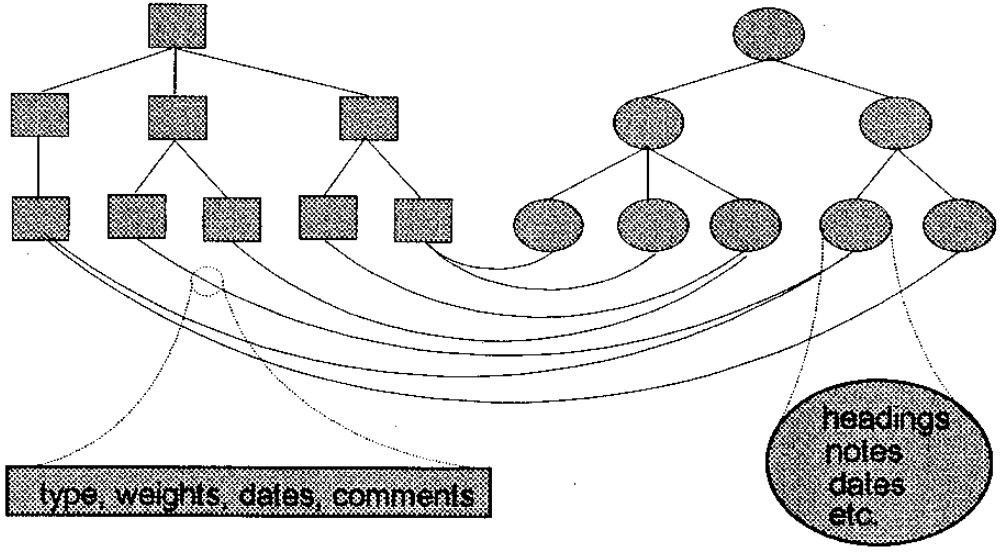
Synapse

an all uses server for classifications



INSEE 18, Boulevard Adolphe Pinard 75675 PARIS cedex 14 (France)

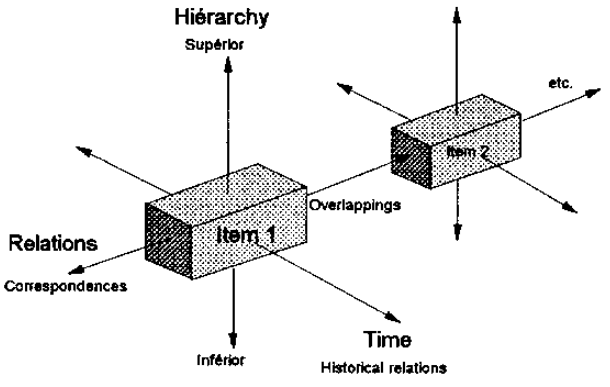
General outlook

Our more and more computerized and normative world needs to classify and locate objects of all sorts, generating a multiplication of classifications, codes and lists in every domains. They respond to varied concerns (statistical, managing, administrative or other). Very often, a classification (i.e. a list or a code) is defined in relation to another classification or must be linked to another one for needs of codification, comparisons, etc. At last classifications "live", develop ; they must be managed and dessiminated of the most efficient and user friendly way as possible.

An organism producing, managing and disseminating classifications, INSEE has developed a modular tool to ease its management, coordination and dissemination of classifications, which, innovates in four levels :

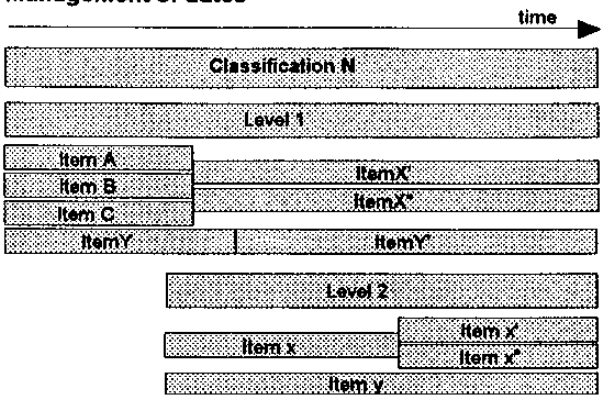
1 - In the data base, classifications are organized in a network. This allows, from a given item, to access all the items that are linked to it : hierarchically by correspondence or overlapping or else historically (previous or successive items). Then by transitivity to sweep the base from a single entry point if necessary.

Navigation in the base



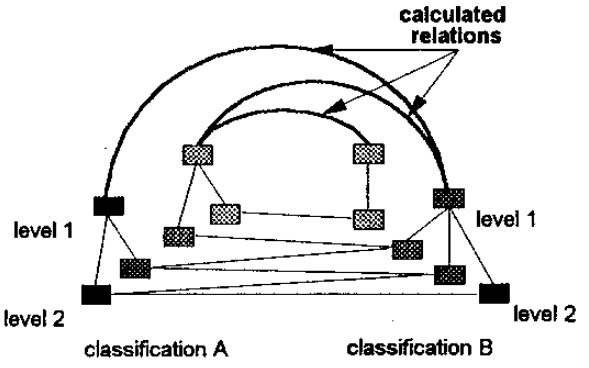
2 - As classifications develop, in a partial way and without a fixed frequency, every entity in the base is dated and each element (classifications, levels, items, tables, relations, headings) has its own validity period. Thus, for a query, a print or a file production, the choice of some reference date selects only the valid data at this date. Such an arrangement authorizes historical queries and limits the dimensions of the data base, thus improving the performances of the system.

Management of dates



3 - Relations between classification items can be weighted (from source and target). For the system calculated tables (combination of several other tables), weights are calculated taking into account the weights of elementary tables. For lack of explicit weights, the system calculates pseudo weights as a function of the number of relations kept by each item (weight is equal to 1/n if the number of relations is n).

