WP.16
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE EUROPEAN COMMUNITIES (EUROSTAT)

ORGANISATION FOR ECONOMIC COOPERATION AND DEVELOPMENT (OECD)
STATISTICS DIRECTORATE

Joint UNECE/Eurostat/OECD work session on statistical metadata (METIS)
(Geneva, 9-11 February 2004)

Topic (i): Functions of metadata in statistical production

NOMENCLA: A TOOL TO MANAGE, DISPLAY AND DISSEMINATE METADATA

Contributed Paper

Submitted by INSEE, France[1]

Summary

By the beginning of the 90s, considering expected strong changes in classification uses, INSEE decided to build up a tool to manage and disseminate classifications. Along the development period it became a tool for lists and tree structures linked together or to other data through correlation tables.

By the same time, a European group of statisticians developed an EDIFACT message to exchange, tree structures, lists, correlation tables and updates of these objects: CLASET (Classification set message). Along the time this message was enlarged and translated in SGML, XML, HTML (for browsing) and some other private formats. A dedicated tool-box (CLASET-tools) allows to cross as well from any format to any other ones as to extract Text files from the messages.
Today, as well the message as the tool-box are implemented in some NSIs DB including NOMENCLA.

NOMENCLA (including CLASET) is a triple application:
- a “Manager application” to manage the metadata under control and a precise process;
- a “Presentation application” to display, extract and compose XML messages covering any types of information in the DB;
- a “Linguistic engine” to search in the DB and to code external corpuses (including a function helping to build up correlation tables only based on textual information). Today this multilingual NLP application is in French and (partially) in English and open to German and Spanish.

A data model enough generic to be adapted to a lot of purposes, targets and types of data, structures the DB with details, optional or mandatory, which allow a flexible management depending on the wills of the users.

The present paper is synthesized in a PowerPoint presentation which underlines the targets, the specificities and the limits of the tools. A possible demonstration on a laptop shows the various functionalities, inputs and outputs.

[1] Prepared by Emile Bruneau (emile.bruneau@insee.fr).

I - INTRODUCTION

1 - By the beginning of the 90s, in order to prepare the strong changes on classifications the statisticians would have to face in 1993, INSEE decided to develop a tool whose the targets were

- to widely disseminate the new system of classifications (to statisticians, Ministries, professional bodies and enterprises);
- to link it to the present national one;
- to ease the understanding of and the searches in these new classifications with linguistic tools able to "enter" classifications not only through the wordings present in the classifications but through the semantic content of the descriptions.

2 - By the same time, as well Eurostat and some countries perceived the need of common tools to exchange information concerning the classifications and their explanatory notes, the concepts and their definitions and the various links between such reference metadata. Within the European Expert group for EDI and Statistics (EEG6) a working group was created with the target to develop a normalized message for tree structures (and flat lists) and correlation tables.

3 - As well the IT systems as the EDI formats strongly changed and advanced during the last twelve years. By the beginning of the 90s:

- INTERNET was not yet so widely shared;
- but in France, the MINITEL (using videotext) was widely disseminated;
- the mainframes remained the support of complex applications;
- the screens were terminals not really user-friendly;
- the languages used were not the same ones.

4 - So, the writing of such an application was done two times: for the previous environment (terminals, mainframe) and for the present one (graphic screens, Internet, servers). Apart from the targets and the basic models, about everything was changed and moved to "modern" solutions during the second run, the result of which is today presented.
Along the time the "possibilities" of NOMENCLA shew the first targets could be enlarged:

II - THE PRESENT CHARACTERISTICS

5 - **NOMENCLA a precise and powerful management tool**
The Manager is enough flexible to allow as well batch uploading and individual data management. But it is also « restricting » meaning a manager needs to provide clean work the application verifying what he does and the consistency of data before to upload them. Sure the human error always remains possible but is strongly limited.

