STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

Working Paper No. 29
English only

CONFERENCE OF EUROPEAN STATISTICIANS

UN/ECE Work Session on Statistical Metadata
(Washington, D.C., United States, 28-30 November 2000)

Topic (ii): Metadata modelling and terminology issues

METADATA STRUCTURES IN EUROSTAT

Submitted by Eurostat [1]

Contributed paper

I. INTRODUCTION

1. The aim of this contribution is to give concrete insights on how metadata are treated in Eurostat, in particular in the NewCronos Reference Environment. Metadata will be described from two complementary points of view: contents-oriented and tools-oriented. This will allow us to give an overview of the major types of problems we encounter, and of the solutions that are currently developed - or will be developed - in the framework of our projects (focusing mainly on the "reference" and "dissemination" units).

II. REMINDER: THE EUROSTAT PICTURE

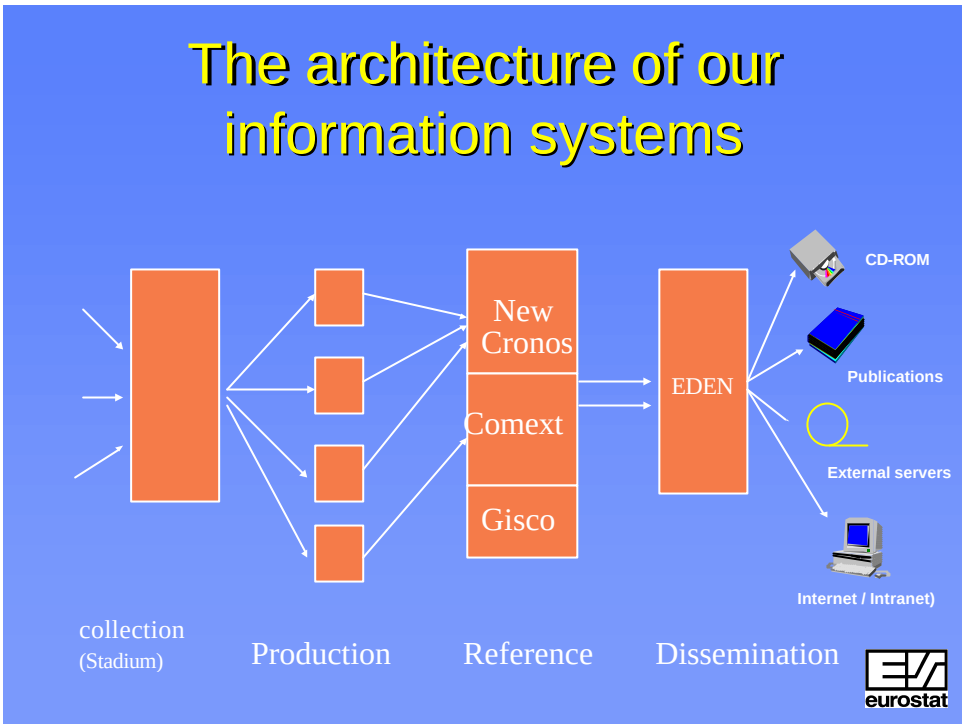
2. The Architecture of the information systems in Eurostat is structured in three layers: Production, Reference, and Dissemination.

3. Each one of these layers can have its specific metadata. The reference databases must contain all the information that can be disseminated, their metadata will be aimed mainly at the satisfaction of users' needs, inside and outside Eurostat.

4. The reference layer gets its metadata from production, but there is also specific creation of metadata in order to add value (quality, harmonisation) to the disseminated information, taking into account the users' types.

---

[1] Prepared by Dominique Groenez, Stefano Paganoni and Bart De Norre.



5. The dissemination layer can be seen as a "superlative" form of the reference layer, satisfying strict quality criteria (like timeliness, reliability, harmonisation, documentation, etc) and remodelling the data, offering additional facilities, covering different platforms,…

6. We will focus now on the reference layer. It comprises three databases: NewCronos, Comext (External Trade) and GISCO (Geographical Information System). Our main concern will be NewCronos' metadata, which is the responsibility of unit A3 (Reference Databases). Note that Comext and GISCO may have specific metadata structures.

7. NewCronos is a huge set of statistical tables, structured following a hierarchical classification (themes, domains, collections, etc.).

