



**Economic and Social
Council**

Distr.
GENERAL

CES/AC.71/2001/4
30 October 2000

ENGLISH
Original: ENGLISH and
FRENCH 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 (i): The impact of data warehousing on the management of statistical offices

NEW ARCHITECTURE FOR STATISTICAL INFORMATION SYSTEMS IN EUROSTAT

Submitted by Eurostat ¹

INVITED PAPER

I. INTRODUCTION

1. Eurostat, the statistical office of the European Communities, currently manages over one hundred computer systems. These contain statistical data and metadata, mainly from the Member States, which are worked on until, harmonised and often aggregated, they are made available to internal users, European institutions, Member States and citizens in general. These systems cover any number of diverse operations, such as data reception, checks, estimates, harmonisation, various transformations, analysis and dissemination, to mention only the main ones.

2. For reasons of economy and interoperability, it is imperative that the development and use of

¹ Prepared by Daniel Defays.

systems are closely coordinated. One means of achieving this coordination is to define a general architecture for information systems, and Eurostat has recently felt the need to update its architecture. This document sets out the reasons for these new developments, the general principles which they will have to uphold, the main elements of the architecture and the planned migration strategy.

3. This project is now at the start-up stage. What follows, therefore, is not a success story, but a plan which is still under discussion and which had not yet been validated when this document was written. Therefore these proposals commit the author alone.

4. There is no doubt that most statistical institutes face similar problems, and exchanges concerning any planned solutions or any which have been implemented are welcome. The participants at the MSIT meeting are therefore invited to comment on these proposals in the light of their personal experience.

II. THE REASONS FOR CHANGING THE ARCHITECTURE

5. The current architecture is structured in environments. An environment, in Eurostat jargon, is defined on the basis of a set of data, the activities/functions linked to it, and the corresponding operators. Three separate levels can be identified: production, reference and dissemination. Upstream, standard collection tools are made available to users. The production organisation is decentralised, while specific units in reference and dissemination ensure substantial coordination.

6. This model needs to change for several reasons:

- ◆ The decentralised management of production is reaching its limits; the associated costs have to be better managed by better coordination, and the architecture has to contribute to this process.
- ◆ The ever-increasing volume of confidential data to be processed and growing concern for their confidentiality create the need to rethink how these data are processed, and thus to review the concept of a production environment.
- ◆ At present, the data which Eurostat receives are not routinely archived in any coordinated way. If ever a problem arises, it is no easy matter to rerun a processing sequence on the original data.
- ◆ The growing need to document data and to associate meta-information with them to facilitate processing and access means that statisticians are increasingly dealing with non-numerical data, and the rôle these play and their place within the current architecture are not precisely enough framed.
- ◆ The linear model of the statistical process (collection, production, dissemination) underlying the architecture does not always fit the bill in the event of shared production, when Eurostat provides data from other sources without any internal processing or dissemination of metadata. The Internet also plays a part in exploding this linear and sometimes closed view of the statistical process, and increasing numbers of administrative applications do not sit well with this breakdown, either.
- ◆ The current version of the reference environment is more geared to the needs of production than those of dissemination: producers can locate their data easily, but clients often face real difficulties in finding what they are looking for. This is because there are not enough metadata to help a search, and there is no single, client-oriented model for data presentation.

III. SOME CHARACTERISTICS TO BE DESIRED OF THE NEW ARCHITECTURE

7. Analysis of the life cycle of data in Eurostat and of the problems encountered with the current

organisation point to the following conclusions.

8. Essentially, four sets of strategic data have a role in structuring the architecture of Eurostat's information systems and organisation:

- ◆ raw data and metadata from information providers;
- ◆ cleaned-up, harmonised data which constitute the basic ingredients for any statistical product; called internal reference data, these may be confidential;
- ◆ external reference data which users can consult;
- ◆ disseminated data.

9. Data production is heavily decentralised. The architecture has to go fit this form of organisation and afford users maximum flexibility. On the other hand, there are areas in which the production process is more rigorously coordinated (data reception via Stadium/Statel, data provision in Comext and NewCronos). These are Eurostat's interfaces with the outside world and, for the sake of consistency and efficiency, the current degree of centralisation has to be maintained.

10. The secure environment concept needs to be adapted to make it easier to work with data collections containing any confidential information. Ideally, a sizeable part of the production environment should be located in a secure area.