Automatically calculated tables



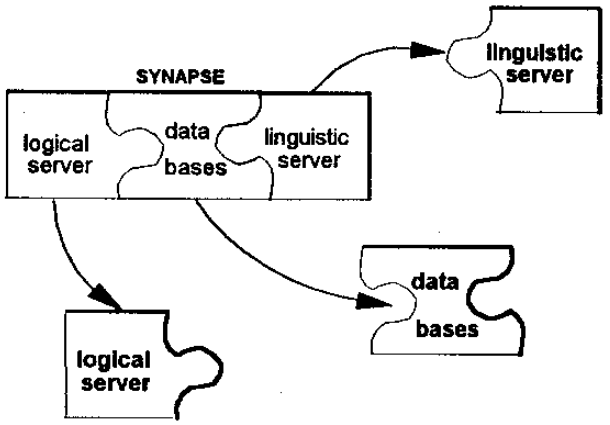
4 - For the family of activities and products classifications, a linguistic analysis module allows the access to informations through natural language queries. Not only it eases the query (by an informed or not public), this tool allows mass automatic codifications (censuses, large surveys) without a previous normalization of the language. It may yet help to the construction of correspondence tables between classifications (i.e. lists, codes).

A modular architecture

SYNAPSE has been designed to answer varied needs : management and query, extraction and printing, user friendly search and capacity to take into account the evolution of language.

Two main parts structure the classifications server:
- a logical server which concerns the formal structure of the classifications and the relation tables including extraction and management functions of these entities,
- a linguistic server which concerns the semantic structure of the activities and products classifications including search, codification and management functions of the linguistic data.

A modular architecture



SYNAPSE makes a whole but every part (logical server and linguistic server) is an application in itself which "works" on its own answering specific needs. Data bases are also products to be integrated in other applications (for more details, see the item "Derived products from SYNAPSE").

Complete functions

The very object of the development of the classification server was to build a tool designed for management (1), coordination and dissemination of any set of classifications, and associated relation tables. It was thus needed to answer any types of needs whatever be the operator. The proposed functions are thus adapted to different potentially concerned publics (classification specialists, statisticians or ordinary users) knowing well, partially or not, the classifications.

On-line documentation

Every screens are documented by :
- a quote of the environment already selectioned and the reference date,
- an on-line help (what is the use of the panel and how to use it ; concept, definitions, precisions),
- an information on the last selectioned entity (short description, statistics or rules concerning it).

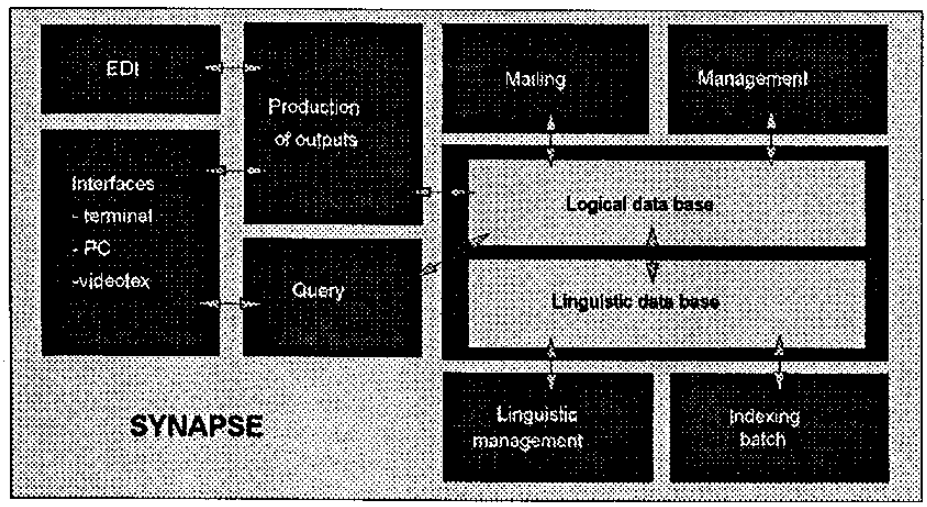
Reference date :

The choice of a reference date is the key element of a query or an output production : every entity of the base has its own period of validity. The choice of a date (by default today's date) selects only entities valid for this date. If by error, omission or ignorance, the chosen date is exterior to the validity period of the selected classification or table, the system brings automatically the nearest possible date to the reference date (beginning or end of validity).

Query

The aim of the query is the quick and exact access to desired informations and their environment.

General organisation of SYNAPSE



(1) Important remark : "management" does not mean "construction". Construction of classifications and tables is the object of an independent but complementary project of SYNAPSE : MIN (Informatic handling of classifications) will be the toolbox of the classification specialist to help him construct in a consistent and controlled way, under constraint or not, classifications and relation tables.

- the operator may choose the classification and level, to query among all the classifications (and their levels) present in the base.
- to access the items, three solutions are offered : he knows the classifications well : he can access the items by theirs codes,

he knows less well or is looking for a set of items : he can define a list of items (by exact codes or using an abridged formulation with the use of a joker),

he does not know it or he wants to verify the way an item is classified : he types a descriptive text in natural language.
(for more details see the item "Linguistic server").
- he can access all the informations interesting a chosen classification item : history of the item, explanatory notes, different headings, historical relations, other relations.
- if the operator is interested in the relation tables concerning the chosen level he can :

access all the existing tables in the system,

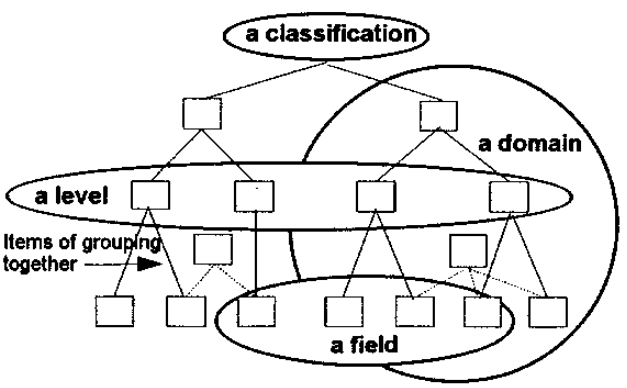
select a subset of relations by defining a field in the source level and/or in the target level,

define a new table which will be constructed by the system.