III. METADATA IN NEWCRONOS

8. For this presentation, we have adopted two descriptive frameworks for metadata that will help highlighting the different kinds of problems we are faced with:

9. Contents (type/use of metadata)

- Descriptive (basic titles/dimensions of a table)
- Semantic (explanatory texts on the data: methodology, quality etc.)
- Administrative (not directly linked to the data: contact persons, help files, etc.)
- Selective (for example menus or keywords used for data retrieval)

10. Management tools: the other classification used is based on the tools that are needed for efficient metadata management. They constitute the building blocks of an (ideal) integrated metadata management environment. For each of these systems, we describe the present situation and the ongoing developments.

- Nomenclatures management system
- Texts management system
- Thesaurus management and indexation procedures

These classifications have both a pragmatic basis; they are based on our experience as documentalists in the daily management of numerous metadata. Other classifications can of course be used, relying on other approaches.

11. The challenge: Since data and metadata are the two sides of the same coin, and given the high update rate in NewCronos, the Big Challenge for us is to keep and ensure the coherency between data and metadata. This is furthermore made difficult because of the multilingual context that is specific to the European institutions: we have to put metadata at disposal in three languages (German, English and French) and the updates have to be made in parallel. (mostly, not for all texts)

IV. INVENTORY OF OUR METADATA: THE CONTENTS' POINT OF VIEW

12. A complete list of these elements is given in annex.

IV.1 Basic descriptive metadata

13. The tables represent the fundamental data objects in NewCronos and are the leaves of a hierarchy that consists of the following levels: theme, domain, collection, group and subject. The two last levels are not mandatory.

14. A statistical table in NewCronos is a set of statistical observations structured in a multidimensional table (the cube model). This means that an observation is described by a value (a position) for each of the composing dimensions (criteria, indicators). Therefore, for understanding the observation and his value we need to know the phenomenon the table is going about and the definition of the dimensions and their positions. The dimensions are fundamental metadata for a statistical table.

15. Elementary metadata are:
- title of the tables and their levels (the hierarchy)
- dimensions (dictionaries) which define the structure of the cube :
- title of the dimension
- for each position : labels (in 3 languages)

16. In fact, many dictionaries can be compared to nomenclatures - be they official or not. They should be as much as possible derived from the official nomenclatures in the production environment.

17. Important derived metadata for our clients are the classification plans: a standardised and structured description of the various hierarchical levels of the database like domains and tables. (these are available in pdf and html.)

18. Other types of "basic" metadata are:
- Flags (standard list)
- Special values (standard list)

In practice, these are managed as data.

19. Identified Problems
- Lack of harmonisation, redundancy in concepts: the same concept is expressed in different ways, which leads to ambiguities and difficulties in searching. Moreover nomenclatures "live" (version management). And again, the fact that we manage three linguistic versions makes things more difficult. A conceptual harmonisation is needed in order to overcome this problem.

20. Ongoing developments
- Codes harmonisation is a first step to reduce the number of dictionaries; but it is not sufficient. A conceptual harmonisation is needed, which in turn leads to the definition of a standard descriptive

framework for the dimensions and their corresponding dictionaries. This allows the user to build interactively any table he wants, combining various indicators.

- Implementation of hierarchical views and structures within the dimensions of a cube in order to make the presentation more user-friendly for viewing and searching.

IV.2 Semantic metadata (explanatory texts)

21. The fundamental metadata can be too difficult or complex to understand or too general for all interpretations. That is why supplementary metadata in the form of explanatory texts and footnotes are so useful for explaining the meaning and relevance of the observations. We call these metadata "semantic metadata". In this class we distinguish three subclasses: Basic definitions, pure methodological texts, and texts that help interpreting correctly the data (which has to do with quality). They are in the form of free text, html or pdf formatted, and can be attached at each level of the hierarchy. They are also in three languages, and are currently managed in an ad-hoc way

22. The quality report, an important initiative in Eurostat, should be an output of these texts. (Internal quality report - short form -pilot phase Eurostat)

23. Databases of national data collection methodologies in the Member States are an important element of metadata (for example in business statistics, social statistics, …) and can play an important part in quality assessment.

