STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

COMMISSION OF THE EUROPEAN COMMUNITIES

EUROSTAT

Joint UNECE/Eurostat Work Session on Statistical Metadata
(6 - 8 March 2002, Luxembourg)

Working Paper No. 2
English only

Topic (i): Infrastructure issues for statistical metadata

THESEUS

A MULTILINGUAL THESAURUS FOR ACCESSING EUROSTAT'S REFERENCE DATABASES

Submitted by Eurostat[1]

Invited paper

ABSTRACT

Metadata have several roles to play: a descriptive, a semantic and, last but not least, a search assisting role. It is mainly this third role, aimed at helping users to find the statistical data and associated information, which gets priority when talking about a thesaurus for our reference environment: **THESEUS**.

To fully understand this thesaurus and its use, we will first highlight its place within **EUROSTAT**'s Statistical Information System and metadata architecture. Afterwards, we will explain the objectives, content and structure of **THESEUS**.

The next chapters talk about its application in relationship with our major reference database, **NEWCRONOS**. This chapter explains semi-automatic and manual indexation procedures as well as synchronisation issues.

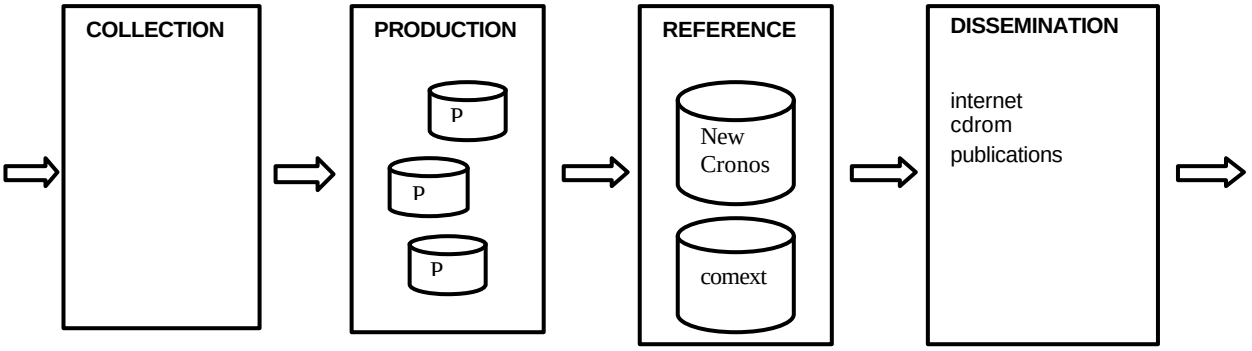
The logical continuation is the use of the **THESEUS** and the indexation results in search functions for the enduser. And coming back to our metadata architecture, we will address some future issues of the thesaurus.

[1] Prepared by Bart De Norre and Dominique Groenez

I. EUROSTAT'S STATISTICAL INFORMATION SYSTEM ARCHITECTURE

1. Let's start with a general picture of Eurostat's Statistical Information System:

Chart 1



2. In this context we will focus on our reference environment and more precisely **EUROSTAT**'s most important database for endusers: **NEWCRONOS.**
From a user's and metadata point of view **NEWCRONOS** shows several drawbacks:

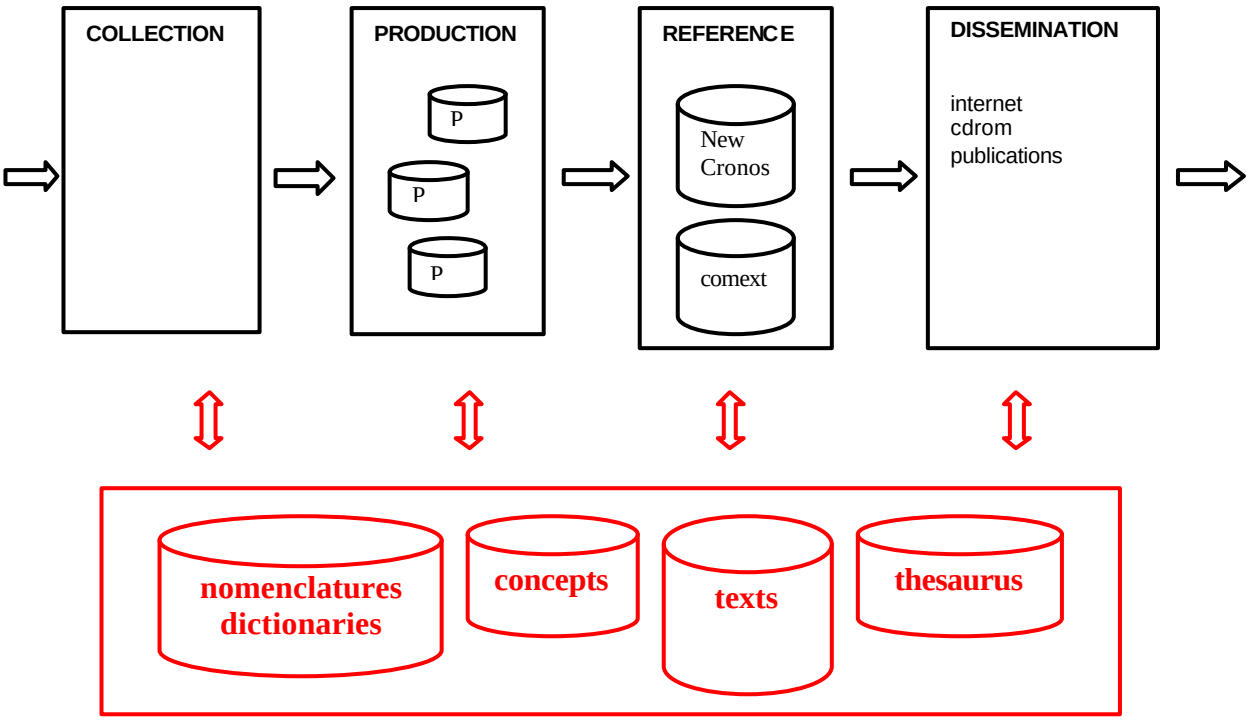
- There is an insufficient harmonisation in the terminology of its basic descriptive metadata (conceptual and linguistic harmonisation).

- **NEWCRONOS** is a huge hierarchy of statistical tables (hypercubes) with very little search functionality and difficult cross-domain comparison.
  NEWCRONOS consists of themes, domains, collections, groups, subjects and tables. The multidimensional tables - that contain the "real" statistical data - are grouped together according to several points of view: geographical (regions, candidate countries), thematic (population, industrial branches…), or strategic (euroindicators, structural indicators).
  This classification also implies that the same kind of data (e.g. employment) can be found at different places and that the user has no means to know it directly - unless he is explicitly informed about it, or if he has time to spend browsing the entire database!

- **NEWCRONOS** covers all statistical domains; there is very little assistance for tuning or orienting the user in his search process.