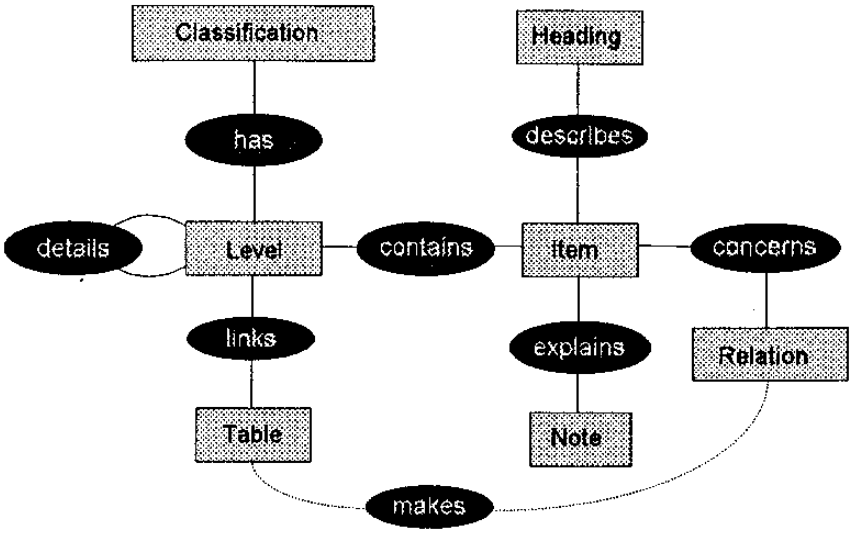
- from a table, the operator can access the elementary relations which make the table then, either the source items (in the source level) or the target items (in the target level).

Thus from tables to relations, from relations to items, from items to explanatory notes, etc, every base entity more or less linked to the interest centers of an operator can be accessed in the same query session.

Define useful subsets



The simplified data model



Outputs production

An operator may wish to recover (on paper or under file form) an information set (visualized or not in a query). He may thus select exactly the information he wants to print or save on file. This avoids too important prints (much paper for little useful information) or tedious files that are to be worked again.

It also allows to answer immediately, on a fitted medium, every request of partial informations on a classification or a table without being compelled to print, give or sell a whole document.

The variables that may be selected are the following :
- for a classification : the levels (one or several), the field (the code zone, possibly disconnected), the variables (sundry headings, notes, validity periods, etc).
- for a table : the headings of the items, in relation the weight (source and/or target), relations validity periods, possible comments.

The producted files are automatically named and archived in the working base. The prints, structured by a SGML layer are today adapted to the IBM architecture and require a POSTCRIPT printer. For other architectures, the printing format must be revised and adapted to the available hardware environment.

Transfers and exchanges

Product informatic files may still prove insufficient. It is necessary to give the possibility to an exterior request to access the base, extract data and transfer them automatically to another computer in a preestablished format.

In the reverse way, it is easier to receive automatically such data (modifications, updates, creations) according to well defined formats than to wait for information or diskettes, tapes or paper in varied format or even quaint format.

To answer this type of needs, the classification server will contain an EDI function (Electronic Data Interchange) which :

- extracts the requested data
- formats the received data
- sends or receives normalized messages.

This function however will not be possibly put on before the designing and acceptance of an international message normalized and adapted to classifications and tables. This work is being done at present in the frame of the EDIFACT-Board and should open into tests during the year 1995.

Management

Classifications develop (new levels, changing in items, headings, explanatory notes, introduction of jurisprudence). It is necessary to bring in all those modifications in real time, even by advance.

Management functions allow the updating of all the entities by creation, modification, cessation or deletion. The system carries on consistency checks for every intervention, warns the user of the consequences of the intervention and carries on himself some when the overall logic allows.

In spite of programmed checks and some preestablished operating sequences, this first version of the management functions remains still very "atomized". It will be enhanced by better sequences and more numerous checks thanks to the experience of the first complete campaigns of updating that it will have allowed to realize.

Formatting of data and loading

The quality requirement either for the data presentation or their exactness has lead to finalize a set of utilities on PC to help the management clerks to format the data and avoid rejects of the system generated by errors.
Loading programs checks the consistency of data (either in an internal way or with their environment) before doing the loading of the data in the base. They warn the management clerk of the gaps, anomalies, omission or other found inconsistencies. It is then the operator's job to correct or validate the data that cause the message;

The interest of a unique software to manage and disseminate different families of metadata is evident : a single maintenance, a single working method, a communication and query homogeneous for sundry operators and further an interconnection with comparable tools set up in varied places.

Interfaces

Access modes to every software rely on informatic environments. In the case of the classification server they rely on the choice of architecture (IBM and UNIX).

For its own needs, INSEE has developed three different user interfaces (terminal, under WINDOWS, VIDEOTEX) and planned the possibility of an all UNIX environment under X-MOTIF.

This variety thus allows access via four types of hardwares : 3270 screen (in an IBM architecture), PC, work station (UNIX) and VIDEOTEX.
The ergonomics is common to 3270 screens and VIDEOTEX on the one hand, to PC and work stations on the other hand.

---oOo---

The classifications complex tools at the daily disposal of statisticians, but also of other operators (general government, entreprises, scientifics) deserved a particular development fitting to their specificities.

Codes and lists, connected objects are also concerned by this management and dissemination tool box. It turns out that other objects comparable as for the structures can also be taken into account.