11. The architecture has to be designed in such a way as to encourage the interoperability and reuse of different systems or applications. This will permit economic, decentralised operation. Interoperability will be guaranteed by defining standard interfaces and appropriate exchange mechanisms between information systems.

12. Standardised metadata should enable the consistency of data to be guaranteed and permit these to be documented. They will be defined according to the needs expressed by clients, producers and others all the way up the chain to the information providers.

13. In the interests of effectiveness, reliability and saving resources, it is also important:

- ◆ only to adopt solutions which have been tried and tested in other environments;
- ◆ to provide for gradual change based on thorough knowledge of the current situation.

This last condition is bound to constrain subsequent investigations.

IV. BENEFITS FOR DATA PRODUCERS AND USERS

14. A job as big as the revision of the general architecture of the information systems can only be taken on if it is reasonable to expect substantial benefits for the institution itself, its data producers and its clients.

15. The new model has to offer:

- ◆ substantial rationalisation of computerised tools;
- ◆ tools which are better focused on the needs of the different client groups: internal users in Eurostat, the

Commission's Directorates-General, outside suppliers (National Statistical Institutes etc.) and other clients, whether specialists or generalists;

- ◆ better access to metadata and better linking of databases;
- ◆ greater traceability of the processing operations carried out;
- ◆ easier processing of data, and of confidential data in particular.

V. THE FOUR ENVIRONMENTS WHICH MAKE UP THE NEW ARCHITECTURE

16. An analysis of the life cycle of data, a survey of the operations which data undergo and the definition of the environment concept together point to a four-level structure which is slightly different from the existing one. It is important to note that the environments are not necessarily separate physical constructs.

V.1 The reception environment

17. This is organised around the storage of the data and metadata provided by our national correspondents. This "production data repository" constitutes Eurostat's fortune in a way, and is the starting point for any processing carried out. This environment has links with the activities of "data capture", "validation" and "error correction".

V.2 The production environment

18. This is organised around internal reference, and is the sole source of input to the reference and dissemination environments. This is where primary data are stored once they have been validated and corrected, and where estimates and derived data are constructed by combining data from different domains as appropriate. Access is open to Eurostat staff only.

19. This environment contains all of Eurostat's microdata and macrodata, including the confidential parts. It also contains all the metadata necessary for production and for the external reference and dissemination environments (dictionaries, nomenclatures, key words, multilingual headings, footnotes, methodological notes, various links between statistical objects, formulae, etc.)

20. It is linked to "transformations and derivations", "estimation", "data inspection and editing", "statistical analysis", "nomenclature preparation and housekeeping" activities.

V.3 The reference environment

21. This is organised around external reference, and contains quality, non-confidential data and metadata for dissemination outside Eurostat. It is accessible via a single interface.

