Proteus:

a concept browsing interface towards conventional Information Retrieval Systems

Oreste Signore
CNUCE - Institute of CNR - via S. Maria, 36 - 56126 Pisa (Italy)
Phone: +39 (50) 593201 - E.mail: oreste@ICNUCEVM.CNUCE.CNR.IT

Alfredo Maria Garibaldi* - Maurizio Greco*
*Presently: *TECSIEL - via Enriques, 28 - 00146 Roma*
Phone: +39 (6) 505091

3rd International Conference on
Database and Expert Systems Applications

DEXA'92

September 2-4, 1992

Valencia

CONSIGLIO
NAZIONALE
delle RICERCHE
Istituto *CNUCE*

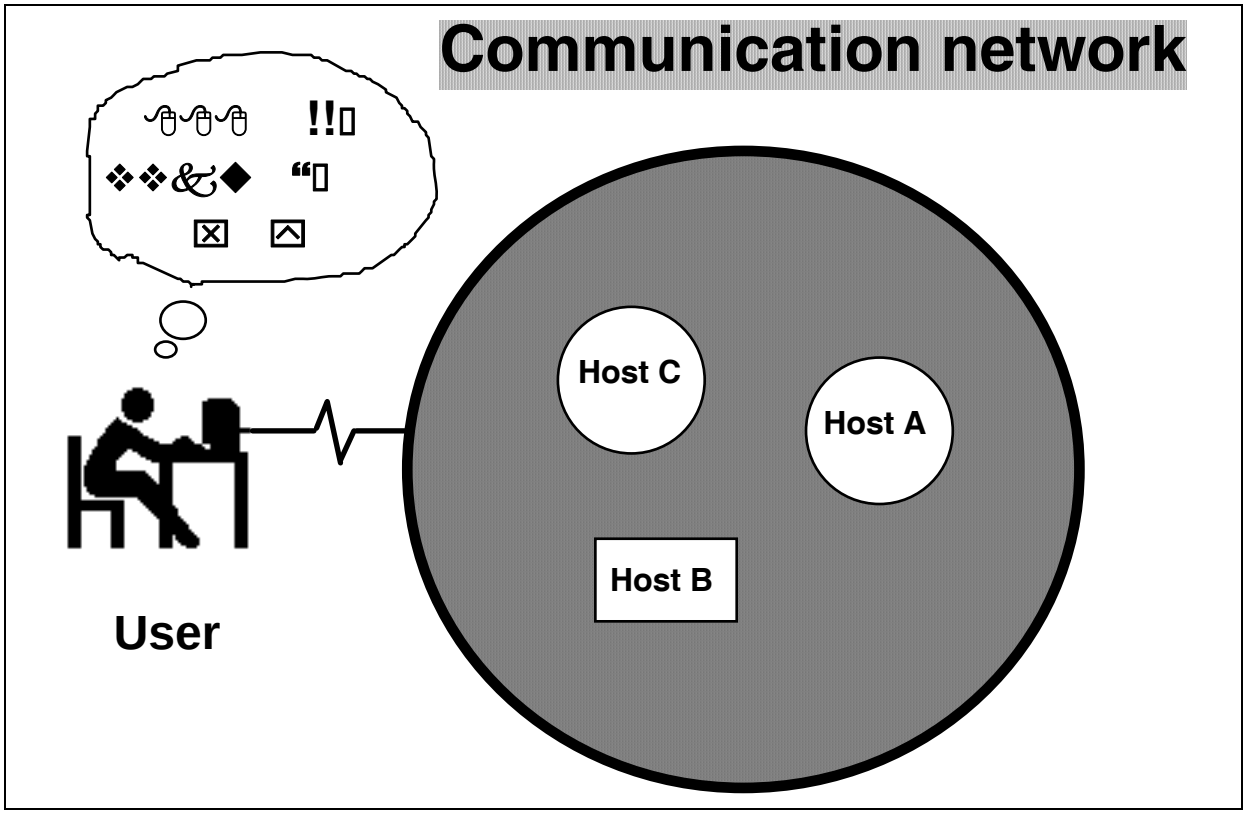
Contents

The myth of the on-line information

- I intend to have access to an on-line data bank...

- **No problem**, *all you need is a* **modem** *and a* **NUI**
  That's all! Just connect to the data bank of interest.

... and so ...