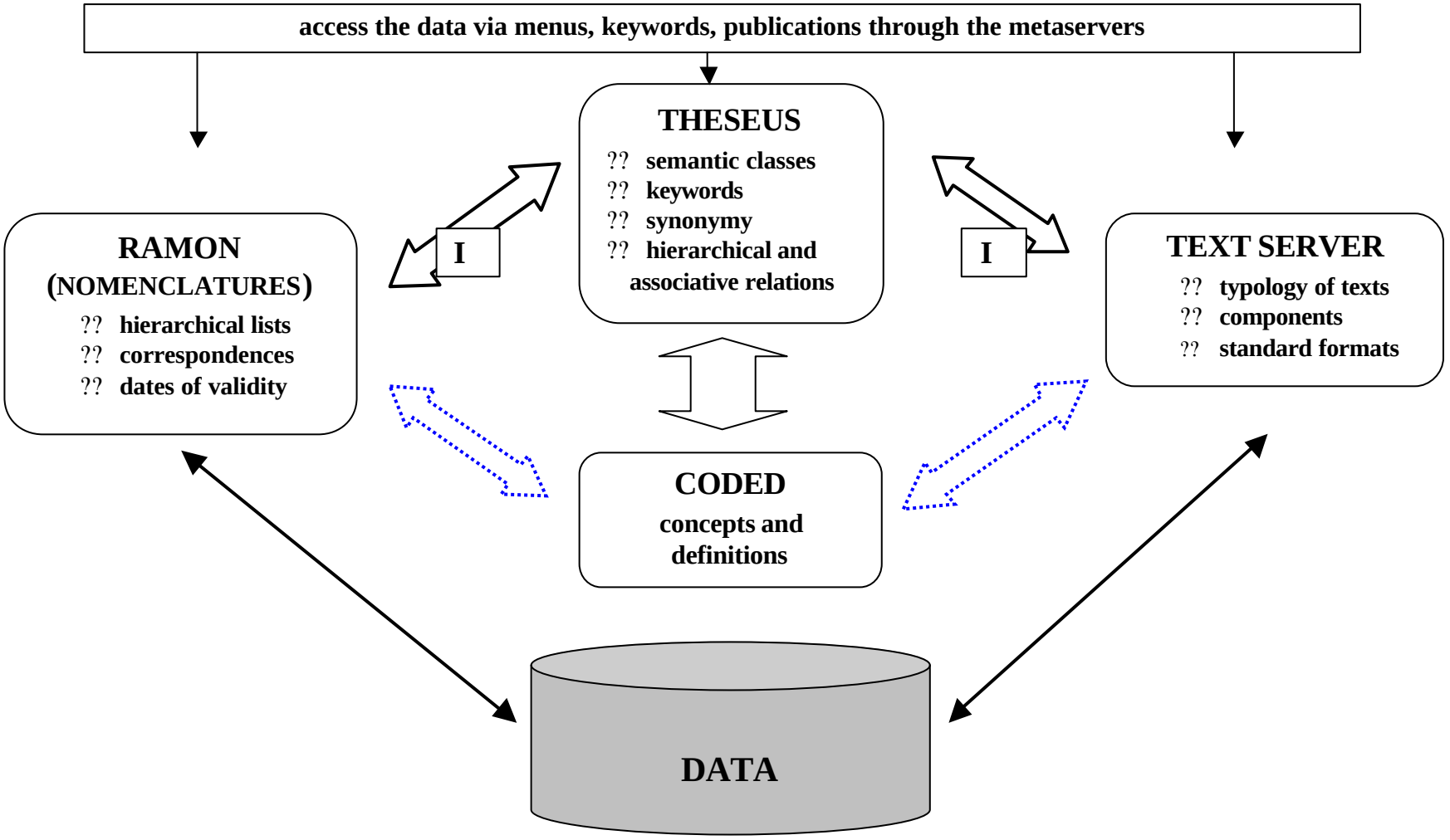
3. Our metadata architecture and policy should address all these problems. The next picture illustrates the conceptual view of a metadata server.

Chart 2



And the "metadata server" that we will put in place for the reference environment looks like:

Chart 3


access the data via menus, keywords, publications through the metaservers
THESEUS
?? semantic classes
?? keywords
?? synonymy
?? hierarchical and associative relations
I
I
RAMON
(NOMENCLATURES)
?? hierarchical lists
?? correspondences
?? dates of validity
TEXT SERVER
?? typology of texts
?? components
?? standard formats
CODED
concepts and definitions
DATA

II. THESEUS: EUROSTAT'S REFERENCE THESAURUS

2.1. THESEUS: objectives and definition

4. A thesaurus can be defined as a structured list of the expressions used, on the one hand, for representing the document's content (indexation process) and on the other hand, for searching these documents in a documentary system. In the case of **THESEUS** the documents take the form of statistical tables.

5. **THESEUS**, being the thesaurus for **EUROSTAT**'s reference environment **NEWCRONOS**, has 2 major objectives:

- Improving the information retrieval process of the end user
- Offering a multilingual tool (for the moment: English, French and German), listing in a naturally structured way, the list of the terms and their variants describing the contents of the database.

6. **THESEUS** will also contribute to a harmonisation in the reference environment, especially by its indexation procedures with the other metadata systems (dictionaries in **NEWCRONOS**, concepts in **CODED** and the nomenclatures in **RAMON**).

7. The focus here on the reference environment and the enduser implies also the use of a more "natural" language instead of a rather "technical " language of the statistician.
The thesaurus also aims at suppressing the ambiguity of natural language, and in particular synonymy and homonymy.

- Synonymy: the same concept can be expressed by different terms or expressions *(e.g.: biology - biological sciences)*
- Homonymy: the same term can have different meanings *(e.g.: stone (fruit) - stone)*

8. **THESEUS** structures the domain of knowledge in the following way:

- Semantic classes:

  At the first level, the thesaurus is structured in semantic classes that group together all the terms that are specific for a given sub-topic. Semantic classes may be hierarchically defined (which is not applied for the moment)
  Example: agricultural production, demography

- Descriptors or keywords:

  They are terms or univocal expressions designating concepts used in the field covered by the thesaurus. If there are several synonyms, only one term is chosen as a descriptor (for example, according to its frequency). The indexation of the documents (tables) is done only with descriptors.

