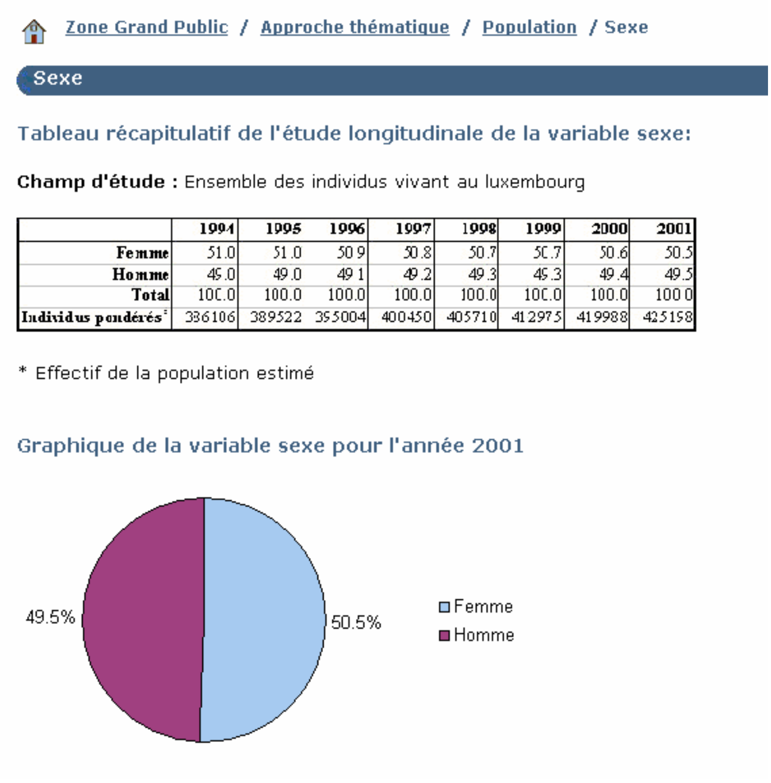
Note:
Var.: I = Individu / M = Ménage
Type: M = Métrique / N = Nominal / O = Ordinal

Code	Description	Fichier	Type	Année	
M01222	Année d'emménagement du ménage dans le logement actuel	M	M	2001	□
I95001	Nationalité	I	N	1995	□
I96001	Nationalité	I	N	1996	□
I97001	Nationalité	I	N	1997	□
I98001	Nationalité	I	N	1998	□
I99001	Nationalité	I	N	1999	□
I00001	Nationalité	I	N	2000	□
I01001	Nationalité	I	N	2001	□
M95222	Année d'emménagement du ménage dans le logement actuel	M	M	1995	□
M96222	Année d'emménagement du ménage dans le logement actuel	M	M	1996	□
M97222	Année d'emménagement du ménage dans le logement actuel	M	M	1997	□
M98222	Année d'emménagement du ménage dans le logement actuel	M	M	1998	□
M99222	Année d'emménagement du ménage dans le logement actuel	M	M	1999	□
M00222	Année d'emménagement du ménage dans le logement actuel	M	M	2000	□

Vous pouvez ajouter une variable à votre panier en tapant directement sa clé dans cette boite de saisie:

Ex:I98525

Figure 2: Management of the variable basket



Zone Grand Public / Approche thématique / Population / Sexe

Sexe

Tableau récapitulatif de l'étude longitudinale de la variable sexe:

Champ d'étude : Ensemble des individus vivant au luxembourg

	1994	1995	1996	1997	1998	1999	2000	2001
Femme	51.0	51.0	50.9	50.8	50.7	50.7	50.6	50.5
Homme	49.0	49.0	49.1	49.2	49.3	49.3	49.4	49.5
Total	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Individus pondérés*	336106	389522	395004	400450	405710	412975	419988	425158

* Effectif de la population estimé

Graphique de la variable sexe pour l'année 2001

Figure 3: Output of the Sedo navitator

For the lower level of expertise, Sedo uses bibliographic information with figures extracted from the survey databases. The service contains a protected area for experts, with user identification. In this area, baskets of variables can be stored to query the tables and perform statistical analysis, results can be downloaded, and a special memory keeps tracks on the server of the operations done and the preferences of the user so that it can retrieve his customized configuration from any place on the Web.

Conclusion

The Sedo navigator makes statistical information and analysis available on demand to users. The effort will now be put on the integration of the service. First of all, we have to ensure the effective customization and reutilization of the platform, which will be internally and externally tested at the CEPS/INSTEAD. Second, for the sake of improving exchanges of scientific data, we want to concentrate on the interface with other statistical tools, and in particular with the Nesstar software used by most of specialists of the specialized community [As02]. Sedo will offer export and import functionalities for the Nesstar format.

References

[Tr02] Truong S.: Communicating Data Effectively with e-Data and e-Notes, Meta-Xceed Inc, http://www.meta-x.com

[As02] Assini, T.: NESSTAR: A Semantic Web Application for Statistical Data and Metadata; WWW2002 Conference, Hawai, USA, 2002; http://www.nesstar.org/papers

[Su97] Sundgren, B.: Making Statistical Data More Available, Workshop on R&D Opportunities in Federal Information Services, Virginia, USA, 1997; http://www.isi.edu/nsf/papers/sundgren.htm

[Un04] The Multilingual UNBIS Thesaurus of the United Nations: http://unhq-appspub-01.un.org/LIB/DHLUNBISThesaurus.nsf

[Br98] Browne, S.; &al.: Technologies for repository interoperation and access control; Proceedings of the third ACM conference on Digital libraries, Pittsburgh, Pennsylvania, United States; Pages: 40-48, 1998.

[Me04] The Metadata Architecture: http://www.w3.org/DesignIssues/Metadata