- **Data banks reside on hosts, and are managed by**

Information Retrieval Systems

Accessing on-line data bank services: some drawbacks

The growth of the information market has been limited by several factors:

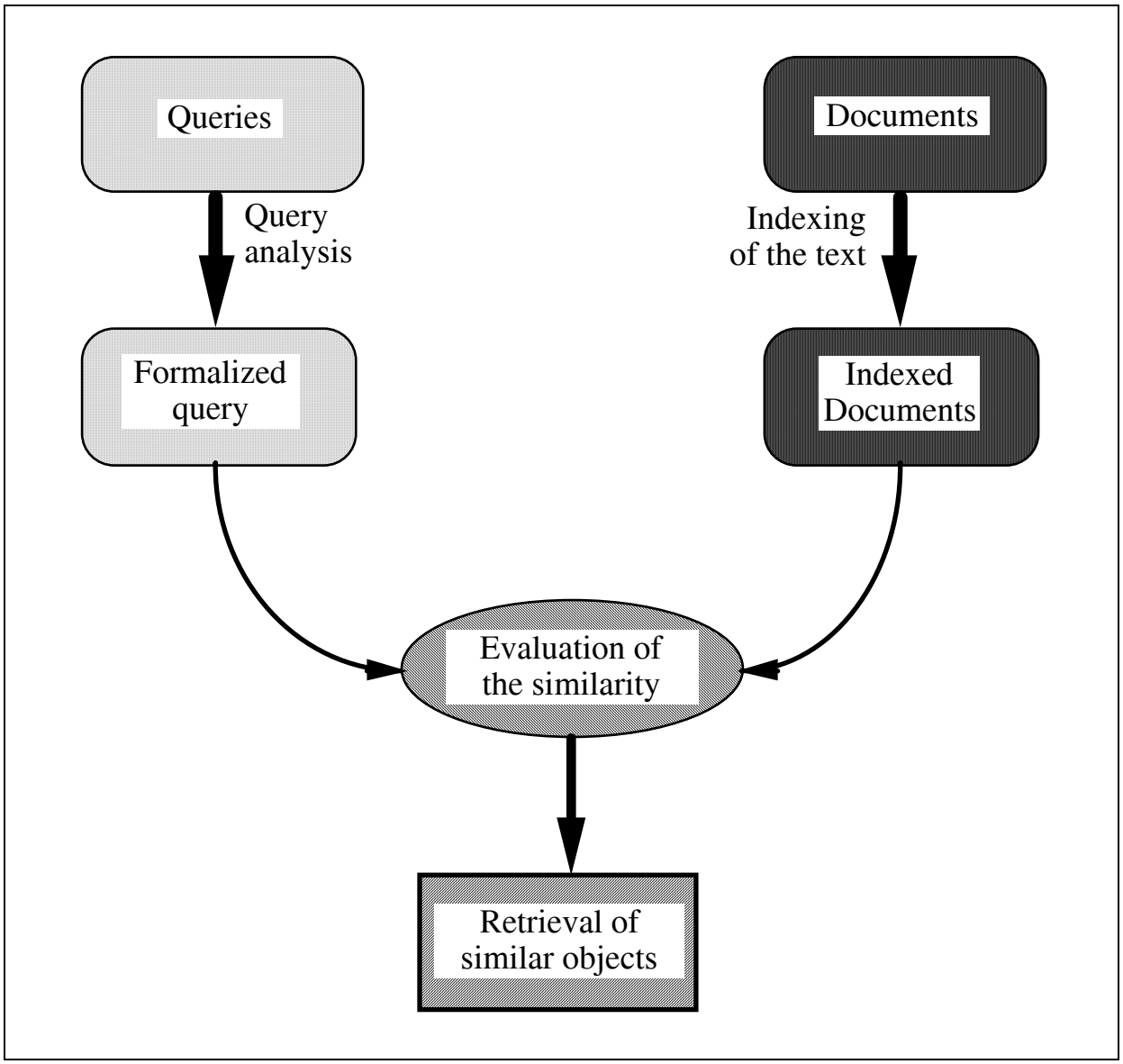
- **technological problems:**
  - **reliability and speed of the transmission networks**
  - **TTY based dialog**
  - **accounting problems**
  - **high costs (transmission and access)**

- **User interface problems:**
  - **complex login procedures**
  - **differences between query languages**
  - **data bank structure**
  - **contents of the fields**

Some solutions proposed at EEC level:

- **Common Command Language (CCL, also proposed as an ISO standard)**
- **special purpose workstations**

Functional description of an IRS



- **Stored documents are represented by a set of:**

  index terms

The user interface

p **In the Information retrieval approach it is necessary to:**

- identify the content of the documents **to be stored**
- identify the content of the user's query
- **begin a search based on a** specific topic

p **Users need support to:**

- define or refine **the topics which interest them**
- **improve their** knowledge **of the subject**
- **clearly** express the concepts **to be found out**
- refine or enlarge the concept **they were starting from**

?! A menu based interface?