- organization charts (structures, objects, name of the persons and functions)
- metadata "dictionaries" (such as EDIFACT dictionary, ISO standards, etc).

The "classification server" product has thus a much larger calling than the one detected by the initial analysis of needs associated to management of activities and products classifications in an unsettled international environment.

The linguistic server

The SYNAPSE linguistic part has been constructed for three goals :
- ease access to activities and products classifications for non informed users ;
- bring an aid for codification for other operators ;
- allow mass automatic codifications.

A fourth interest has been confirmed by the tests :
- give an aid to the construction of relation tables by using the language and not only exogeneous logical choices.

What is the linguistic server ?
It is an application which analyzes descriptive texts written in current language, turns them into formulae and compares these formulae to the indexes of descriptive texts defining classification items (headings and explanatory notes).

Two basic differences and the important consequence that follows) with the keyword systems are to be underlined :
- the linguistic server does not seek, to compare them, character strings but concepts (one word can cover several concepts)

- the descriptive text can include (and it is desirable that it be so) link words, prepositions and punctuation. These are words which, meaning empty for keywords systems, give all their meaning to descriptive texts.

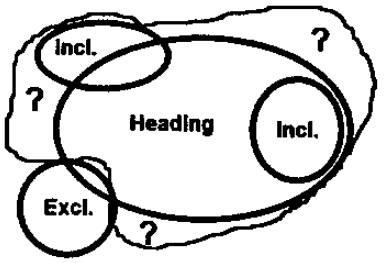
- consequence : the answers to queries are not items the headings of which have a set (or a subset) of keywords in common with those of the queries, but the items the semantic contents of which defined by the headings explanatory notes, even the jurisprudence is the nearest.

How is working the linguistic server ?

The basic components are :
- systems of reference (dictionary, thesaurus, index base) ;
- algorithms (grammars, navigation in the network of classifications).

Every descriptive text is analyzed according to the following :
- a text is transformed in a formula containing one or more interpretations ;
- a formula is described by terms (or concepts) which make the semantic network of the thesaurus ; attributes precise the vocation of terms in the formula (and thus in the descriptive texts),
- the thesaurus terms are associated together by links of domain, proximity and genericity-specificity,
- the thesaurus terms are associated to the words of the dictionary which have properties (utilizations, domains, synonym links).

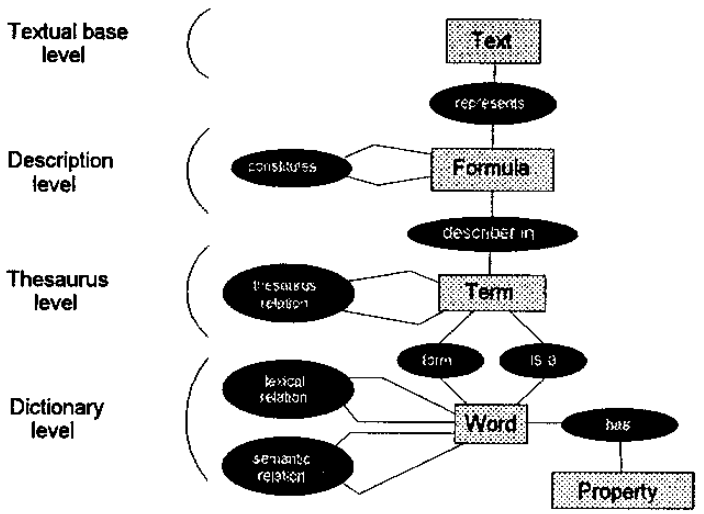
Recall on the definition of an item



- a heading "sweeps" roughly the content of an item and only seeks to name it,
- explanatory notes precise this content and partly define its border in inclusion as in exclusion,
- shadows subsist that only jurisprudence can fill in with time passing.

To define the semantic field of the items, all the known descriptive texts (headings, explanatory notes, jurisprudence) are analyzed and indexed on the corresponding items. Exclusions send back to the good items even if these items do not have the corresponding descriptive texts in inclusion. This set makes the index base.

The breaking up of a text

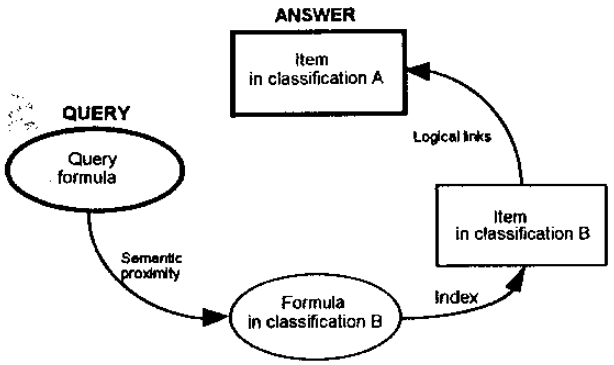


Every formula representing the text of a query is compared to the set of formulae of the index base (with an optimization algorithm).

After comparison, according to a scale of semantic proximity computed by another algorithm, the server sends back as an echo, for each found interpretation, the item(s) answering the query.

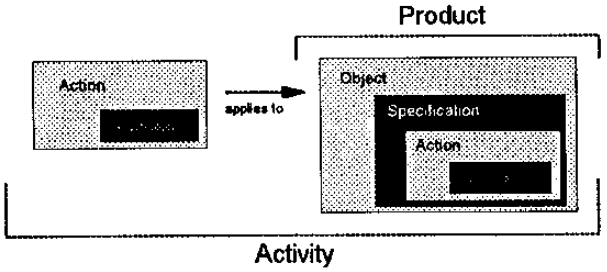
If equivalent (or near) indexes of the query formula are found in other classifications, the system uses the relations between items (logical network) to answer in the classification and at the chosen level.