24. Identified problems

- Many problems in the definition, the modelling, the management, the presentation and the consultation of texts are identified : lack of a structured datamodel, difficult incoherencies detection, missing workflow management, unsatisfying linking mechanism to the data, missing standardised presentation formats, poor integration in the consultation process, etc.
- Quality indicators are missing

25. Ongoing developments

- The contents, structure and length of the texts are very diverse. From the qualitative point of view, in order to ensure a correct understanding of the data, it is important that a minimum set of mandatory explanations is put at disposal for each domain, in a common format. This kind of standardisation is realised through small projects: SDDS format for Euro-SICS, Euro-indicators (a common format and common elements)
- Introducing a more detailed classification and a more formalised description in the texts (XML technology, RDF-schema, Dublin core)
- In fact we started a project "textserver" (see infra)
- Eurostat has also introduced quality assessment for its data, via the development of a "quality report" comprising a number of standard texts. This report will be implemented and systematically put at disposal as an output of the text server.

IV.3 Administrative metadata

26. These are not related directly to the data. Example: price, update calendar, contact persons, Help files…

27. Identified Problems

- A growing demand for update calendar information and well described releases of updates.

28. Ongoing developments

- The EDEN (Eurostat's Dissemination Environment) project foresees an update calendar.

IV.4 Selective metadata

29. There exist for any statistical information system many possibilities for information retrieval (data and metadata) that can be classified into two broad classes:

- Hierarchical search
- Horizontal search: via keywords

30. Hierarchical search is a very simple way to retrieve information, but it has the defaults of any classificatory approach: it does not provide a horizontal, cross-domain view of the information available for a certain topic (because it may appear in different themes, domains etc.)

31. Identified Problems

- In order to improve it, we have a basic index search based on automatic keyword extraction. This system presents several defaults and is presently being improved through the development of a thesaurus-based system, for consultation, indexation and retrieval (see THESEUS).
- The descriptive harmonisation (see above) will of course also contribute to the user-friendliness of the access.