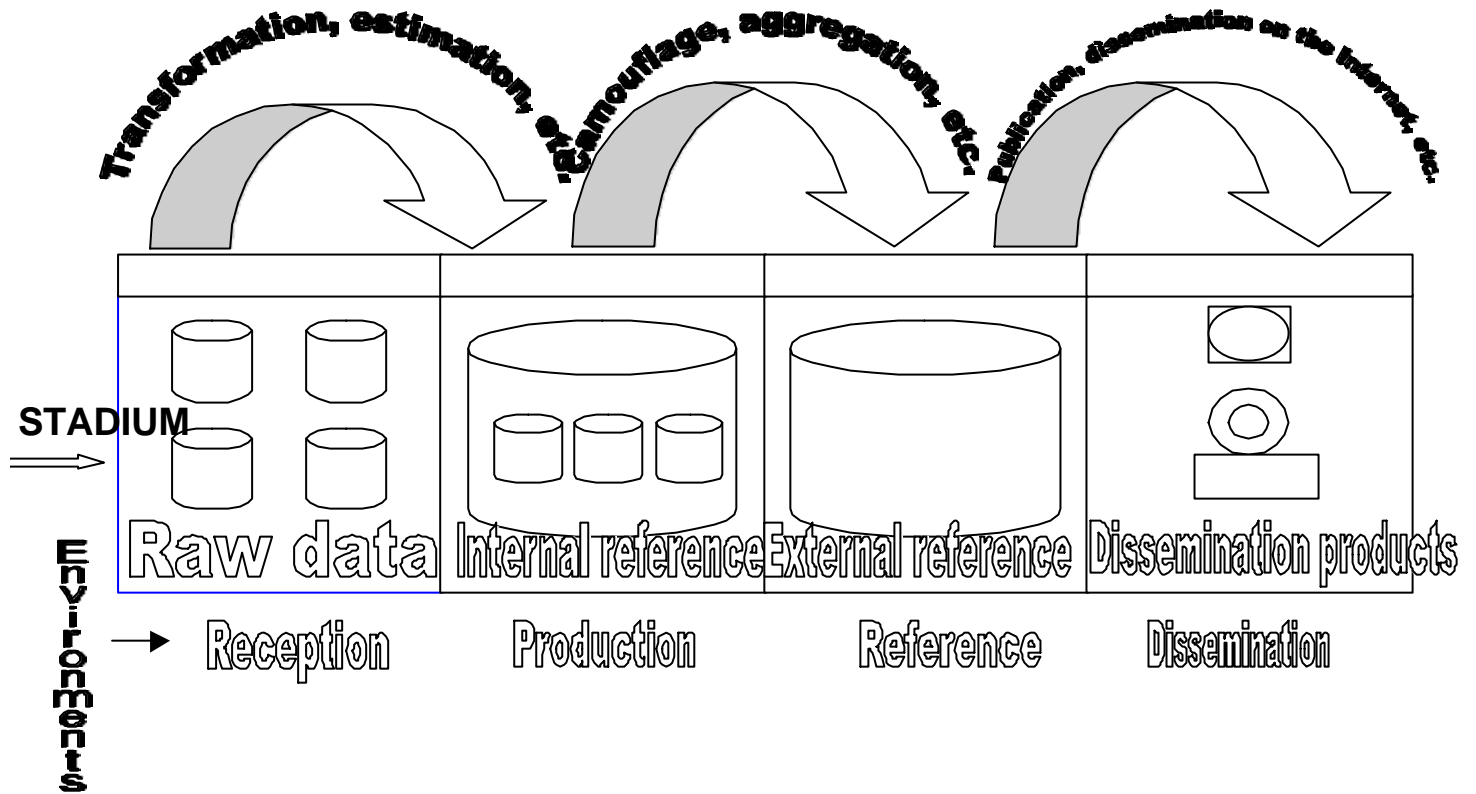
22. The public concerned is any individual or organisation with a regular interest in detailed statistical data, European and international institutions, ministries, data shops and data providers.

23. This environment is linked to the activities of "camouflage", "aggregation", "application of flags and footnotes", "supply of data to reference DBs", "nomenclature preparation and house keeping" and "statistical analysis".

