



**Economic and Social
Council**

Distr.
GENERAL

CES/AC.71/2001/18
30 October 2000

ENGLISH ONLY

**STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE**

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

CONFERENCE OF EUROPEAN STATISTICIANS

Joint ECE/Eurostat Meeting on the Management of Statistical Information Technology
(Geneva, Switzerland, 14-16 February 2001)

Topic (ii): Challenges and opportunities for statistical offices working in a network environment

**THE TECHNOLOGICAL EVOLUTION IN STATISTICS INSTITUTES:
RISKS AND OPPORTUNITIES BASED ON THE ISTAT EXPERIENCE**

Submitted by ISTAT, Italy ¹

CONTRIBUTED PAPER

I. THE IMPACT OF NEW TECHNOLOGIES

1. The inexorable growth of the technological evolution is significantly modifying the production process, profoundly affecting the information technology function. The statistics institutes in a position to adapt to innovation can make use of new tools, new functions and new services in order to benefit from the opportunities presented to them.

2. ISTAT represents a context in which an evolution process is underway. This process is characterized by the changeover from centralized systems (mainframe), where all the statistical production phases took place, to distributed systems, in which PCs and UNIX work stations, from the client side, and specialized processing power for services, on the server side, share a fast and open network.

¹ Prepared by A. Guarino, P. Paparo and S. Ronzoni.

3. This adaptation has paved the way for a new phase in which technological innovation increasingly affects the internal organizational structure as well. In fact, since this shared environment of resources and applications has made it possible to distribute the work load among the various local production organizations, the technical personnel who once belonged completely to a centralized function can be distributed today throughout the other production sectors.
4. The information technology function tends to be integrated in the production areas whereas the technological sector assumes a new role of orientation. This new role is implemented through specialized centers which, through the technological evolution, establish standards of architecture, HW and SW tools and development methods.
5. Communications play a fundamental role in this context and this is the reason that ISTAT has focused on opening up to the Internet, offering all personnel the chance to access the worldwide web. While on the one hand this makes it possible to develop individual professional skills, on the other it creates the risk that the different production units will independently develop solutions with the most diverse range of technologies on the market. The uniformity and integrality of the different systems can be guaranteed only by issuing and observing specific and centrally-dictated standards.
6. On a more serious note – for ISTAT in particular, but for any statistics institute in general – the technological evolution affects the production process during the main phases of collection, processing and dissemination.
7. While the changes that the new network technologies have brought to the production process make it possible to grasp new opportunities and take on new challenges to improve services, such changes also represent elements of risk that, if not assessed in advance, can lead to undesirable effects that may even damage the very image of the institute.
8. Therefore, the risks and opportunities faced by an ISTAT online are analysed in the following sections, presenting several implementations that may prove useful as case studies.

I.1 The collection phase

9. A technological environment based on the telematic network affects all phases of statistical production and, in particular, considerably modifies the procedures and the operative flow of the data production phase. In fact, traditional measurement methods have been joined by new methods such as data collection by transferring files using the “ftp” protocol or “e-mail”, the development of specific “client/server” applications and the introduction of “web-based” applications operating in Internet and Extranet environments.
10. The strong point of online data collection is undoubtedly its timeliness: the collection time is shortened, and the data can be used immediately by the institute. The respondent, availing himself of an application that is accessible on the Internet, for example, sends the data to the statistics institute which, through automated procedures, is already in a position to organize the data and make them available for analysis. The natural link between the Internet and relational databases makes it possible even in this early phase to create an organized data structure that is transformed for researchers into a homogeneous source valuable for the purpose of analysis.

11. While all this requires new skills and instruments, it is carried out with a significant decrease in various administrative expenses. Tapes, magnetic media and the various questionnaire sheets used for surveys are gradually being replaced by hardware and software, reducing publication costs as well as the costs required for the physical organization of the survey network.

12. Another aspect that must be considered is the enormous flexibility of web-based applications. They make it possible to create user-friendly interfaces with more effective control over the typing of data, thus minimizing input errors. Furthermore, they permit the collection procedure to be monitored directly.

13. Obviously, the consequence of the new technology is that research institutes must create online help functions to further facilitate completion of the questionnaire by the user. It also means that direct contact should be maintained between the responder and the statistics institute. In this context, tools such as FAQ's, news groups, chat clubs and so on can be extremely helpful.