6 - **NOMENCLA reference basis**
Any list, simple or complex classification, fully or semi finished tree structure, can be uploaded including all hierarchical links, internal or external historical links and links with other data. This can be a reference area for metadata (as soon they can be structured):

1. classifications (structured and lists);
2. lists of concepts and their definitions (possibly different depending on the environment), sources, if necessary, calculation methods, roughly all metadata;
3. geographical codes of any nature and their links between them;
4. classifications, physical units, questionnaires, survey managers and the like to manage the surveys (for instance, production ones);
5. organization charts and staff lists in order to know the successive allocations at a given position or the successive positions of a given person, as well the organisation chart and the list moving along the time;

6. a DB given including arrays, graphics and/or time series, all objects identified. Every individual entities (raws, columns, cells, curves) include information which can be described (definition, source, methods, etc.). Such a DB associated to NOMENCLA leads to a system (near) fully describing statistical data and allowing to search any datum or any information describing that datum: the meatadata supporting the arrays, graphics and/or time series.

(Apart from the last point, sure but not tested, all the other ones were verified).

7 - **NOMENCLA, "made-to-measure" extraction tool**
The filtering and selecting functions are designed in order to allow a user to only extract the information he exactly needs and not as often, all the information related to a list or a classification. This leads to limited sized files without having to delete 90% of the downloaded file after having sorted not useful information.

8 - **NOMENCLA, tool helping searches**
Either through « logical » functions (along the tree structures, by level or codes) or through linguistic functions, from the less expert user to the specialist can find or extract what he wants.

9 - **NOMENCLA, complement of automatic coding systems**
Generally the "automatic coding systems" only partially encode batch files (from one to two thirds of the data depending on the capacity, on the wording complexity and on the reference complexity. And then?).
The NOMENCLA linguistic part can be used as an extra tool in a further coding step when an automatic tool does not code. Such a step can help the users providing a limited list of "proposals" sorted by score from the best score to the minor ones in order to avoid manual mistakes and or long manual searches.

10 - **NOMENCLA, help for matching corpuses**
Nor rare as well in NSIs or in Enterprises are the data coming from various sources linked to not usual or unknown metadata (e.g. in the INSEE's macro-economic database, the thousands of data are linked to more than 600 classifications more or less derived from some reference classifications. The "help to build table" function of the linguistic application provides to the classification specialists and to the users an help which proposes draft tables avoiding long manual searches to match the various corpuses.

11 - So, who can be concerned by NOMENCLA? Anybody being concerned by one of the above listed works: from a simple search to precise analysis of links between corpuses; From the extraction of simple files of codes and headings to the re-uploading of a DB; From the management of a flat classification to the management of a complex network.

12 - As any sophisticated software, NOMENCLA needs to be « learnt » (at least the Manager) through a time-limited training in order to understand the internal rules. Concerning the Linguistic application a real investment is necessary about the natural language processing but only some persons are concerned (if the linguistic management is not externalised…).