- Non-descriptors or synonyms:

  This involves terms or expressions, which cover more or less the same concept as the corresponding descriptor. Each non-descriptor corresponds to only one descriptor. In contrast, a descriptor can have many synonyms.

- Semantic class relationship (introduced by the initials «SC»):

  This relation defines the membership of a descriptor or keyword in a semantic class. A keyword may belong to one or more semantic classes.

Example: male labour force SC employment and unemployment

?? Synonymy or equivalence relationship:

Descriptors and non-descriptors are linked in a symmetric way by the synonymy or equivalence relationship:
Example: male labour force
UF male active population
UF male worker
During search, the user can use equally the descriptor or the non-descriptor.
Example: male worker
USE male labour force

The equivalence relationship can include several types of relationships such as:

?? true synonymy,
Example: value-added tax
UF VAT

?? quasi-synonymy:
Example: livestock
UF herd

?? antonymy:
Example: irrigated area
UF non-irrigated land

?? inclusion: when a descriptor includes one or more specific concepts which are considered as non-descriptors because their use is not frequent:
Example: apple
UF cooking apple
UF desert apple

9. The **THESEUS** thesaurus introduces a distinction between "true synonyms" and "false synonyms". The latter can be the plurals of descriptors or of non-descriptors, spelling variants etc. The search on the false synonyms will be possible and will re-direct automatically towards the corresponding descriptor.
This feature is essential for the user-friendliness of the thesaurus, since it can provide the user with more "natural" expressions - for him - of a statistical concept.

?? The hierarchical relationships:

The hierarchical relationships (BT and NT) implement the hierarchical view of the domain covered.

?? Relation «BT» (broader term), between a specific descriptor and a generic descriptor.
Example: wine-growing region
BT1 agricultural region
BT2 region
The number indicates the number of hierarchical levels between the specific descriptor and each one of its generic terms.
The descriptors that do not have generic terms are also called "top terms".

?? Relation «NT» (narrower term), between a generic descriptor and a specific descriptor.
Example: region
NT1 agricultural region
NT2 less favoured area
NT2 mountain region

NT2 wine-growing region
NT1 frontier region

?? Associative relationships

The associative relation sets up a symmetric relationship between two descriptors, which invites the indexer or the user to «see also» another descriptor. Associative relationships are less strictly defined as the other ones, and allow therefore much flexibility. These associative relationships may link two descriptors across different semantic classes.
The associative relationships (marked by the initials «RT» - related term) are not yet used for the moment.
Example: calorie "RT" human nutrition in the SC Health

Polyhierarchy, a natural feature of language, is implemented in several ways:

- ?? A descriptor which may belong to more than one semantic class
- ?? A descriptor which may be a narrower term of two descriptors in the same or in different semantic classes.

10. Finally, scope notes show the indexer or the end-user how to understand and use a descriptor. Definition notes can specify the descriptor's meaning. These notes are not available for the moment.

11. **THESEUS** also foresees some standard reports. We can mention:

1. Thematic overview:

   The thematic presentation shows the hierarchical structure of the keywords (or descriptors).
   Within each semantic class, top terms (i.e. keywords not having themselves a generic descriptor) are classified alphabetically.
   Under each generic term, the descending hierarchical classification appears, going from the generic to the specific levels (preceded by the initials «NT» with the number of the hierarchical level). The lower hierarchical levels are also indicated by an indentation to right. Inside each hierarchical level, the specific descriptors are classified alphabetically.

2. Alphabetic list

   It contains the complete and structured alphabetical list of keywords and synonyms. Keywords are presented with all their semantic relationships, while synonyms have only a reference towards the corresponding descriptor.

3. Multilingual alphabetic presentation

   The keywords are listed in 3 columns: the central column is the one for the chosen language and defines the alphabetic order. The synonyms are listed for this language. The left and right columns show the corresponding labels in the 2 other languages.

12. To summarize from what precedes, we can see that the contents of a thesaurus, thanks to its concepts and relationships, constitutes a road map of a specific domain and is close to the user's "natural" or non-specialised language.

13. The thesaurus in itself is a pedagogic tool: its structure in semantic classes and the semantic relations between the terms allow to place the term in its semantic context, and to show related concepts that might also be of interest for the user.
The synonymy relations offer linguistic variants and there is a possibility of including definitions.

2.2. THESEUS: content building, multilingualism and rules

14. The **THESEUS** thesaurus is based on a census and a structuring of all the terms used for the description of the **NEWCRONOS** tables.