V.4 The dissemination environment

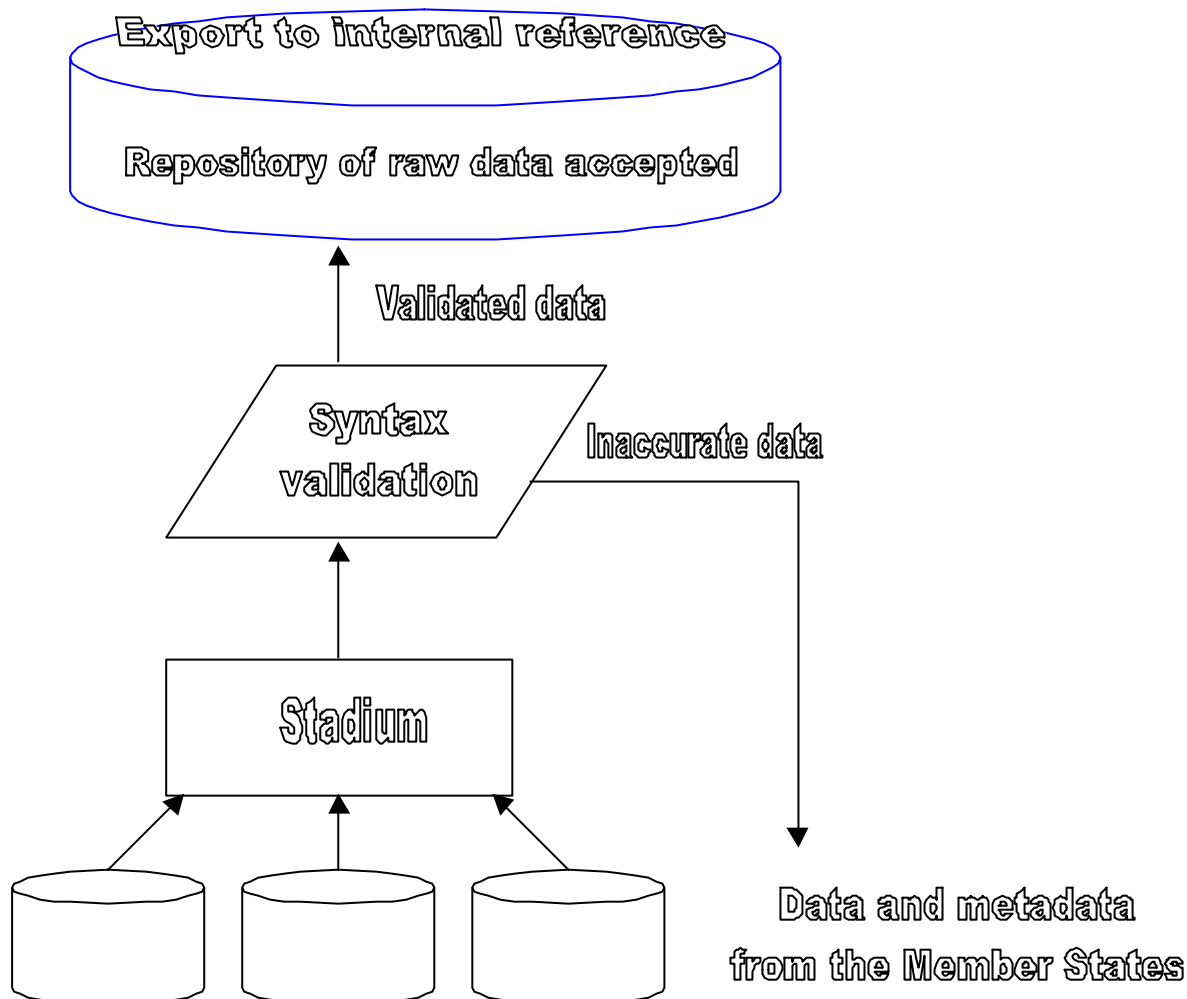
24. This is organised around an electronic version of all the data disseminated by Eurostat. These data are Eurostat's shop window. This environment is linked to "preparation for publication" and "dissemination" activities.

Graphique 1. Les environnements constitutifs de l'architecture



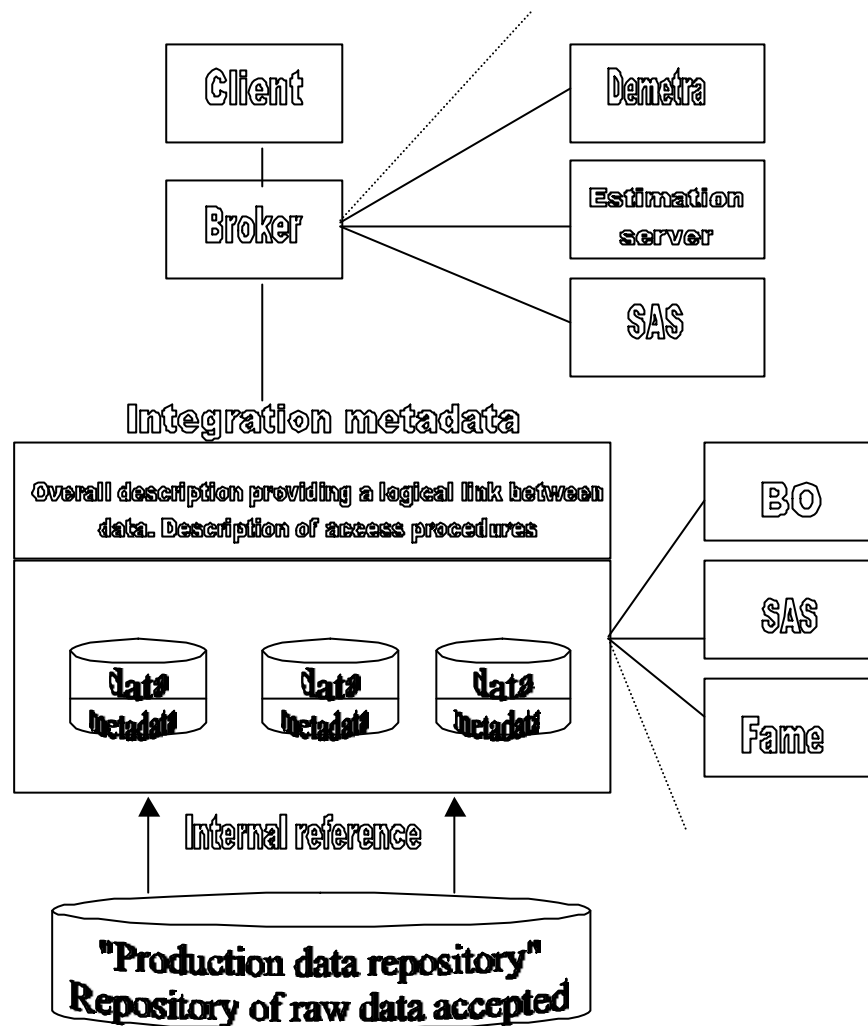
VI. THE COMPONENTS OF THE SOFTWARE INFRASTRUCTURE