Example : search in an A classification



In case of multiple interpretations for the same query
- for a consultation, the operator disambiguates by choosing the good interpretation
- for an automatic codification, the system keeps the most probable, that is to say the semantically nearest of the query.

Descriptive texts and multilinguism
In almost all of the european languages (at least), the descriptive texts for activities and products follow the same schema :



- an action (possibly specified) either describes a single activity or describes it in applying to a product
Examples : trade
cutting
wholesale trade
mounting for a third party of X

- a product is described by an object possibly specified. This specification can be itself an action (specified or not)
Examples : furniture
machine tools for wood
agricultural machinery
car equipments
spare parts of Y

- the union of both parts (action + object) leads to the expression of sometimes complex descriptive texts (not taking into account synonyms, periphrases or inversion)

Example : design, making and repair of machines aiming at food preserving.

This generic formulation of activities and products descriptive texts lead to a significant economy for the use of another language. Actually, the structuring of the concepts in the thesaurus and formulae representing descriptive texts for classifications are independent of the language. In the other hand, the vocabulary (dictionary) and its properties and grammars depend on the language.

The linguistic server includes a thesaurus of multilingual type which avoids a new start from zero in the case of the taking into account of a new language and allows to limit the works to those specific to this new language :
- grammar developments
- making of the adapted dictionary
- translation of the thesaurus
- indexation of new classifications (except the international and european classifications)

The complete set of functions and linguistic tools being set up, the global economy for the processing of another language is around 50 to 70 % according to the languages.

How is developing the linguistic server ?
A toolbox for linguistic management and indexing is integrated to the linguistic server.
It allows to :
- complete the dictionary,
- enrich the thesaurus,
- index new classifications,
- over-index, modify or correct the indexes of the already loaded classifications,
- analyze the construction of formula to correct indexes, grammars, even descriptive texts...

A consequence of the linguistic server is to allow the classification specialists to correct the headings or explanatory notes written by themselves : what the server does not understand may of course be linked to a bad indexing or imprecision of grammars but is also very often due to a bad writing of the descriptive texts.

At the end of this item will be found "Some thoughts, questions and advices" concerning linguistic servers.

What results may be expected ?
Be it an expert-system, the linguistic server should not replace a human's analysis above all if he is an expert in classifications. Otherwise, as the system works using knowledge base, its quality depends on its enrichment with time passing and not only of its initial status. This been said, its performances are nonetheless interesting.

To measure the possibilities of the server, INSEE has accepted a battery of 7 tests (8 for the further versions) holding 8300 descriptive texts of variable complexity, coded in three classifications (NAF : French activity classification, HS : Harmonized system, CPA : Classification for products associated to activities). These descriptive texts are : well formulated or badly written activities, products or corporate names including an indication of activity.

The tests have been broken up in "complex", "simple" and "heterogeneous" queries. Two environments have been taken into account : query (a single screen of thirteen possible answers is queried ; the good answer is in it) and automatic codification (only the first given answer is interesting).

Four indications have been accepted :

- efficiency = $\frac{\text{number of codified descriptive texts}}{\text{total number of descriptive texts}}$

- quality = $\frac{\text{number of well codified descriptive texts*}}{\text{number of codified descriptive texts}}$

* among the queried 13 or the first codified

- performances :
  - average time for an answer
  - average number of answers

The results are the following :
(with version 2.0 delivered in july 1994)

Corpus	Efficiency in %	Quality in % Query	Codific.
Complex	82,2 à 85,7	63,1 à 70,0	40,3 à 45,5
Heterogen.	87,8	73,3	55,6
Simple	97,6 à 97,8	87,0 à 95,3	65,7 à 73,9

Corpus	Average time in sec.	Average number of answers
Complex	) < à 1s	3,4 à 5,1
Heterogen.	) (de 1/10	4,6
Simple	) à 13 s)	3,9 à 5,2

It must be underlined that tests carried on two complex corpuses and done by "specialists" (not experts) of NAF and a keyword system used in the Register of entreprises give the following results.

Complex corpus	Efficiency in %	Quality in % Query	Codific.
Specialists	93,4 et 100	///	58 et 67
Keywords	62,0 et 91,4	56 et 62	25 et 29

If no statistical value must be given to such a comparison it is nonetheless a "performance indicator" of the linguistic server.

Expected improvements (developing of grammars, algorithms and indexes of new classifications, introduction of jurisprudence) will only have a minor effect on efficiency (answers rate). On the other hand, quality either in codification or in query must still improve of 10 to 15 points ; thanks to these improvement end up to the approximate awaited results :

Corpus	Efficiency in %	Quality in % Query	Codific.
Complex	85 à 90	75 à 80	50 à 60
Heterogen.	90 à 95	80 à 85	60 à 70
Simple	98 à 99	95 à 100	75 à 85

One will not hope better, firstly at least.

Some thoughts, questions and advice
After having invested in a linguistic system, INSEE has not become for that reason an expert in the domain. The experiment of the development of the linguistic server has however generated an accumulation of information which can be useful to others.

- the linguistic problems concerning activities and products descriptive texts have been clearly identified,
- the questions to be asked concerning the methods, the algorithms and the performances of the linguistic systems are well listed ;
- the possibilities and the limits of such systems have been analyzed in a realistic way ;
- the battery of tests aimed at judging the results (included analogous servers) is varied if not complete,
- the development and maintenance costs problems are well known.

All these informations can be communicated after asking INSEE.