Enter the required values

Author = _____ OR _____ OR _____

Year = _____

Argument = _____ OR _____ OR _____

No, thanks!

(I am able to understand the boolean logic, what I'm looking for is the content and the organisation of the data bank)

Thesauri and concept browsing interface

p **Typical problems:**
- **lack of precision of the utilized terms**
- **incorrect identification of the element to be found**

p **An approach based upon knowledge:**
- **assures a consistency in the representation between indexers and users**
- **provides a semantic system of concepts accessible to indexers and users**
- **creates a mechanism of communication that fits the key concepts.**

p **The user may be supported by the:**
explicit representation of the relationships of synonymy, preference and hierarchy between terms.
That is building a

thesaurus

p **A *concept browsing interface:***
- **allows the user to identify the descriptors**
- **creates a communication mechanism to fit the key concepts**
- **takes advantage from the availability of a high quality thesaurus.**

- **Some prototypes:**
  - **CANSEARCH (Pollitt, 1983)**
  - **CALIBAN (Frei and Jauslin, 1983)**
  - **COALSORT (Monarch and Carbonnell, 1987)**

Proteus

p Aims:

- **easier interaction with conventional systems**
- **interaction with a "conceptual document", independent from the physical implementation**
- **formulation of the query in ISO standard language**
- **tools for the identification of the relevant concepts**

[

more effective queries

p Tools:

- **Dictionary manager**
- **Graphic browser on structured thesauri**
  (graphical representation of the concepts and of their hierarchy, synonymy and preference relationships)
- **Management of special fields**
- **Windows for the insertion of the query values**

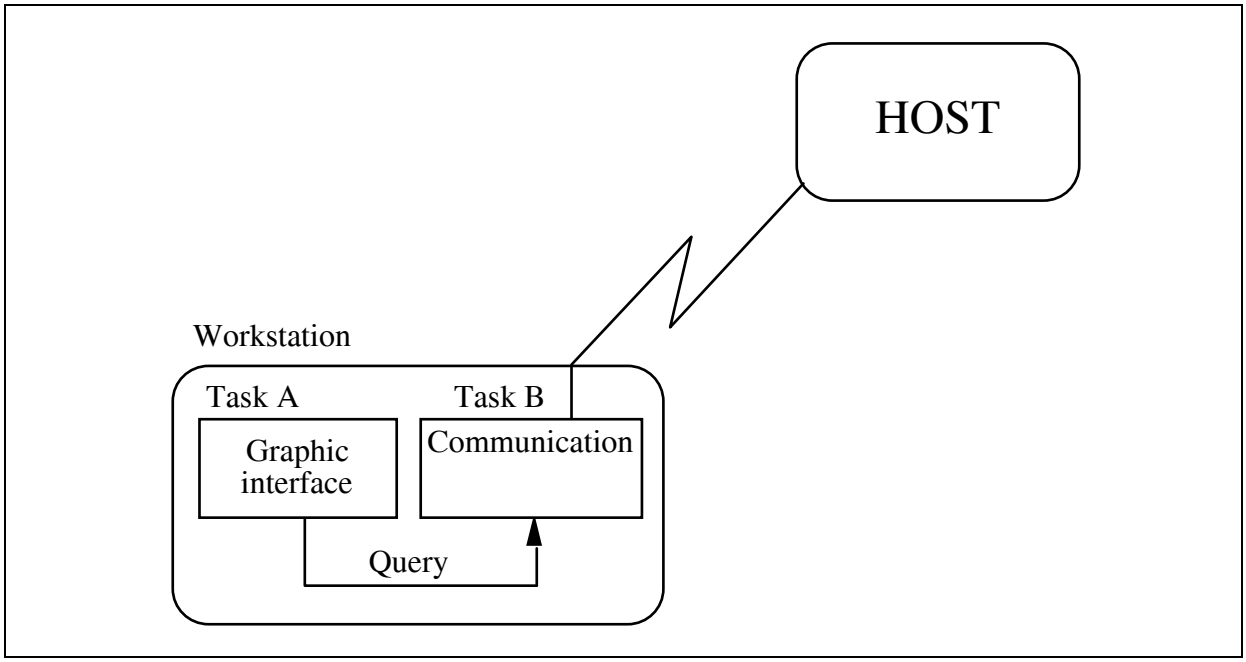
General architecture

p **The basic idea:**

- **prepare the query in the ISO standard syntax**
- **translate it into the query language of the system available on the host the user is connected to**