25. Each of the environments described above is reflected in databases (real or virtual), toolboxes, software components, etc. This paragraph describes what we are aiming for and where the systems stand at present.

VI.1 The reception environment

26. This environment is organised around existing tools for data reception/recording (Statel, Ediflow and Stadium). The data are sent using GESMES. The producer receives the information transmitted via Stadium and checks the syntax. The file may then be returned to the sender. Once validated, the data are recognised by Eurostat and the sender as the entry point to the processing chain. They are archived. At present, this takes the form of a pointer to a file which is directly controlled by the producer. In time, storage using a single DBMS would be desirable in the interests of backing up and interoperability

VI.2 The production environment



27. The production environment, and internal reference in particular, should cater for interdomain communication and easy, secure access by external processing software. Its design will draw on the existing set-up, and add a descriptive layer to permit virtual integration of data. The contents of this layer, called "integration metadata", will include the following,

- ◆ descriptions of pivot forms and name correspondences for each type of data accessible;
- ◆ pointers from physical objects to composite objects;
- ◆ descriptions of the types and physical formats of accessible data, including the type of storage software;
- ◆ information on the location of data and on elements to permit harmonised queries (dimensions, standard values, etc.);
- ◆ a description of the basic behaviour of the software queried (return errors, etc.);
- ◆ the names of statistical functions, their input parameters, types of results and the location of the application programme.

This is quite clearly the hub of this environment.

28. Furthermore, insofar as possible, the functions will be encapsulated in reusable software tools. The operations should proceed as follows: the user accesses the information contained in internal reference (physically stocked in the SAS, Oracle, Fame bases etc.) via a client component, which would be capable of constructing requests, initiating analyses involving different application servers; and browsing results. The client sends the query to the "broker", the front-end software which essentially receives orders, routes them to the application servers or data servers, receives messages on the progress of the work and then makes the appropriate decisions.