Some rules and standards:

- ISO 2788-1986: Guidelines for the establishment and the development of monolingual thesaurus
- ISO 5964-1985: Guidelines for the establishment and the development of multilingual thesaurus
- Descriptors and non-descriptors are in general in the singular form. The plural form is used whenever it is the common form in the language of economists and statisticians.
- Abbreviations (for example: «VAT») are generally non-descriptors and are linked to the complete form. They are however preferred to the complete form in case they refer to international organisations, for reasons of facility. The complete denominations are given as synonyms or in the definition notes.
- Initials used: English initials are used in all the versions of the thesaurus.

Sources used:

- The **NEWCRONOS** database and its metadata.
- **EUROSTAT**'s official classifications
- **EUROSTAT**'s publications
- the EUROVOC and ECLAS thesauri
- EURODICAUTOM for the translations of the concepts, and the enrichment of synonymy.

15. The multilingual constraints in the European context has created difficulties; we had to set up a multilingual thesaurus, the structures of which having to be strictly parallel, i.e. identical for the 3 languages.

16. Even in one language, as explained before, you can find for example various expressions of the same concept and homonyms.
When dealing with 3 languages in parallel, things become much more difficult. There is sometimes no one-to-one correspondence between statistical terms in different languages, which makes it difficult to find precise equivalencies.

17. As a remark, these problems can also be seen as an opportunity: it forces us to establish linguistic equivalencies between concepts, and to have a high quality terminology, useful not only for our external users, but also for the metadata producers.
So, **THESEUS** can be seen as a sort of multilingual dictionary, which is structured on the basis of semantic criteria, and contains equivalencies between languages, while taking into account the nuances of each of them.

18. A thesaurus is a living tool: it has to evolve with the data it indexes. Certain concepts appear while others disappear and others change.
To stick as much as possible to this development, a team of "documentalists" examine systematically all these changes within the reference database and will reflect them, if necessary, in the thesaurus.

2.3. THESEUS as an information system

19. **THESEUS** is set up as a relational database with a client-server application to manage the contents and with a web interface for consultation. This web interface is for the moment only available at our intranet.

20. The database is implemented in Oracle. The management application is developed with Powerbuilder and the web interface is made with Coldfusion and Javascript.
The web interface offers equally reports in pdf format generated by Business Objects.

21. Some numbers (approximately):

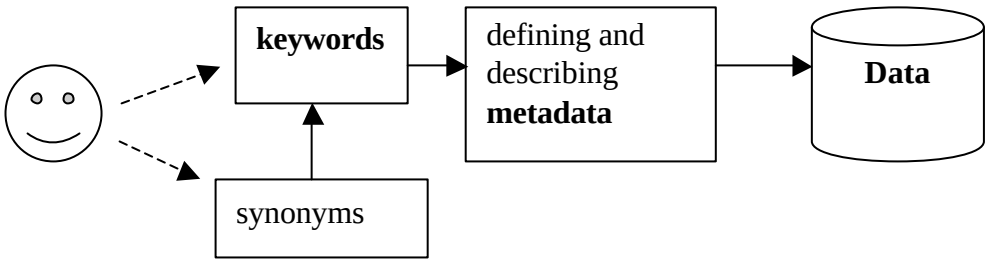
semantic classes	70
keywords	3800
synonyms English	8000
synonyms French	7900
synonyms German	8300
SC relationships	3820
BT/NT relationships	3500

III. THESEUS INDEXATION PROCEDURES:

3.1. What

22. **THESEUS** is not meant to be a "stand alone" system. **THESEUS** should be linked to the metadata, which define and describe the statistical data and their structure in our reference database **NEWCRONOS**. This process is called indexation and it is the basis for implementing the search facilities.

Chart



Remark: this chart illustrates 3 linking mechanisms: the first one is the indexation between **THESEUS** its keywords and the metadata, the second link is the use of the metadata in the data and the third one is the definition of synonyms for keywords.

3.2. Approaches

23. We considered, roughly speaking, three possible approaches: the complete automatic [2], the semi-automatic and the manual method.

[2] A fully automatic indexation, in our opinion, is not compatible with a high quality indexation. It should always require manual validation, e.g. for inclusion of implicit concepts.