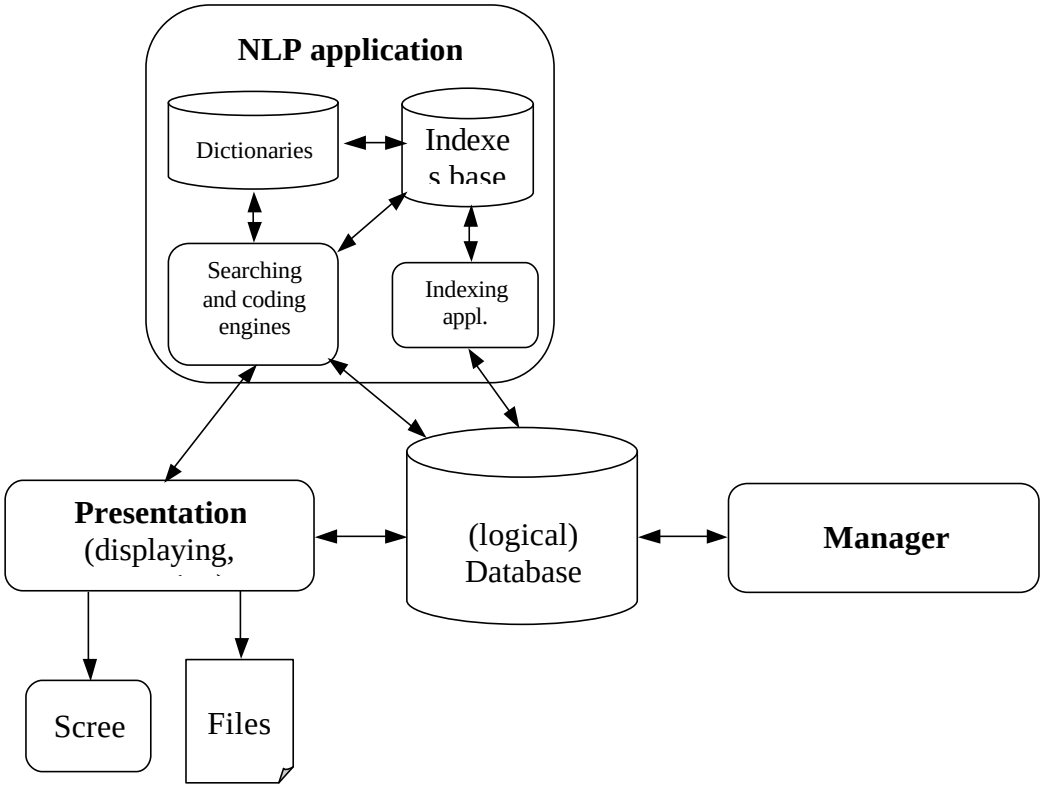
13 - Sorry, but NOMENCLA does not make coffee or prepares tea… but it can manage classifications of the various coffee and tea types. Even generic such a tool is not devoted to nobody because generic: the management and consulting of tree structure problematic is not depending on the data but on the types of data.

III - THE APPLICATION STRUCTURE

14 - NOMENCLA is an application composed of three sub-applications:
- a Manager software to *manage* the database. This client-server application composed of about one hundred functions shared between sixty screens allow to create, modify, stop, delete any objects (entities) and their attributes;
- a Presentation software to *access* the database in order
    - to *display* any present information at a given date (photography) or between two dates (updates) or
    - to extract needed data in files to be downloaded;

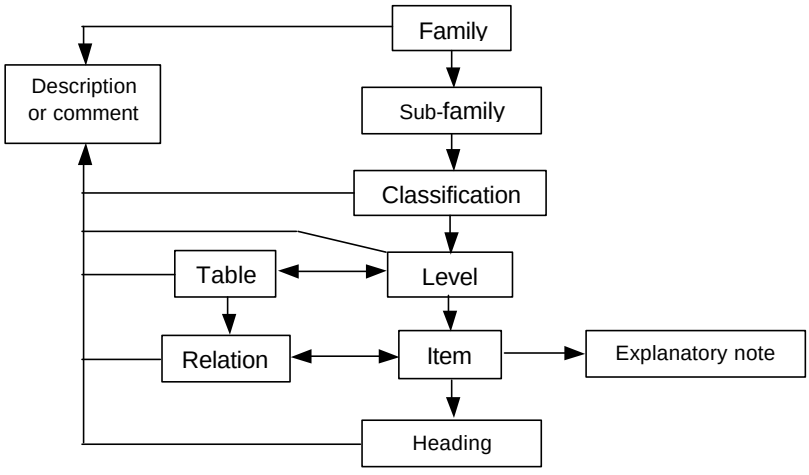
- a Linguistic application (using NLP : Natural Language Processing) to access the *textual* information allowing to enter free natural language, not only based on the headings, not only the wordings present in the database (mandatory restricted), but various expressions describing the same concept or object even with different syntaxes or different wordings.
  Such an application can run alone (for batch coding purposes or as help to the building up of tables) because it also includes the knowledge of the networks (structure of the tree structures and correlation tables) but the best use includes as well the linguistic part associated to the logical one.