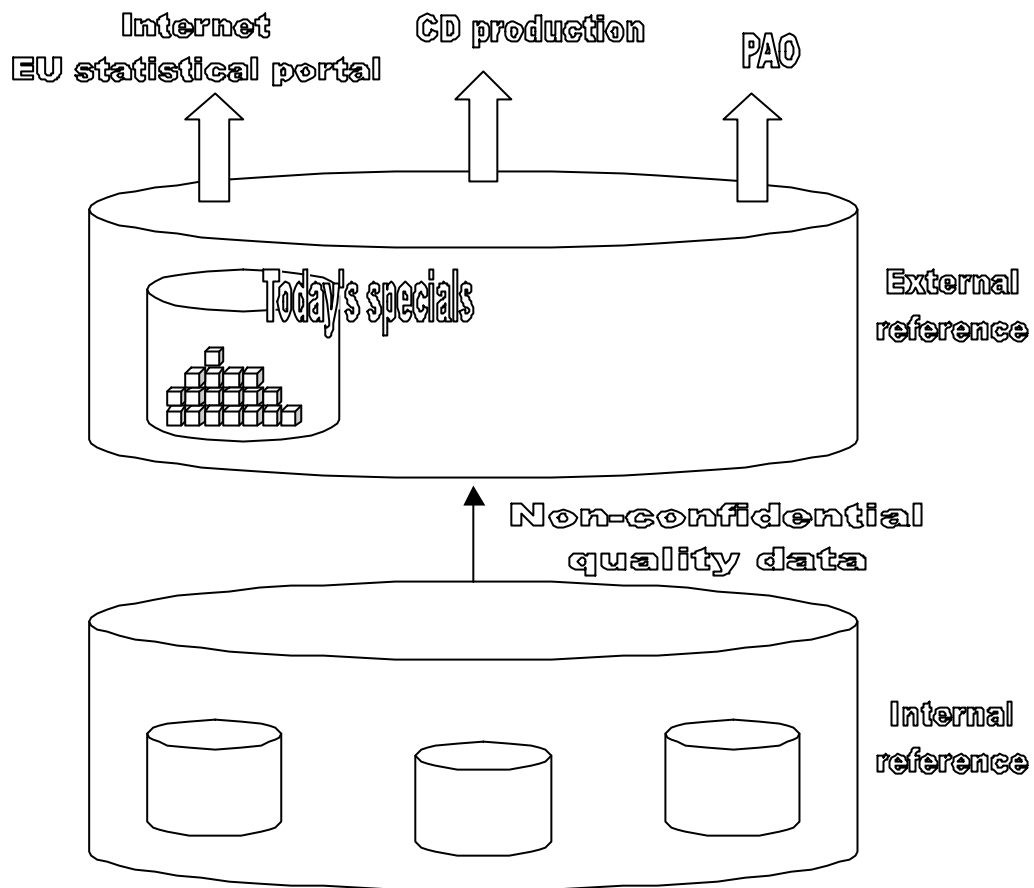
29. Generic processes are performed by reusable components in the form of what have been called application servers (Demetra, estimation servers and SAS applications in the figure). An application server is a piece of software, frequently with no memory, which can be used separately. The inputs and outputs are standardised to permit process sequencing.

30. It is also planned to keep alternative software types (open transactions - with no return to source) which do not use the broker. These are products like SAS, Oracle Express, Business Objects, Fame, TPL or ACCESS/EXCEL. One feature common to all these is the use of SQL and the creation of their own repository. They are another essential part of the production environment. At present, they are widely used for analysis and for performing complex functions in the production environment (indices, sampling, etc.), and they would appear to be irreplaceable.

31. The "integration metadata" concept has been tested in Comext. There is currently an interface which enables data from Eurostat's main reference base, NewCronos, to be combined with external trade data using the principle proposed in this document.

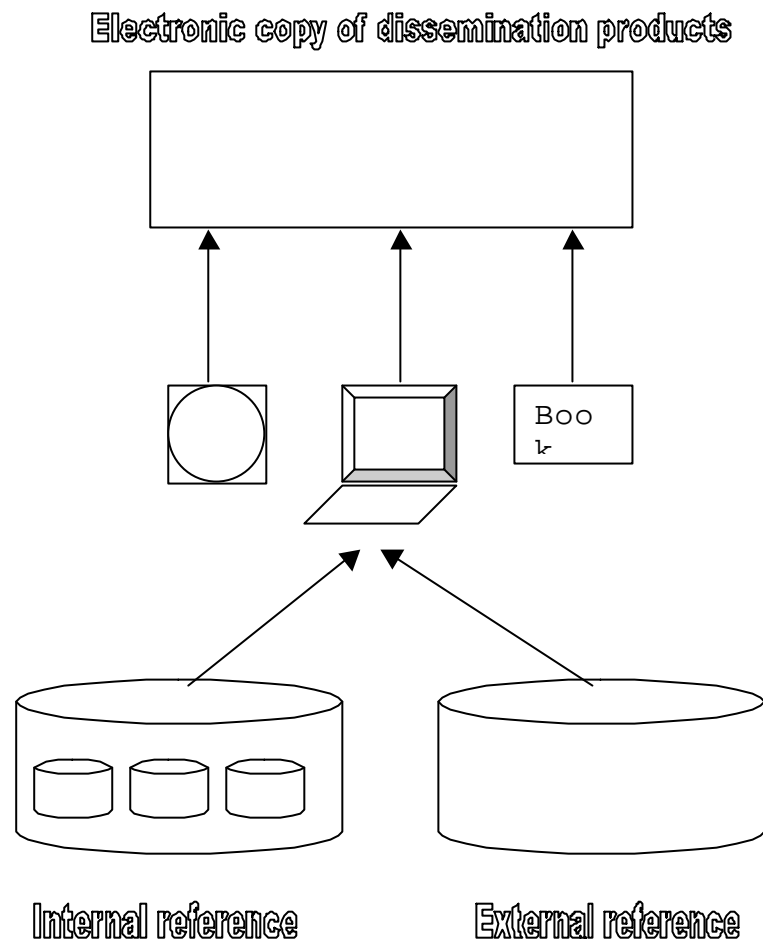
32. Application servers exist in embryo: Demetra for seasonal adjustment, a camouflage server, CIF, data preparation software (DPS) written in SAS. The broker-based architecture is going to be tested in the process of developing an estimation server.