[

the interface and query formulation support functions are separated from the data bank query phase

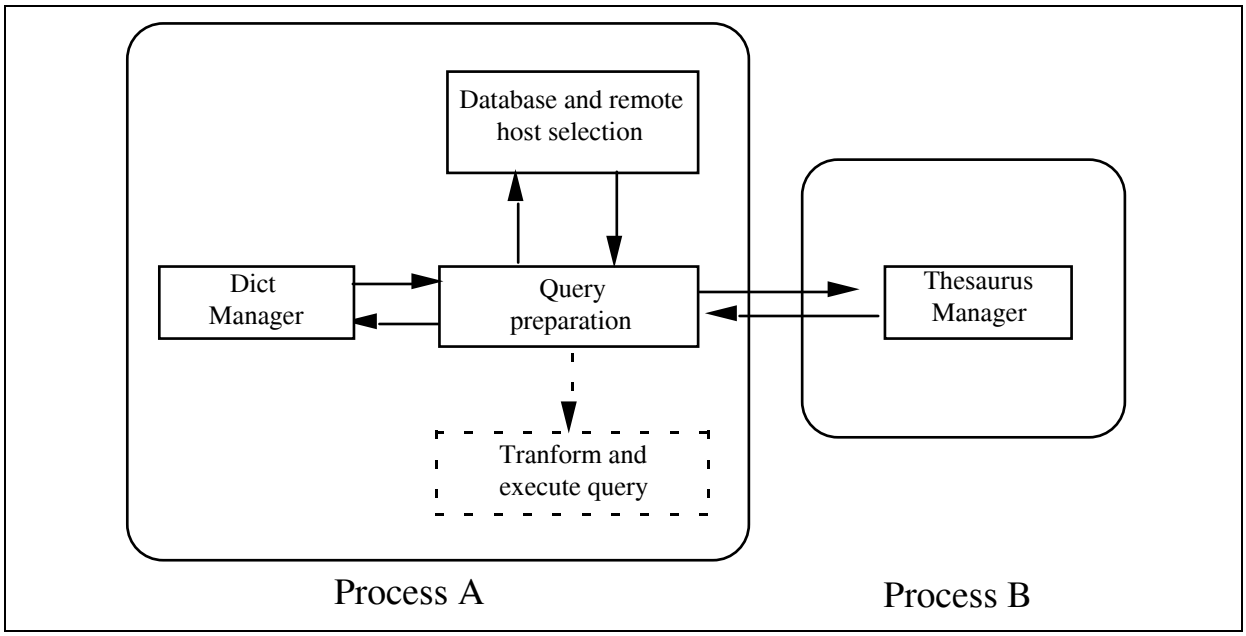


Two tasks

- **Adaptable to every Information Retrieval System**
- **Available on widespread personal computers**
  (PS/2, running OS/2)

The software architecture

- **Two different programs (A and B)**
- **Their executing instances are the process A and the process B.**
- **For each thesaurus a thesaurus manager has been coded, but they have the same basic structure.**