Informatic choices

The informatic architectures

The informatic architectures have been studied depending on the INSEE informatic environment and of the will to have the best possibility of conversion. This strategy leads to present the different derived applications of SYNAPSE under two architectures :

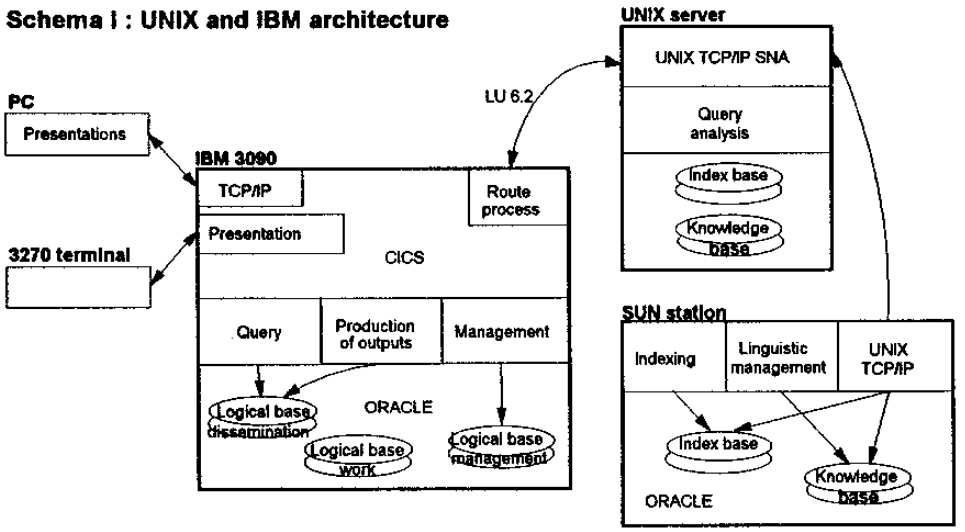
- on IBM 3090 under CICS for the logical part and UNIX RS 6000 server (or DPX20 BULL) from the linguistic part (see schema 1)
- on UNIX RS 6000 server for the whole (see schema 2).

In both architectures, linguistic management and indexing are set up on SUN workstation.

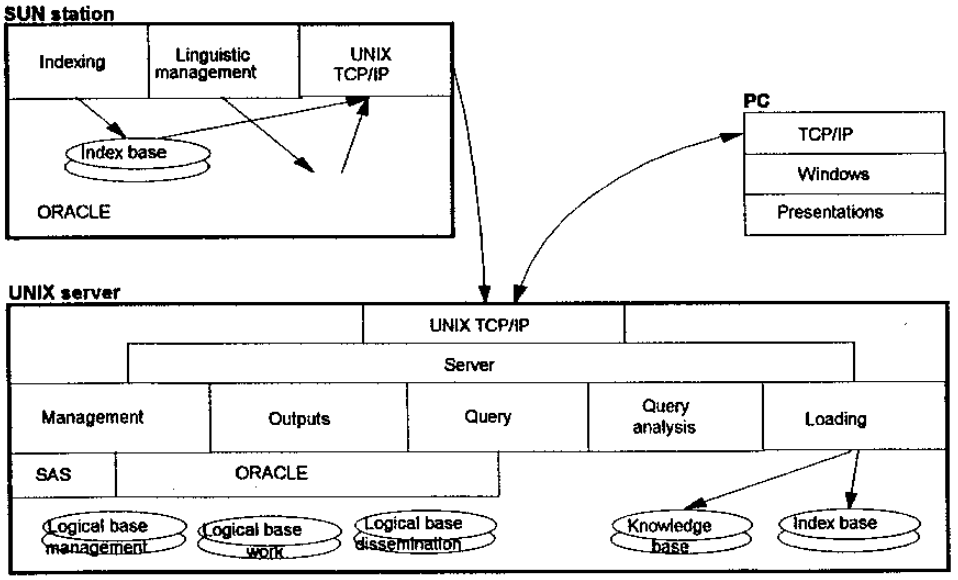
If one of the applicative products must be installed in another architecture and/or on other hardwares than those described above, conversion costs are to be contemplated.

In every cases, hardwares are provided by the buyer and must be sufficiently sized to welcome the application(s) and respond to the awaited traffic.

Schema 1 : UNIX and IBM architecture



Schema 2 : all UNIX architecture



The main characteristics

The choices of languages and softwares are the following :

- functions are programmed in C language
- the calculations of relation tables weights are SAS macros,
- connections use LU 6.2 protocol,
- data bases are under ORACLE,
- the linguistic toolbox is ALETH software (from society GSI Erli).

The user interfaces offered are at the same time linked to architectures and open :

- for IBM architecture : 3270 terminal interfaces under CICS for query and outputs production, under D I A L O G - MANAGER for the management of logical data.
- for both architectures for both families of screens :
  PC interfaces under WINDOWS
  X-MOTIF interfaces for UNIX

This double choice (of architecture and interfaces) broadens the possible field according to the existing informatic environments.

The dimensions
Some figures can give an idea (simplified) of the dimensions of the classification server.

	Logical server		Linguistic server*	
	At present	Later	At present	Later
Code lines	80 000		150 000	
Number of screens	60		///	
Infor-mations	35 000 items	95 000 items	65 000 indexes	150 000 indexes
Dimension of the bases	153 Mo	320 Mo	13 Mo	17 Mo

(compiled)

* Except management and indexing software installed on SUN station : twenty screens ; roughly 200 000 lines of code

Installation and maintenance
The transfer of the rights of use of any software imposes a certain number of obligations and after sale services. The complexity of the software as its foreseen developments (shared or not) imply a minimum of coordination.

After having taken down the general operations common to the different applications, we must distinguish between the problems linked to the logical server from those linked to the linguistic server.

General operations
INSEE commits (with the help of a subcontractor)
- to install and test the softwares in the chosen informatic architecture ;
- to guarantee the operation of the applications for a six months period,
- to train a team to manage the classifications in the logical base, to use the classification server (query and production of outputs),
- to translate the interfaces and/or adapt the screens to the classifications to be loaded,
- to supply the user documentation (in french and/or in english),
- to deliver yearly the updatings of the logical data base (if necessary),
- to deliver the enhancements (interfaces, algorithms, functions, enrichments) which would be brought to the softwares (be they realized by INSEE or by a user).