VI.3 The reference environment



33. This environment should permit standardised data presentation, strict access control, a unique interface and single language for data extractions and storage. For the sake of access and robust overall architecture, it is proposed to make this a physically separate environment (at the price of data redundancy). It is constructed around the existing reference base, NewCronos, or its successor. In the short terms, there appears to be no need for any major change other than making the interfaces uniform (Comext/NewCronos). In time, there is a good case for abandoning the current storage format, which is proprietary and makes cross-referencing fairly laborious.

VI.4 The dissemination environment



34. Disseminated data will have to meet quality standards. Access management will be developed, with portals corresponding to different types of clients. Interactive subscription, automatic notification and personalised access facilities will be developed. Eurostat site 2 is the Internet part of this environment at present, and developments on the strength of advances in e-commerce are to be expected in coming years.

35. The major change on the current situation will arise from the need to keep a centrally stored electronic copy of all the products disseminated.

VII. THE MIGRATION STRATEGY

36. Updating architecture while meeting the demands of production and keeping within a constant budget is a risky business, and these two constraints have played their part in curtailing ambitions. The new model will be implemented incrementally, beginning with tried and tested robust modules, as already mentioned.

37. The strategy adopted takes the following approach:

- ◆ Widespread involvement of informed users in defining the project. A general guideline document describing the reasons for this project and its main features has been discussed within Eurostat.
- ◆ Involvement of management. The Management Committee has given its agreement on the thinking behind the plan and on a first analysis of the data cycle in Eurostat.
- ◆ Detailed analysis of statistical objects and the required functions. The different types of data and metadata handled in Eurostat have been thoroughly analysed in the light of work by B. Sundgren and METIS. In parallel, the operations which these data undergo during their life cycle have been catalogued.
- ◆ First rationalisation of the tools used for production. The range of software and specific applications used in Eurostat has enabled statisticians to satisfy most of their needs flexibly and promptly. The costs for the institution, however, are considerable (licences, training and maintenance). While introducing the new architecture, therefore, it has been decided to limit the number of development tools and to make the most of existing modules. This should simplify the existing production environment and facilitate migration.
- ◆ Construction of prototypes. Prototypes of certain key elements of the architecture, such as the "integration metadata" layer, the "broker" and the application servers, are currently being developed.
- ◆ Examples. In order to draw on others' experience, certain Member States' information systems and international recommendations have been studied.
- ◆ Incremental approach. The proposed architecture permits an incremental approach. Efforts will focus on the main information systems currently running under FAME, ACUMEN, Oracle/Express and SAS.

The various applications will gradually be integrated.

38. One of the main benefits of the planned solution is that it basically fits the existing set-up like a glove. This leaves plenty of room for manoeuvre in deploying the new model.

VIII. CONCLUSIONS

39. This article has looked at the discussions on the development of a new architecture for statistical information systems within Eurostat. The success of an operation on this scale will depend not only on the pertinence of the proposed model but also on the institution's ability to change. Patterns of behaviour will have to change, and that is not the least challenge.

40. In the solution now envisaged, the major concern has been to retain a balance between information technology at the service of the users, which can react swiftly to new needs and new constraints by giving users a large measure of autonomy in developing and using the tools required, and a more corporate approach which uses generic solutions as far as possible and guarantees the interoperability of local systems.