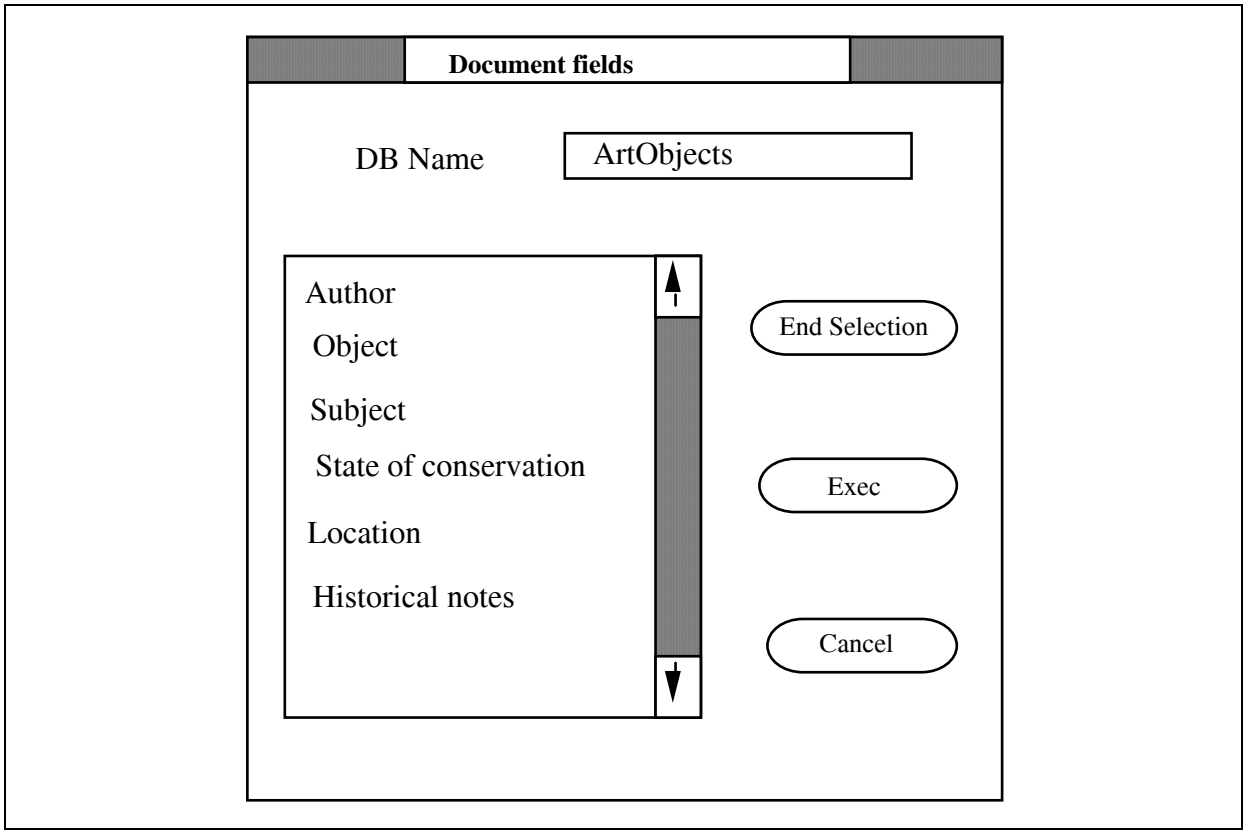


Two processes

(A can activate B and both can continue in parallel)

- **Multitasking implemented both at the process level and at the thread level.**
  Each process is made by two execution threads:
  - **a *first thread* receives the user input and passes the time consuming tasks to the second thread;**
  - **a *second thread* executes the time consuming tasks (as a series of queries to the support database) avoiding locking the system waiting for the completion of them.**
- **Processes are syncronized by semaphores, threads are syncronized by an exchange of messages.**

The “conceptual document”



- **Four different types of field:**

  Text *free text field*

  Cod *the field takes values that are a coded representation of the perceived values*
  (splitting in several subfields, codified values, etc.)