Concerning this last point, in order to take into account each one's desiderata and to coordinate the software developments, a "User club of SYNAPSE" is created. Its vocation will be to improve the quality and performances of the applications in providing the users of the new developments. It also aims to avoid to have to maintain too different versions simultaneously.

Logical server
Except the general points quoted above, maintenance only concerns the so called software on the basis of the last known version (that is to say without any unilateral modifications) The maintenance is "on request", owed every time it is needed with a minimal yearly fixed sum ;
It is realized :
either by telemaintenance when possible
either by shipment of corrected versions (on diskette or cassette) installed by the user
either by "on the site" fixing (with travelling expenses in addition).

Linguistic server
The conditions of the software maintenance and of the grammars are the same as for the logical server. As for the knowledge base, two problems come up :

- maintenance of the data (dictionary, thesaurus, index) may be realized either by the user or by linguists or shared.
The choice of a self-maintenance implies the training of linguistic management clerks. We have also to keep in mind that knowledge bases identical at the start but managed by different teams will quickly diverge.

- the taking into account of another language on the linguistic server (translation of a thesaurus, adaptation of a dictionary, development of grammars, enrichment of the index base) implies moreover, a sharing out of the property rights :
- the adaptation of the linguistic server which the addition of another language represents
is realized by the user or subcontracted
is a co-ownership
can be transfered to third parties under conditions jointly accepted by the co-owners.
- the maintenance of the grammars and of the bases in this language are chargeable to the users (if INSEE had to belong to these users, the same obligations would impose to it).

Derived products from SYNAPSE

The modular architecture of the classification server developped by INSEE allows to give five products adapted to the needs of users concerned in different ways in the domain of classifications.

- an application of management and dissemination of any set of classifications
- an application of query in natural language and automatic coding, for activities and products classifications
- an application integrating the two previous applications (SYNAPSE)

For each of these products there is an item describing the product and associated services.

Technical notes:
The three applicative products can be installed on the following architectures :

- for SYNAPSE and the logical server : on IBM 3090 and/or UNIX server of RS 6000 type, loaded (if necessary) with the data on activities and products classifications.
- for the linguistic server : only on UNIX server of RS 6000 type.

Other hardware configurations are possible. In this case, the hardware with sufficient dimensions to welcome the chosen applications must be provided. Conversion costs are to be contemplated

User interface may be of Windows type or of terminal type (3270 sceens) for all the functions (query, outputs, management)

Item 1 - Logical server (management and dissemination application of a set of classifications and relation tables)

This application provides management and query of any set of classifications, codes and lists as well as relation tables which are associated to them.

Query
Access to items can be made :
- from code lists
- or codes
- or relation tables (hierarchies, correspondences, overlappings)

Items are described by one or several headings (official, normalized, in other languages, etc.) and by explanatory structured notes, official or not (jurisprudence).

All the entities (classifications, levels, items, headings, tables, relations) are dated with a validity period proper to them. Choosing a reference date allows the selection of valid information for this date.
The network structuration allows navigation in the data base following the relations (hierarchical, correspondence, overlapping or historical) between the items.

File production
The application allows extraction and construction of information sets on classifications or tables corresponding exactly to user's needs (fields definitions, selection of levels, choice of variables).

A printing interface and a connected printer allow the printing of these informations.
A module of Electronic data interchange (EDI) will be developed in 1995 as soon as normalized EDIFACT messages will be available.

Management
All entities may be created modified or deleted.
A set of functions allows to format the data for their loading.
Checks allow the verification of the loaded and and modified data consistency.
A module of table computing allows the construction of relation tables synthesis of several tables.

Data bases
Three ORACLE bases are formatted to welcome the classifications data :
- a management base for operators in charge of introducing and managing the classifications and tables,
- a dissemination base, stabilized copy of the previous one, for queries and file requests,
- a working base for the application itself to save the files in a temporary way.

(1) A printing module on a Postscript printer with a SGML layer is available on IBM 3090 in the architecture chosen by INSEE. This module must be modified if the printing environment is different.

Item 1bis - Logical server adapted to activities and product classifications

This application is the logical server (see Item 1) in charge of a set of classifications of activities and products which will hold later (end of 1995) : 28 classifications (in which 11 international and european ones), 70 levels, 140 tables, 100 000 items, 175 000 llinks (in 300 files roughly).

All the entities are described ; for each item, the base contains the following informations (when they exist) : headings (official and normalized in french, official in english) and explanatory notes (structured, official or not).

In the tables, "weights" measure the overlapping parts or the correspondence parts of the items in partial relations.

Item 2 : Linguistic server (application of query in natural language and automatic coding concerning activities and products classifications.

This application allows the access to activities and products classifications through descriptive texts written in natural language more precisely than with keywords systems

Query
The system allows the access to items of any classification (described by headings, and explanatory notes) and indexed.

Queries are analyzed by a grammar capable to detect ambiguities (to be solved by the user) . Answers are classified by likelihood order.

To improve precision of the answer, the system uses the semantic links with the vocabulary used as well as logical links between items (1) (2).

Putting questions on products, one can have as an answer the different activities concerning the product.

Automatic coding
In the frame of works on a great scale (files, surveys, censuses), the linguistic server can codify automatically activities or products or be a codification aid (selection of items then manual validation).

It is also a tool to aid table construction (descriptive test analysis, codification then manual validation).

Linguistic management
The linguistic server includes a linguistic data management module (dictionary and thesaurus) and a classification indexation module (headings, explanatory notes, jurisprudence). Thus accepted language can be enriched and items definition can be improved (3)

Data bases
The linguistic server is made of two data bases : an index base (links between indexed descriptive texts and the classification items) and a knowledge base.

The latter information set constitutes the lexical and semantic network of the concerned domain.