14. During the collection phase, the network helps streamline the operating procedures while also taking the technological limitations for good system design into account. In fact, the network service must include elements such as reliability and performance of the transmission system, essential in defining the architecture of online applications. For a properly designed collection system, it is important to assess the network load as well as the hardware and software that are utilized, also setting up control and recovery mechanisms in the event of malfunctions during the transactional phase.

15. Another element that cannot be neglected concerns security. Collection involves the permanency of data for a given amount of time in external media open to the Internet (database, file system, etc.), and an online transaction of the data.

16. Concerning security, automatic and technological systems must be adopted to minimize the risk of the data being read or lost. Instead, with the online transmission of data, there are problems related to authentication, data confidentiality and integrity; in other words, there should be a guarantee that the data will not be altered during transmission. There are specific protocols and instruments such as digital certificates and digital signature to solve these problems.

17. However, it is important to avoid the risk of proceeding superficially in order to speed up innovation, and thus failing to confront critical aspects such as careful design of the systems, their security and the standardization of development instruments.

18. In addition, the fact that the new collection methods can address a broader range of consumers means that making the service available to the user in an unlimited time frame and dealing with technical problems such as reliability, efficiency, security or online help-desk functions can guarantee quality in the collection process and create trust among users of these new work methods. All of the above represent a valid opportunity for the institute in terms of the constant extension of statistical analyses which investigate the increasingly diverse aspects of the country's socio-economic situation.

I.2 The processing phase

19. The hardware and software architectures have changed extensively, progressing from a model where all the operations were performed by a central processor (mainframe) connected to non-intelligent terminals, to a distributed architecture where the work load is subdivided among the various machines, yielding advantages in terms of processing power and the optimization of resources. This architecture, defined as "client/server", nevertheless requires careful design and management in order to avoid

disappointing results, because there are many factors at play: the work load on the server, the client's processing power, the network occupation, the number of users and of operations on databases, etc.

20. Client/server applications are oriented more to the end-user because they furnish a rich user interface and permit multitasking activities. However, the client/server architecture entails installing the applications at each desktop, with high installation and administrative costs. A server-based distribution solves these problems, shifting the complexity from the desktop to the server where the applications can be managed centrally and downloaded to clients on demand.

21. The natural evolution is thus the use of a web-based architecture in which the work load is typically distributed among several servers (application server, database server, web server) and a set of clients who simply require browser software so they can work with the applications. The leading benefits of this type of architecture are its platform independence (since the only requirement for the client is to install a browser), and the fact that there is a single environment to access the applications with an ensuing improvement in terms of learning and utilization on the part of users.

22. In this type of architecture, it is important to attain the proper balance between attractive interfaces and simplicity of content in order to offer a service that is attractive for the user but that can also be supported by communications systems.

23. One of the main strong points of net-based architectures involves the use of relational databases. An important advantage is gained from the sharing of data by different corporate information technology systems. In the case of a statistics institute in particular, it is possible to share metadata and data common to several surveys. In fact, often the data monitoring and validation activity must be performed using information drawn from other surveys. These advantages are guaranteed through the constant updating of the database by the services producing the information.

24. Yet another positive element is a drastic reduction (theoretically, the elimination) of data redundancy, which can be achieved through proper design integrating the information coming from the various surveys. Another important characteristic involves the unique formatting of the data and their meanings. In particular, with metadata, users (analysts, researchers, students, etc.) can avail themselves of a dictionary on the data which specifies the source, the reference dates, the units of measure that were adopted, etc.

25. A further development is the creation of the data warehouse, i.e. the set of data structures and instruments required to obtain summarized information, starting with heterogeneous data received from corporate IT systems. These systems initially became widespread as a decision-making support for managers and marketing operators, but they are also highly adaptable for research offices and thus for statistical researchers. In fact, the availability of the data warehouse permits the circulation of processed data in a format that is not prepackaged. Modern software products make it possible to perform analyses such as data mining, a data-filtering technique designed to discover or verify correlations (or patterns) among available data, plus what is known as "what if", a generic term for techniques with neural or statistical algorithms to forecast future data from series of input data.

26. The changeover from traditional systems to distributed systems has fostered a propagation of development software available on the market at increasingly accessible costs, not to mention the ones that can be downloaded from the Internet free of charge. This fact has greatly expanded the possibilities of developing ad hoc software with versatile and effective instruments, also permitting the development of programs by users who are not computer specialists. Moreover, there is a proliferation of non-standard

software, thus making it difficult for EDP personnel to guarantee the necessary servicing and optimization of resources.