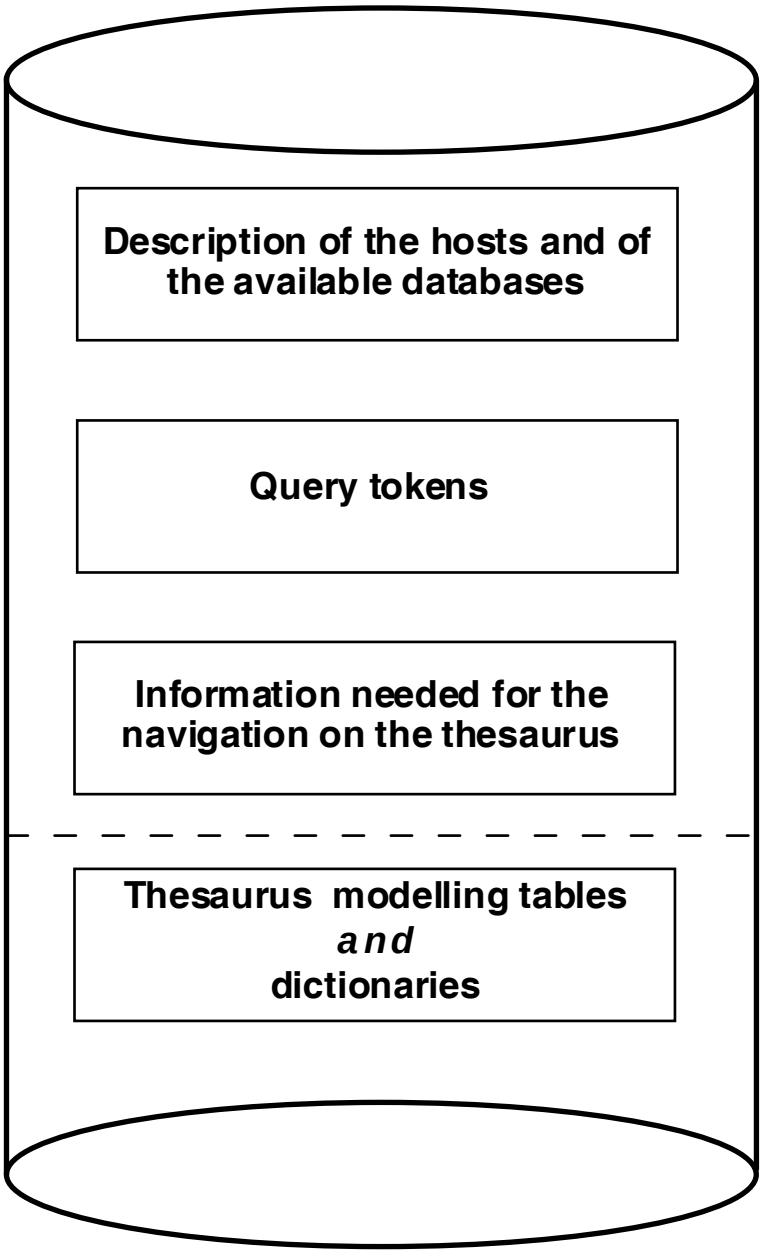
  Dict *the field can take values from a definite list of values*

  Thes *the field values are semantically arranged according to their relationships of preference, equivalence or hierarchy*

- **Others (*picture, sound, etc.*) can be added**

The supporting database

- **Three groups of tables (absolutely general)**

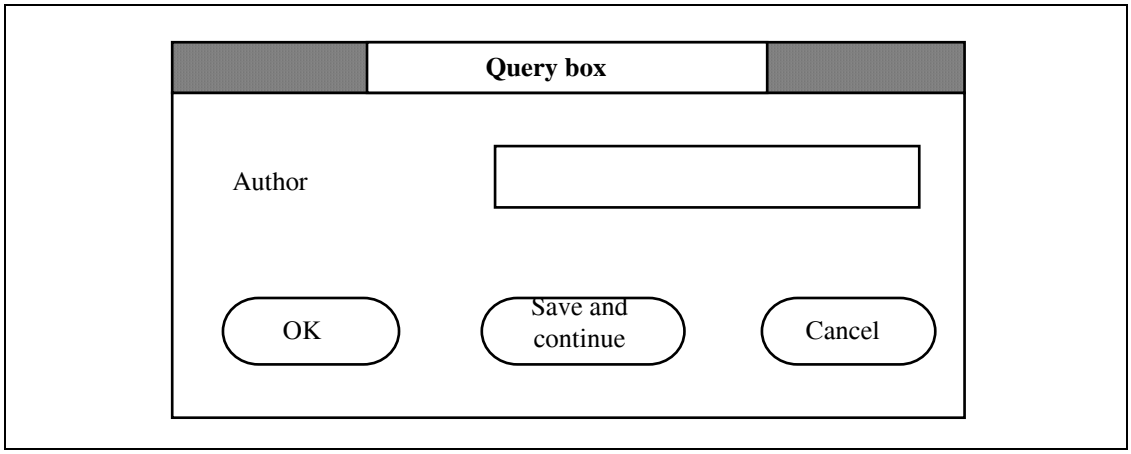


- **The thesaurus modelling tables represent the semantic relationships between the terms**

The query formulation phase

- **The user can get a description of any document field**

- **Field values can be entered in the appropriate window**



- **Free text field values are accepted as entered**

- **Coded field values are translated by a field dependent module** *(user coded)*

- **Dictionary controlled fields are automatically checked or the values are choosen from a multiple selection window**

- **Thesaurus controlled fields are automatically checked.**
  If the user needs to refine the concept he/she is interested to, a graphic browsing on the thesaurus can take place.