The thesaurus is structured by roughly 12 000 concepts describing activities and products. The dictionary will contain later roughly 20 000 words covering the domain and bound to the thesaurus.

Two restrictions limit (voluntarily) the richness of the vocabulary taken into account :
- the professionnal slangs are not integrated (except particular cases),
- the domain of chemistry is covered by words that can be defined but does not cover formulae.

Dictionary and thesaurus are enriched by new words or concepts that the development of economic activity and language will introduce.

(1) The linguistic server is at present exclusively in french but can be enriched by other languages, the thesaurus being of multilingual type.
Adding a language implies the building up of a dictionary, formulation of a grammar, the translation of the thesaurus and the indexing of the loaded additional classifications
(2) A morphological analyzer corrects typing errors or orthograph. the phenomena of synonymy, genericity and specificity are taken into account.
(3) Query and indexation grammars as well as search algorithms can also be managed but remain of the linguist's competence.

Item 3 : SYNAPSE (complete application concerning activities and products classifications)

This application integrates logical and linguistic applications (for more details, see items 1bis and 2) and respond to all users needs concerning :

query,
dissemination,
extraction of informations,
automatic codification,
management,

of activities and products international, european, national and particular classi-fications.

Some examples of screens

Welcome

For the frequent users : a selection of the more used classifications and tables ;
For all : the other accesses to the base and the general information.

The reference date (put by default to the date of the day) can be modified.

```
                              SERVEUR DE NOMENCLATURES

  Identifiant  : <          >    obligatoire pour Production d'une sortie
  Mot de passe : <          >    facultatif pour un accès standard

 1 Recherche dans la NAF (au niveau le plus détaillé)
 2 Recherche dans la NC
 3 Recherche dans la CPV (au niveau le plus détaillé)
 4 Recherche dans les Nomenclatures d'enquêtes (nouvelles)
 5 Listage de la table NA73-NAF
 6 Listage de la table NC-PRODCOM
 7 Listage de la table Nom. Enquête ancienne-Nom. Enquête nouvelle
 8 Autres consultations
 9 Production d'une sortie
 A Listage des dernières mises à jour
 B Informations générales sur SYNAPSE

 Choix :  8
 Date de référence des informations :  14/09/1994


 F1:aide   F3:sortie du serveur   ENTREE:suite
```

```
SERVEUR DE NOMENCLATURES: Choix d'une nomenclature     (données au 14/09/1994 )

     Type                                        choix du type :  P

A  Activités
P  Produits
E  Echanges extérieurs

- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -

     Nomenclatures                          choix de nomenclature :  cpa

CPC      Classification des produits centrale
CPA      Classification des produits associés aux activités
CPF      Classification des produits française
NP73     Nomenclature de produits 1973
PRODCOM  PRODCOM
NOMEP    Nomenclature détaillée de produits


 F1:aide   F2:info   F3:retour   F4:accueil   ENTREE:suite
```

Choice of a classification

The selection of a type of classification displays the list of those of this type present in the base. The choice of the levels is made in a following screen.

The types are, here, adapted to the activities and products classifications.

Interpretations

In case of ambiguïty, the linguistic server returns the different interpretations of the query together with a proximity indicator (semantic and/or logic) and the number of concerned items.

```
SERVEUR DE NOMENCLATURES: Interprétations          (données au 14/09/1994 )

Nomenclature source : NAF                                          page  1/1
Nomenclature d'activités française
Niveau source : Niveau 700
Validité du 01/01/1993  au aucune

    Distance  Nombre de      Interprétation
    interpr.  postes

1    00        2      PRODUCTION (D'UN BIEN) FRAISE (FRUIT)
2    01       20      PRODUCTION (D'UN BIEN) FRAISE (OUTIL)
3    01        7      PRODUCTION (D'UN BIEN) DIAMANT (PIERRERIES)
4    02        1      PRODUCTION (D'UN BIEN) DIAMANT (OUTIL)
5    03        1      PRODUCTION (D'UN BIEN) DIAMANT (TETE DE LECTURE)



Choix : 4

 F1:aide   F3:retour   F4:accueil   F7:page préc.   F8:page suiv.   ENTREE:suite
```

```
SERVEUR DE NOMENCLATURES:    Table                  (données au 14/09/1994 )

Niveau source :    NA73       Niveau 600                          page 3/7
Niveau cible  :    NAF        Niveau 700
Type : Historique
Date de création :       01/01/1974
Date de fin de validité aucune

LISTE DES RELATIONS                              pondérations
      Poste source      Poste cible       à la source  à la cible; commentaire
 19    24.04            29.1A             0.33         0.25
 20    24.04            29.1J             0.33         0.33
 21    24.04            34.1Z             0.33         0.14
 22    24.05            29.1C             0.33         0.17
 23    24.05            29.1D             0.33         0.25
 24    24.05            29.2D             0.33         0.13
 25    24.06            29.1C             0.50         0.17
 26    24.06            29.1D             0.50         0.25
 27    24.07            27.2C             0.50         0.25

choix de la relation : 19

 F1:aide F2:info F3:retour F4:accueil  F7:page préc.  F8:page suiv. ENTREE:suite
```

Tables

The display of a table can be limited to a field defined in the source level as in the target level.

The weights (part of an item covered by another) allow to "measure" the importance of the link.

(In the opposite screen, weights are pseudo-weights calculated by the system)

For more information, please contact :

Emile BRUNEAU
Département des normes
statistiques et comptables
Timbre D230

Tél : (++33) 1 41 17 52 74
Fax : (++33) 1 41 17 68 49

INSEE
18, Bd A. Pinard
75675 PARIS cedex 14
(France)

Elisabeth BARTHELEMY
Département des Projets
Timbre C561

Tél : (++33) 1 41 17 35 70
Fax : (++33) 1 41 17 78 44