27. A statistics institute includes professionals with widely different roles. In terms of information technology, in addition to the EDP technical staff, there are users who utilize programs and procedures to produce the results of surveys, with all the data loading, correction and validation operations that are involved, as well as users who utilize non-proceduralized software to analyse the data and conduct studies. The presence of distributed software that is easy for non-expert users to run greatly accelerates the processing and publishing of the results when it is not possible to develop procedural applications.

I.3 The dissemination phase

28. Within the production process, outgoing communications, and particularly the spread of information to the public, is the phase that benefits most from the changeover from "host-centric" architectures to "network-centric" ones. In its broadest possible meaning, the communications network effectively becomes a circulation instrument with unthinkable potential that is capable of carrying the load of information, data and analyses of statistics institutes on the desk of each individual, regardless of whether he is an ordinary citizen or an expert researcher. It also makes it possible to bridge the cultural gap between countries, permitting an efficient exchange and complete circulation of information in order to move gradually towards an overall view of socioeconomic life.

29. The spread of information over the Internet toward levels of consumers with different needs and demands has evolved rapidly from static information to dynamic information through the development of complex IT systems characterized by opening up to the Internet and easy navigation. Here, the producer of static information is the one to publish the results, thus drastically reducing circulation time and costs. The user himself creates his own demands for processing and extracting data at different aggregation levels by interacting directly with the databases available from the institutes. Data exist in different representation formats, ranging from a pure chart form that can immediately be processed to a structured and georeferenced form.

30. As statistics institutes turn increasingly towards the web for the circulation of information, there is likewise an increase in the benefits for the community, in terms of decreasing the time and cost required to publish results and in terms of the amount of information directly available. However, there is also an exponential increase in the potential lack of security of the system. In fact, both data and system integrity are at increasing risk and, through them, not only the statistical information but the very image and credibility of the institute can be damaged.

31. The issue of increased security of network systems is a priority, since the careless design of security measures represents a high risk factor. There has been a real cultural revolution in the spread of information, from the time that electronic information existed only as a backup to printed documentation through to today, when it has its own role as primary information. This is one of the greatest opportunities offered by the use of open networks, in which e-publishing and e-commerce make it possible to reach user levels that were previously unfathomable and to develop new business possibilities.

32. In order to maximise such opportunities, it is necessary to supply a truly attractive system for customers, paying attention to the publication of data constantly and punctually updated. It is also necessary to develop services based on the effective needs and demands of customers, including security and the ease of on-line payment service.

33. However, it is also becoming increasingly necessary for statistics institutes to protect their e-publishing business by relying on techniques that will guarantee the copyrights of their publications. A lot of opportunities for and solutions to confidentiality protection can be found by adopting techniques based on digital certificates and digital signature. This is particularly true since, as is the case in Italy, they are recognized as legally binding, thus the electronic document represents primary information that is legally valid.

II. THE EVOLUTION PROCESS IN ISTAT

34. In 1999, from an information technology standpoint, ISTAT focused its attention as well as its human and financial resources on the Y2K transition, and this coincided with the adaptation of its processing environment. Thus, on 31 December 1999, it disengaged the centralized systems, simultaneously activating the UNIX system.

35. This new environment, based on servers and workstations distributed in the various production sectors, together with the use of Internet technologies, has made communications the focus of the system and one of the foundation elements of the entire production of statistics. This is precisely the reason that, after the transition was completed, attention and resources were focused on the technological innovations needed to develop new systems and new procedures. The paragraphs below describe some of the developments implemented at ISTAT which highlight the risks and opportunities of these technologies; they could be valuable topics of reflection for those who intend to follow the same course of action.

II.1 NEW PROCEDURES FOR COLLECTING DATA ON-LINE

36. At ISTAT, the electronic questionnaire is a new tool used to acquire data, completed by the responder directly on the Internet. Since we are still in a phase of transition toward the new procedures, this represents an extra opportunity for the user to put the new technologies side-by-side with the traditional collection mechanisms. One of the features of ISTAT's electronic questionnaires is that they are similar to the paper model, so that the responder can use the printed form as a guide. One of the great opportunities offered by an electronic questionnaire is that it permits direct addressing depending on the answers furnished as it is being completed. The questionnaire is tested for its usability, its functional, aesthetic and cognitive parameters because, during the interview, the filter ensured by the surveyor is no longer present.