Two sample cases

p **ICONCLASS (Iconographic Classification System):**

- **a classification system set up by the Leiden University (Holland) between 1944 and 1984**
- **used for the description of the iconography of the western art (about 30.000 subjects)**
- **keywords are attached to the different subjects, and the code supplies a good linguistic independence**

ICONCLASS is tree structured

p **GEODOC:**

- **a multilingual thesaurus agreed by the "Commission on Geological Documentation" (COGEODOC) of the IUGS (International Union of Geological Sciences) as a common terminological thesaurus for Earth Sciences, to be adopted by different documentation centres.**
- **More than 5.000 terms.**

GEODOC is a multi-tree thesaurus

ICONCLASS: a typical hierarchy

Code	Subject
1	Religion and magic
10	(Symbolic) representations in relation to creation, cosmos, cosmogony, universe, and life (in the broadest sense)
11	Christian religion
11 A	Deity, God (in general) in relation to Christian religion
11 A 1	God the creator
11 A 11	God measuring the Universe (with compasses)
11 A 2	Divine Nature
11 A 21	Divinity
11 A 22	Symbols in relation to Divine Nature
11 A 22 1	Cyrcle symbolizing God's perfectness
11 A 23	God's perfections
11 A 3	God's wrath
11 A 31	'Flagello di Dio'
11 B	Father, Son and Holy Gost in relation to Trinity
...	...
11 C	God the Father
..	...
11 D	Christ
...	...
11 E	The Holy Gost
...	...
11 F	The Virgin Mary
...	...

11 G	The Angels
...	...

The graphic display of the thesauri

p **The Thesaurus Manager exhibits a different behaviour depending on the specific thesaurus**

p **The basic functionalities are the same:**
- **the thesaurus is represented as a graph**
  (the nodes are the thesaurus terms, the arcs are the connecting relationships)
- **the arcs can be labelled with the names of the relators**
- **"point and click" interaction style**
- **three levels are shown**

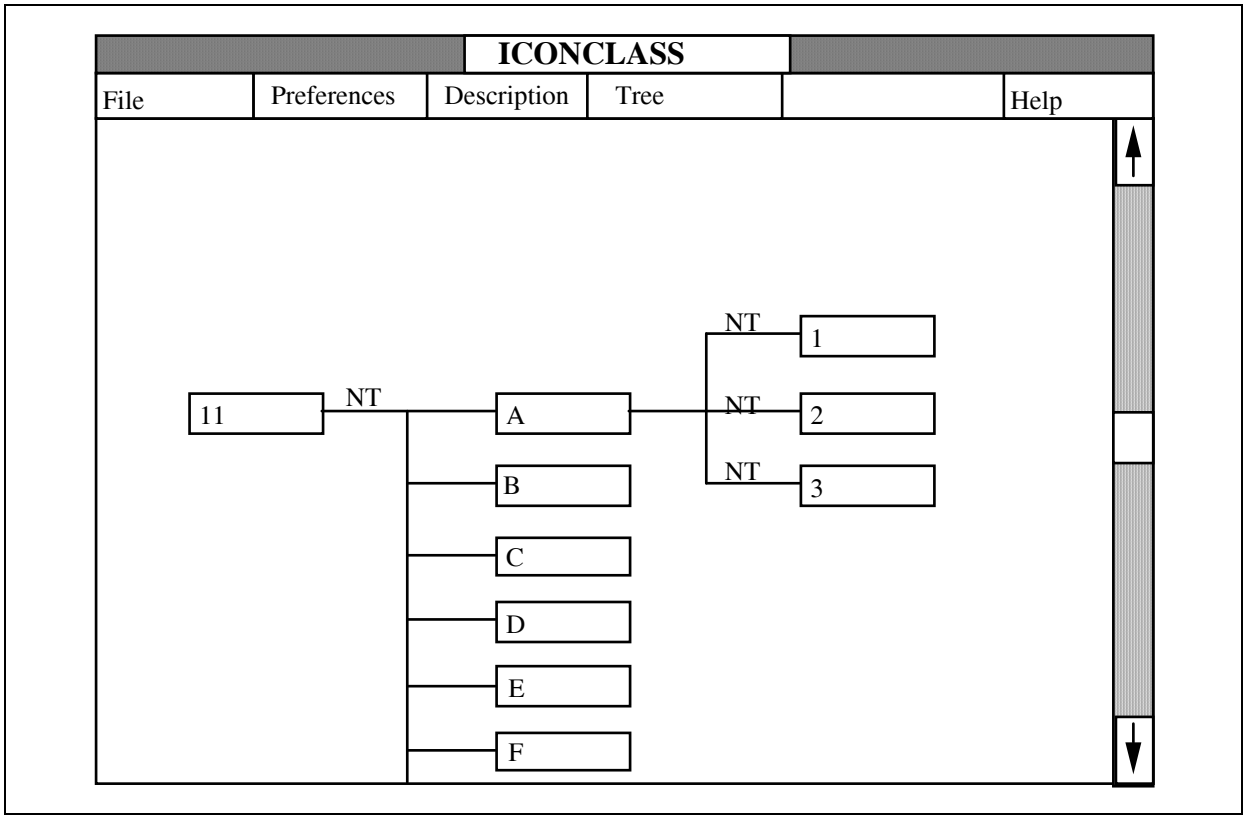
p **Multi-tree thesauri are displayed as a “butterfly display”:**
- **the central level is always constituted by a single node**
- **on the right the domain**
- **on the left the codomain**