15 - The resulting software architecture, which can seem a little complex) is the following:



IV - THE DATA MODEL

16 - A review of the new statistical classification system (specially in Europe) shows very rare are the classifications not linked to another one, either by construction or one defining another one. This review leaded to a data model which creates an *information network* and not “individual objects”.
In order to aim such a target, the following (here simplified) data model was designed and implemented:



A family is a set of classifications or lists directly or not linked together through tables in a network.
A sub-family is an indicator which structures the families in order to avoid "mixtures" of classifications of different types (e.g. activities, products, commodities within the family of economic classifications);
A classification is a tree structure (which can be "flat" as, for instance, lists for any purposes);
A level is one of the partitions of a tree structure (if more than one);
Levels (and not classifications) are linked together by tables;
An item is a building block of a level (these building blocks can be modified along the time);
An item is described by explanatory notes of different types which can change along the time;
An item can have more than one heading (by the same time or along the time);
The items are linked by relations which compose the tables;
All the entities (apart from explanatory notes themselves descriptions) can be described or commented.

17 - All these entities are described by attributes which precisely define the objects (e.g. origin, validity period, maintenance agency, status, type, etc. depending on the object). But the most important attribute of all objects (and, when necessary, of the links between these objects) is the *validity period* which allows to manage the database not repeating the same information and avoiding to create "versions" of classifications or lists: for instance, the Combined Nomenclature -EU foreign trade classification- slightly changes every year (some tens or hundred of items within more than ten thousands). After fifteen years (the CN was created in 1988) the management of versions would imply about 160,000 items in the DB, the management of validity periods only leads to 16,000 items…
The only "constraint" of NOMENCLA is to precise at which *date* we want to display or extract data.

V - THE MANAGER

18 - The various managing processes are necessary in order to assure a high level of DB consistency. The building up of a reference DB including lists and/or classifications linked together and changing along the time, at least needs to conceptualise a network then to upload and to update data. By the same time, depending on the work organisation, to manage a list of managers with their rights (what they are allowed to do ? the general user rights –consultation and extracting- can be eventually managed through different DB if limited accesses exist).

19 - In order to respond to such rules, the Manager offers a lot of functionalities shared between 63 screens helping the managers and controlling their work in order to minimize the "logical errors[2]".

Follows the (simplified) list of managing functionalities which shows all the possible elementary process operations offered:

Managing the system

- *Managing texts (in four languages if necessary)*
  - Texts of the HTML pages
  - Texts of buttons
  - Texts of Helps
  - Texts of system messages
  - Last information on families
- *Managing the "managers"*
  - Managing the rights

Managing the DB (creating, modifying, stopping, deleting)

- *Processing batch files*
  - Parameters
  - Controls
  - Weight of relations

[2] Even with an ultra-sophisticated software, costly and heavy, the human errors can be totally avoided (e.g. orthographic mistakes or bad choices). NOMENCLA only targets to limit logical errors using constraining work rules.

Uploading identifier

Managing individual objects in the DB (all texts in four languages if necessary)

Families
Classifications
Levels
Items
Headings
List of concerned items
Choice of the update type
Relations between levels (hierarchies)
Tables
Re-weighting tables (see below)
Relations
Relations concernées par des cessations de postes
Explanatory notes
List of concerned items
Choice of the update type
Descriptions

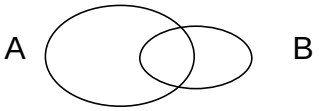
Consulting the managing DB

Uploading the consulting DB (from the managing DB)
(Furthermore, two hyper-links call the linguistic managing and the batch coding functionalities.)

20 - Apart from the management of validity periods instead of successive versions of the uploaded objects, the other specificity of NOMENCLA concerns the tables: any relations of the tables can be "weighted" that means the partial parts respectively covered by the related items can be assessed by two ways and linked together up:

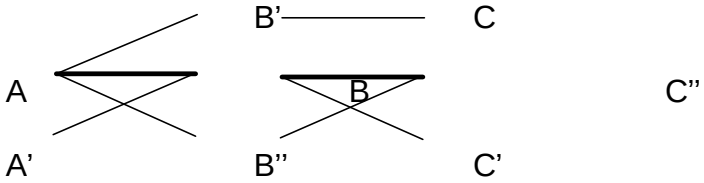
- either statistical data exist the manager wants to apply, which provide the relative overlap of each item in the other one. He can input such weights;
- or no statistical data are available and the system can automatically weight the relations.