37. The institute has a UNIX collection server connected to the Internet and equipped with web and application servers that make it possible to obtain data through electronic questionnaires in order to collect data from businesses and other organizations. At the moment the responder is identified through identification codes. Work is under way to define a security architecture that will guarantee confidentiality, integrity and a powerful authentication achieved through digital certificates and the use of SSL (Secure Socket Layer) protocol.

38. Based on an analysis of the results gained from monitoring the use of these questionnaires, it has emerged that the responders who use these means instead of the traditional procedure amount to about 25%. Another parameter that is satisfied by the electronic questionnaire involves timeliness: in fact, the data from the questionnaires filled out on the web arrive much earlier than the printed ones. Another interesting element is that many structures – about 35% – furnish an e-mail address, and as a result this information is used to create a direct link between the electronic questionnaire and the user and to solicit the submission of electronic interviews in the future.

39. Another type of data that ISTAT obtains through the web involves the monitoring of questionnaire

collection. While the agricultural census for 2001 will collect data exclusively through the printed questionnaire, it nevertheless envisions using the web to monitor the compilation of the forms. The web technology used will be one of the most innovative ones available: Java Servlet on UNIX web server.

40. Another option used by ISTAT involves obtaining consumer prices, conducted by surveyors completing an electronic questionnaire on palmtop computers. Finally, a simple and effective transmission mode, adopted to transfer large amounts of information, is used to transfer administrative and financial data from outside organizations such as the Social Security Agency and the Central Bank to ISTAT. This process takes place via an Extranet with encryption and authentication mechanisms on an IP level. The transfer of information from these organizations is handled through an external ISTAT server and, with timed procedures, it is then filed into a centralized internal server and distributed to the production sectors.

II.2 An example of network processing

41. The new IT system for the production of Foreign Trade statistics has an architecture that faithfully reflects the institute standards with regard to the data structure and the software that is used. The migration process from the pre-existent IT system to the new distributed systems began in late 1988 and has involved about 260,000 statements for over 15 Gb of disk space, thus “upsetting” the operating procedures in the production sector (more than 80 people).

42. The starting point for data analysis and functions was a study of the data structure and functions of the old Adabas databank. An analysis of user requirements and old application limitations helped to complete the project of a new consultation databank. The adopted architecture can be simplified with a three-tier structure: a UNIX server application with SAS instruments, a server database with Oracle and a client PC operating under Windows with specifically-developed forms.

43. The adopted solution significantly decreased the E-R layout and thus the datamarts for data dissemination, simplifying the phases for updating, managing and checking the consistency of the data. The underlying concept was the creation of a single de-normalized structure containing the entire hierarchy of commodity classifications. This solution has proven to offer high performance during the timely request for data, with appreciable response times when total groupings are involved. Furthermore, it has significantly simplified the development phase and the monthly updating phase, as well as application management.

44. On a monthly basis, through the production environment, there is an update of the foreign trade data warehouse, which contains data that have been validated at the maximum level of disaggregation (microdata). The Oracle procedures are then carried out in a client/server mode, creating or updating the aggregation tables that are useful for the dissemination phase. Through SAS procedures, these tables in the Oracle environment then supply the datasets located in the databank that, through the Internet, meet the requests of the institute’s researchers.

45. The data warehouse development of the Foreign Trade Service represents an example of the institute’s policy of organizational and technological innovation. This is reaping benefits not only internally to guarantee greater interrelations among surveys and better data quality, but also externally with the data dissemination on the Internet.

46. Static pages and dynamic web applications have been created. The first mode allows users to access the static HTML pages created by the website provider. Applications of this type usually represent corporate reporting in general, with charts and graphs that remain stable for a given amount of time. The second mode allows users to interact, through their web browsers, with application server sessions active

on the corporate network for interactive data navigation in an OLAP mode.

47. The results achieved through the development of this new IT system are substantial and there is no doubt that they have fully satisfied expectations. The enhanced efficiency and flexibility of the new databank has substantially reduced the time necessary to issue static information and has permitted the online creation of a wide range of queries that formerly required the compilation of specific programs. The development of a statistical IT system in a client/server architecture operating under UNIX with RDBMS Oracle has made Foreign Trade Statistics available within the distributed system of the institute's interconnected databases. The optimization of the procedures has helped minimize the time spent loading, checking and doing automatic correction.