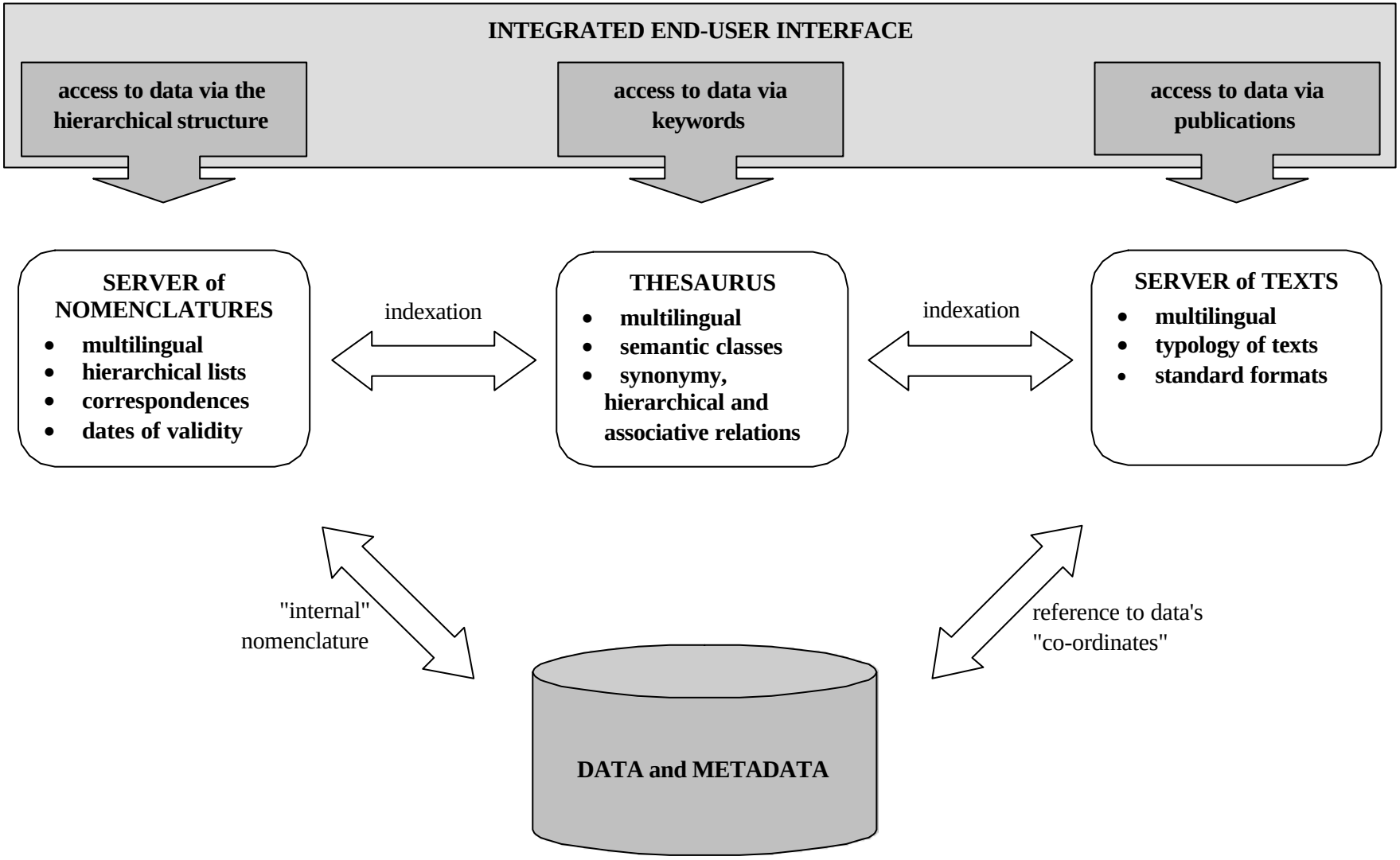
32. Ongoing developments

- See mainly the search facilities in the chapter below

V. METADATA : THE MANAGEMENT TOOLS

33. See the schema on the next page.

METADATA MANAGEMENT : a conceptual point of view



V.1 Nomenclatures management system

34. NewCronos dictionaries are often a mix of official and non-official (ad-hoc) classifications (or nomenclatures). The adopted point of view is again a pragmatic one: we think that all our dictionaries could be managed by a nomenclature management system. Indeed, non-official classifications require the same management functions as the official ones. Both of them are structured (code-label), multilingual and are constantly modified over time, which necessitates version management.

35. A nomenclatures server, offering these functions, would thus help us in the daily management of these dictionaries and would at the same time provide the whole of Eurostat with a sophisticated repository of classifications that could be reused by the domain managers, which would improve standardisation.

36. Ongoing developments

- RAMON, a server containing most international classifications, is running. Further integration with the production and reference environments will be developped.

37. Modelising the different types of dictionaries towards a more user-oriented dissemination : classifying, harmonizing, etc (see also the EDEN project)

V.2 Texts management system

38. The texts are currently managed in an ad-hoc way, which creates many problems, e.g.:

- Incoherency with the data
- Inconsistencies between the three linguistic versions or between similar contents;
- Difficulties to establish coverage, freshness and update statistics
- Diversity of presentation formats, which makes maintenance complex (lack of standardisation).

39. The functionalities that would be required from a text management system might therefore be the following:

- A structured datamodel for texts
- A "workflow" management of texts (from production to dissemination)
- Retrieval the texts on the basis of different criteria: last update date, texts related to a given topic, number of texts per domain, types of texts per domain....Different fields, including searchable fields, have therefore to be defined and used for the management of texts;
- The availability of a common presentation format (template) would also facilitate texts management and ease of understanding; this template could be complemented with editing facilities such as hyperlink update and control;
- Coherency checks (parallel update in the three languages for example)
- Link to other tools containing textual information, such as the CODED database (Eurostat Concepts and Definitions Database).

40. Presently, we do not have such a tool at disposal, but a problem statement for such a text server is defined.

41. As to footnotes, they are a special case. We differentiate some types. Most of these will be attached to the most "atomic" level, a cell in a table, and require therefore the use of a special tool. The management of the multiple relationships and the coherency control are the most difficult aspects. (We developed a software called FOOTMAN to manage in adhoc way the footnotes and their relationships within in the NewCronos cubes.)

V.3 Thesaurus management sys tem

42. THESEUS as a thesaurus management system: A thesaurus contains the terminology of our statistical world presented in a user-friendly, natural way. It is intended to facilitate the access to the information and should be close to the users' language. Terms are inventoried and grouped in semantic ("natural" hierarchical) classes. Other relations such as synonymy, hierarchical and associative relations help guiding the user to formulate his request. THESEUS is more oriented to the end user and less to production.

43. THESEUS as an indexation system : To retrieve the information, we have to index the contents of the database (with a priority on the basic descriptive metadata) and develop a user-friendly interface.
The indexation is carried out via an automatic indexation algorithm, with human post-treatment.
One advantage of such a system is that it allows checking the terminology and the linguistic equivalencies, and thereby it ensures that we share the same language for describing statistics.
The final aim is to put a sophisticated thesaurus-assisted retrieval system at the user's disposal (on-line consultation of THESEUS and selection of the relevant keywords, exploitation of the semantic relations in the formulation of the query…) but we will start more modestly by improving our present index.
This multilingual thesaurus management system is now operational and will be soon put at disposal on the intra- and the internet.