24. We analysed and developed a semi-automatic indexation procedure. This is a procedure in two steps:

- The first one is an automatic indexation program, which compares the metadata texts with the **THESEUS** keywords. The program creates 2 main outputs:
  - candidate index links
  - words found in the metadata but not in **THESEUS**: this output may improve the quality of the metadata or of **THESEUS**.

  The program uses concepts like filtering via an "anti-dictionary", character transformations, recognition of multiple word expressions (keywords and synonyms), filtering non-significant words.
- The second step is the interpretation by the "documentalists" and the appropriate actions: accept or ignore the candidate links, update **THESEUS**, update the metadata, tune the parameters for the program.

25. But the difficulties encountered during an extensive testing prevented us to adopt that possibility: the generated index was too imprecise[3]. For example it did not take into account implicit concepts. And the procedure created a heavy workload during the indexation validation, especially because we had to start from scratch.

26. Surprisingly enough, the manual indexation procedure proved to be much more efficient, especially for creating a first "basic indexation".

27. The metadata involved in the indexation procedure are:

1. The dictionaries defining the **NEWCRONOS** structure (themes, domains, collections, groups, subjects) and tables (titles).
2. The describing dictionaries of the tables, which we consider "interesting". From the point of view of a user searching data via keywords. Examples: a dictionary like "indicators" is very interesting, a dictionary like "age classes" much less and a dictionary like the NACE is partly interesting.
   Deciding to include a dictionary in the process depends on a costs/benefit consideration and on its role in the future search functions which will include a certain stepwise assistance (example: searching on indicators and proposing a refinement via a major "breakdown" dimension like the NACE).

28. In addition, this indexation procedure provided useful feedback on the quality of the metadata, by allowing to detect inter-linguistic inconsistencies or imprecision. These results are of course exploited for improving the quality of the metadata.

3.3. Synchronisation issues

29. A major problem in the indexation procedure is keeping the index coherent in respect with the daily evolution of **NEWCRONOS** and of **THESEUS**.

30. Therefore, a logging and reporting procedure of the changes in the metadata needs to inform the "documentalists" about the different changes. Depending on the type of changes the "documentalists" can whether apply a manual indexation whether the semi-automatic indexation procedure which is here applied in a much more limited scope. (Tuning and overload diminish, once a complete indexation has been done).

[3] The concepts of "noise" and "silence": too many or too few references found

IV. THE USE OF THESEUS IN THE SEARCH PROCESS

31. Different facilities based on the indexation and on **THESEUS** 's content and structure, are or will be offered:

- Searching via the index: the user enters an expression (completely or partially) or the user chooses in an alphabetic index.
- Selection of keywords or synonyms via typical thesaurus presentations: a thematic list by semantic class, or a permuted index (tracking the word in keyword expressions).
- Combining the selected keywords with Boolean operators.
- Feedback to tune or refine the search by offering context information derived from the thesaurus (semantic classes, hierarchical relations, scope notes)

32. These facilities should be available as modules, which we can plug in different applications, and on different platforms. (A strongly parameterised module, which can be, built in the existing interface of **NEWCRONOS** and in the future dissemination applications (web technology and java).

V. FUTURE

33. When we look at the chart describing the metadata architecture, some future developments become obvious.

34. As explained in the previous chapter the immediate future priority is given to functions for assisting the enduser in searching the data and metadata with the help of **THESEUS** and its indexation results.

35. A second development concerns **CODED. CODED** is an information system containing definitions of concepts. It is conceived as a relational database. We will develop a synchronisation procedure between **CODED** and **THESEUS**. The synchronisation procedure focuses on keeping the content between both systems parallel and consistent.
THESEUS and **CODED** can be seen as integration tools at the core of **EUROSTAT**'s Reference/metadata Architecture.

36. Another important evolution in this context is **RAMON, EUROSTAT**'s reference nomenclature server. The goal is that all dictionaries, code lists and nomenclatures will be managed in this system. We foresee an indexation procedure between **THESEUS** and **RAMON.**

37. **NEWCRONOS** also contains quite a lot of explanatory texts. We are now analysing the functionalities and design of a **text server**. The textual metadata will be much more structured using a standard typology, taking advantage of the XML technology and considering some de facto standards like SDDS. The link with **THESEUS** is again an indexation procedure: the user may search on keywords to find appropriate explanatory texts.

38. A last point concerns our future dissemination application (the **EDEN**[4] project family): a major issue in this project is a new Standard Data Presentation Model. Future development again concerns indexation procedures.

[4] EDEN: EUROSTAT Dissemination Environment