48. A new, more intuitive user-friendly interface of superior performance has been developed for reviewers, and new functions have also been added to facilitate the review process for the user. The re-engineering of the production process has permitted both the verification and revision of a statistical method by introducing new rules for automatic data correction and new variables that were not considered before in order to achieve better survey quality.

49. The problem of training was, of course, a consequence of the decision to overhaul the entire production process. It was obvious that such a major transformation would involve personnel on every level. In one year, the professional growth of all the resources of the production sector was substantial: the programmers are able to operate in both the Oracle and SAS environments and the entire staff can operate on a Windows platform with typical office-automation functions. The possibility of accessing Foreign Trade Data over the web will assist all researchers interested in sector studies or particular analyses, making enormous amounts of information available to them.

II.3 New dissemination techniques

50. The service that currently calculates the price variation index has recently modified its IT architecture for the production of these indicators. The previous IT system involved a series of COBOL programs that read files of indexed sequential data, and which were substantially controlled by human operators. So only a few expert programmers were able to conduct maintenance and develop programs for non-standard data extraction.

51. Therefore, the entire production process was renovated in order to implement a new IT system based on relational data, using sophisticated and innovative collection and dissemination techniques. The databases were constructed working in an Oracle environment and a substantial reduction in data redundancy was achieved following this reprocessing phase.

52. One of the most innovative aspects of the new price system involves the legal certification of the indices. By acknowledging technological innovations and Italian regulations on electronic documents, the project's main objective is to make the certification of statistical data available online and to streamline the in-house administrative process.

53. The legal validity of the document produced using the new application entails the sure recognition of the server issuing the certificates; the integrity of the document is transmitted together with its signature. These requirements are effectively managed through the Secure Socket Layer (SSL), the X.509 digital certificates for authenticating the server and establishing secure communications on an http channel (HTTPS), and through digital signature techniques that guarantee the integrity of the document and give it legal value.

54. The application was developed in a three-tier client/server environment, where the client, represented by the browser, interacts with the application server to query the database and activate the procedure for the production and signature of the certificate.

55. Once this project, which is currently limited strictly to in-house users, reaches the final development stage, it will allow subscribers and Italian townships to request and directly obtain the transmission of a legal electronic document certifying the price indices. In this way, the institute becomes increasingly closer to the end user. The time required to issue a certificate has been minimized because the service, previously based on a manual signature and transmission procedure response, now uses a fully automatic response rendering the service request and results virtually simultaneous.

56. The institute is planning to expand the use of X.509 digital certificates to all employees and partners, with the creation of a public key infrastructure (PKI) for the collection and dissemination phases. The institute is also promoting the use of the digital signature to improve and streamline the other processes, such as the distribution and collection of questionnaires for the next population census.

57. It is also important to emphasize certain critical elements that must be faced. A structure involving digital certificates and digital signature procedures also entails a maximum security environment and the definition of a new organizational function that guarantees the operations for issuing and revoking the keys.

58. Finally, it is important to highlight the requirement for widespread training on the techniques and functions that have not yet been fully assimilated by most users, as well as the need to make users aware of the security problem.

III. CONCLUSIONS

59. The inexorable drive of technology, the progressive lowering of costs and the advent of increasingly sophisticated network services mean that institutes who want to benefit from new development opportunities must constantly monitor the market underlying any future decisions.

60. Therefore, it is vital to consider some of the techniques that have emerged recently, not only in the field of telecommunications (GPRS) but also in the Internet development environments (XML). Both GPRS and UMTS, the two data-transmission techniques for mobile telephones, are very interested in all cases involving the collection of data in the field, making them available in real time. XML (eXtensible Markup Language), an extension of HTML whose objective is to become the "lingua franca" of communications between IT systems, can ensure interoperability between different statistics systems.

61. Another opportunity that is now in an advanced phase of research involves obtaining consumer product prices using barcode scanners directly at the sales outlet. In this constantly evolving technological environment, it is worthwhile to choose the solution that is most advantageous for the citizen and for the institute itself; this does not necessarily, however, correspond to the most advanced solution.

Acknowledgments: We would like to thank Mr. G. Budano, Mr. N.R. Fazio, Mr. C. Martelli and Mr. A. Sorce of ISTAT for their collaboration.