V.4 The links between these systems (not yet operational)

44. In the future, metadata should be imported from their specific management systems into NewCronos. The "real" dictionaries (i.e. describing available data) would be subsets of the lists contained in the nomenclature server. The situation would be similar for the texts. The updates of metadata would also take place directly in these specific environments.

45. As to the indexation procedure, it is at present carried out directly on the NewCronos dictionaries. So it has to be updated each time the data (and associated dictionaries) change.
In the futur we plan to index also a number of semantic metadata (well structured explanatory texts)

VI. CONCLUSION

46. We have concentrated on internal metadata management, but we cannot forget that all this information is useless if it cannot be properly accessed, displayed and exported: it requires the availability of user-friendly interfaces:

- For navigation: menus for hierarchical search;
- keyword based search assisted by a thesaurus (free search, search via prefabricated indexes, guidance, exploiting the semantic structures and relationships);
- A browser for displaying and manipulating the selected multidimensional tables (showing also the footnotes); this specific cube-browser is called EVA (Eurostat Visual Application) and is now available.
- Export functions
- Even more fundamental to a user-oriented approach is the datamodel we will develop in the EDEN project.

47. Eurostat started a global strategic project, based on a workflow approach, aimed at obtaining the complete list of metadata that are necessary for Eurostat and the ESS, at each stage of the statistical lifecycle. This project should involve all the concerned partners; therefore any feedback/advice from National Statistical Institutes would be welcome.

ANNEX 1
List of the various types of metadata (draft)

1. Basic descriptive metadata

• Identification: basic description, "name" of the statistical observation, which relies on the use of official classifications (or not), and of dimensions (attributes) that are standardised or not.
• Type: structured text: label-code
• Attributes: reference (e.g. nomenclature NACE 70); type (level: descriptive or structure dimension; list); language; stage (production, reference, dissemination); medium (electronic/paper …)

= Syntax/vocabulary: basic description

LIST
• Name of the table or of the series (hierarchical structure)
• Descriptive dimensions (table level)
standard descriptive framework (syntax): declaring/ indicator/ breakdown/ elaboration (method + unit)/ periodicity/ year + period/ partner
Vocabulary: official nomenclatures
Vocabulary: non-official nomenclatures
• Have a more or less high harmonisation level.

2. Semantic metadata
They consist of explanatory notes of various types:
• Definitions
• Methodology
• Aspects connected with the interpretation/ reuse/ quality (see Quality report): for example circumstances of the survey, non-response rates...
• Footnotes (describing exceptional or specific aspects/ differences)

One can also distinguish various levels:
• Metadata harmonised at the international level
• Detailed metadata comprising information on the national characteristics.

LIST:
• Definition of the observation (indicator and connected concepts): harmonised definition; national definitions
• General Considerations:
Justification: legal basis (OJ etc), strategic/political usefulness (see relevance, items 1.2 and 1.3)
• Methodology
Official conditions for the production of statistics (sources, treatment of confidentiality, national access to the data, official national comments) see IMF
"Nomenclatures" used
Status: harmonised, non-harmonised, official, temporary, non-official
Legal basis, poss.
Design method used
Validity (beginning-end)
Population, geographical coverage, poss. sampling procedure
Units
Periodicity
Calculation methods (indexes etc.)

• Interpretation (closely connected with quality - and with the Quality report)

Geographical comparability : exceptions by country (e.g. definition of an SME); estimates for areas, see Quality report - 5.2

Comparability: with other sources (for "cross-checks" a.o) see Coherence, Quality Report (e.g.: press releases, external bases etc)

Comparability: in time (incomplete coverage, break in series, change of method and impact etc..

Timeliness, punctuality (delays, their causes)

Precision of calculations (Quality report): sampling errors, frame errors, non-response errors, measurement errors, processing errors, model assumptions, overall accuracy.

Comments: national/EUROSTAT/newspaper cuttings

Precautions for re-use

3. Administrative metadata

This involves metadata connected with the use (re-use) of the data. They do not concern strictly speaking the data.
Example: copyright, help files, bibliography, contact person...