Explanation. A relation links two items a priori differently « sized ».



The intersection (A ∩B) represents the common part (overlap) of both items.
The weight of A in B and of B in A can be measured by P(A/B) = A ∩ B / A and P(B/A) = A ∩ B / B which provide an approximated measure of the respective overlaps.
In NOMENCLA, the weights of relations in any direct tables (between two adjacents levels) are calculated considering the opposite numbers of relations owned by each item:

For the indirect tables, the weights are linked up in order to provide « estimated » weights between two non adjacent levels :



Between the level “A” and the level “C” we get the following weights:

A	C	**1/3**=1/3*1	**1** =1*1
A	C’	**1/6**=1/3*1/2	**1/2**=1*1/2
A	C’’	**1/2**=1/3*1/2+1/3*1	**3/4**=1/2*1/2+1/2*1
A’	C’	**1/2**=1*1/2	**1/2**=1*1/2
A’	C’’	**1/2**=1*1/2	**1/4**=1/2*1/2

We can remark:

Σ P(A∀ “C”) = Σ P(A’∀ “C”) = Σ P(C∀ “A”) = Σ P(C’∀ “A”) = Σ P(C’’∀ “A”) =1

More the levels are from each other (multiple tables between them) in a network, more the results are « near an acceptable image of the reality ».

The choice is let to the managers to choose which weighting way they want to follow: either a manual way (but they will be obliged to also manually update such information; or to be assisted by the system and any changes and updates will be don’t by it.

21 - In order to upload the various types of data, the types of batch files to be uploaded are very simple and limited to three types:

1. Code <tab> Heading : for items and various headings (*72.30 Data processing*) ;
2. Code <tab> Code : for tables (*72.30 7704*) ;

3. Code (*72.30)*
   Type (e.g. *1 -central content-)*
   Text (*This class includes:*
   // – database related activities: provision of data in a certain order or sequence, by on-line data retrieval or accessibility (computerized management) to everybody or to limited users, sorted on demand
   – processing of data employing either the customer's or a proprietary program:
   • complete processing of data
   • data entry services
   • scanning of documents
   – management and operation on a continuing basis of data-processing facilities belonging to others
   – web hosting*)*
   for the explanatory notes //

Free text for descriptions and comments.

VI - THE DISPLAYING AND EXTRACTING FUNCTIONS

22 - In order to match with the various end user needs (searches, verifications ; extractions and exports) a broad set of functions is offered:

Consulting at a given date (photo)
- *Choice of a classification*
  - *Displaying related information*
  - *Searching along the tree structure*
    - Displaying all information about the items
      - Explanatory notes
      - Various headings
      - General information
      - Linked items
        - Navigation from items to items
  - *Extracting (part) of a classification*

Selecting restricted fields and data types
Choice of a level
Displaying all information about the level
Displaying included items
Searching items
Searching by texts
Searching by codes
Searching by lists (Limit codes or jokers)
Displaying all information about the items
Explanatory notes
Various headings
General information
Linked items
Navigation from items to items
Displaying linked levels (tables)
Displaying all information about the tables
Listing tables
Selecting (part of) a table
Extracting (part of) a table
Selecting restricted fields and data types

Consulting between two dates (updates)
Displaying updated objects
Choice of a classification
Displaying all information about the classification
Choice of a level
Displaying all information about the level
Listing the updated items
Displaying all information about the items
Explanatory notes
Various headings
General information
Linked items
Navigation from items to items
Extracting (part of) the classification updates
Selecting restricted fields and data types
Choice of a table
Displaying all information about the table
Displaying updated relations
Extracting (part of) the table updates
Selecting restricted fields and data types

On line help
Such functions match as well tree structures as flat lists and correlation tables of any types.

VII - CONCLUSION

23 - As already written, NOMENCLA was firstly devoted to manage classifications so does not probably fit with all the requirements necessary for all types of metadata: only metadata
pre-coded or not,
eventually structured,
described by textual information,
eventually linked together,
can be uploaded and managed.
Apart from the network of families, no drawing can be added. But probably more than 90% of the needs are presently covered. Adding functionalities always are possible. They remain to be developed.