LIST

Contacts (nationals or Eurostat) for feedback/help see IMF

Dissemination formats and periodicity (IMF, Quality report 4.1.1); size, prices (Eurostat)

Restrictions on use; privileges (governments etc)

Treatment rules

Cataloguing information (e.g. for a publication)

Information on software

FAQ

Context-sensitive help

Timetable of the updates; important foreseen updates (changes of methodology etc), important ongoing work.

Last updates

Other databases interesting Web sites

Information of the same type available in other DGs, inter-DGs co-operation...

Bibliography and more general information sources (e.g. Eurostat as a whole)

Symbols and abbreviations

Evaluation (see Quality report): for internal needs.

Relevance: users 'profiles/surveys (quotations, use)

Clarity/accessibility

Specific metadata for management, which ensure a.o. the follow-up of the data and metadata teams work (XML Tags for the texts; filing of work related to the loading/update of a domain; log file, file transcodification, virtual tables, refer table etc.)

Metadata used in data retrieval:

Derived from the previous metadata

Plus additional data

It allows several basic types of search/selection :

"Horizontal" access (key word indexes, thesaurus of descriptors)

Hierarchical selection by browsing the hierarchical structure of the data

Specific menus (profiles "a la carte")

LIST

Hierarchical menus (using the structure of the DB).

Hierarchical views on dictionaries.

A multilingual thesaurus of descriptors, which guarantees the terminological consistency of the database and ensures the precision of the search while allowing major flexibility.

User's queries definitions.

ANNEX 2
THE EDEN PROJECT : SOME EXCERPTS :

DATA PRESENTATION MODEL & GUIDED ACCESS: INTRODUCTION

1. Let us imagine that Eurostat's data were perfect (well documented, relevant, accurate, up-to-date, comparable). Let us imagine too that the database's end-user interface was also perfect (functional, easy-to learn, easy-to-use). In that case, would Eurostat dissemination be perfect or would something still be missing? According to our analysis, users would still miss a functionality that may be called the choice À la carte.

2. The prevalent kind of dissemination databases available on Internet or CD-ROM gives users the choice among a (short or long) list of pre-defined tables. Each of these tables includes a subset of indicators, breakdowns (e.g. by branch, by sex, by age), countries and dates that are available in the production database of the database's authors. The selection is made by the authors themselves who try to guess what the main demands of the users may be and then design the tables accordingly. This approach can be called the Suggestions du chef. The CD-ROM Panorama of Industry is an example of this kind of dissemination database. The structure of dissemination databases of this kind is very similar to the structure of paper publications. Their main advantages are the hyperlinks and the fact that the user can easily extract big amounts of data and import them into other software.

3. A different sort of dissemination database is one which allows the users to make selections in a big (virtual) table that includes all the statistical data available for a specific domain. COMEXT is an example of this kind of database, where the user can define the content of the subset he wants to extract. He selects the indicators (e.g. import), the breakdowns (e.g. products), the countries and the dates he is interested in, and then asks the server to build the table on the fly for him. This approach can be called À la carte.

[New Cronos (which was not conceived as a dissemination database) is a hybrid. Users can find there both ready-to-use tables and big tables from which they can extract sub-tables.]

4. The main difference between these two kinds of dissemination databases is not only the size of the table (one huge virtual table instead of several small real tables). There are two other topics that, in our jargon, are called cross-domain access and sparse array management:

- An À la carte dissemination database (unless it is restricted to one single domain) has to provide users with cross-domain access, i.e. the possibility of defining a table that is made up of data coming from more than one statistical domain. Example of a cross-domain access: data on production, import/export, transport and employment for a specific branch/product. Cross-domain accesses are possible only if the dissemination database adopts a coherent Data Presentation Model, i.e. a unique set of harmonised codifications and a standard data description. The purpose of such a model is not to force an overall harmonisation of data belonging to different domains, but rather to put that variety of data in a coherent framework in order to present them in a proper way to the users.

- It is evident that not all combinations of indicators, breakdowns, countries and dates are actually available in À la carte dissemination databases. Example: the breakdown by size of an enterprise may exist for production and employment, but not for import/export, which can be broken down by partner countries. Employment is available by sex and age (but the other indicators aren't!) and transport may be the only indicator available at regional level. How then can non-specialist users build a proper table from such a sparse array? They need the Guided Access facility that leads them through the process of selecting and coupling the correct dimensions and the correct dimension's occurrences.