p **The user:**
- **identifies the kind of relationships he/she is interested in**
- **selects a term (either directly, or from a list box)**
- **gets the display of a neighbour of the term**
  (only the selected relationships are shown)
- **can use the various options available from the menu bar**
- **can set some default (*“double click”*) actions**
- **can move around the structure**
  (and ask for detailed explanation of the terms)
- **can extend the tree towards the root or towards the leaves.**

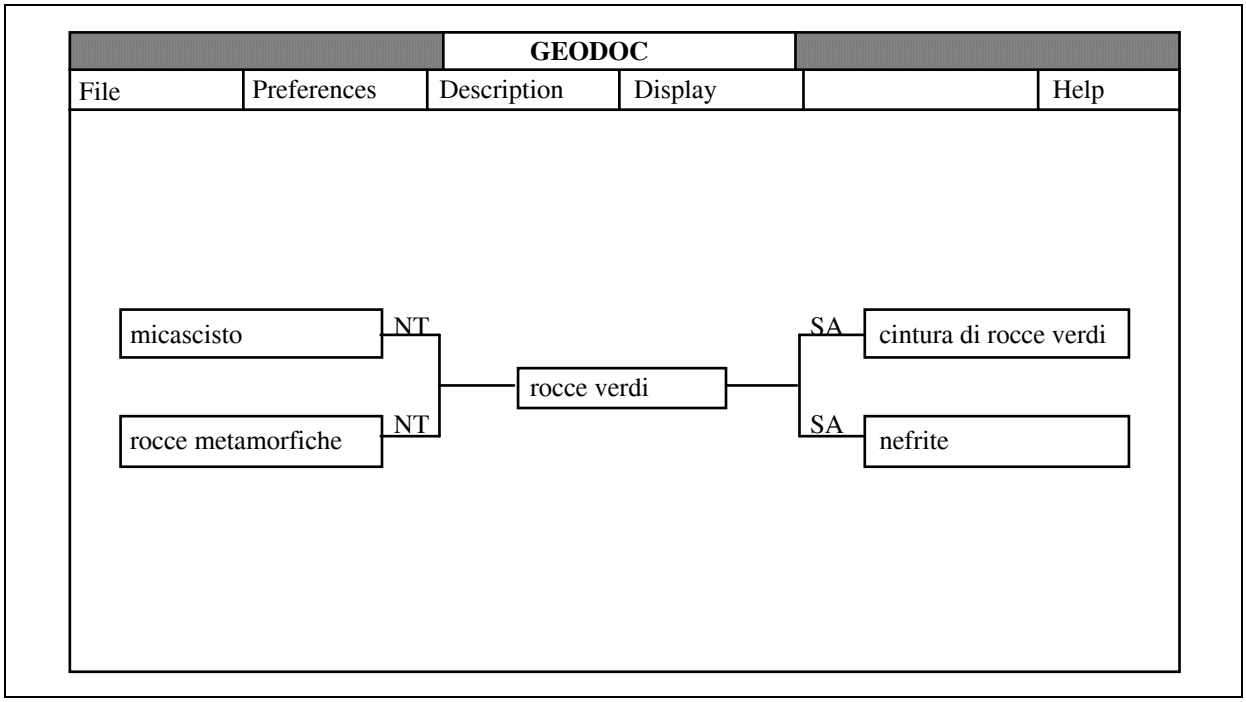
p **Any term can be selected in order to build a query to the database: all the terms which are selected in a single navigation will be OR'ed.**

p **The colours emphasize the role of the nodes**

The two different displays



The graphic display of the neighbouring of the code '11 A'



A “butterfly display” for the GEODOC thesaurus (italian version)

The query

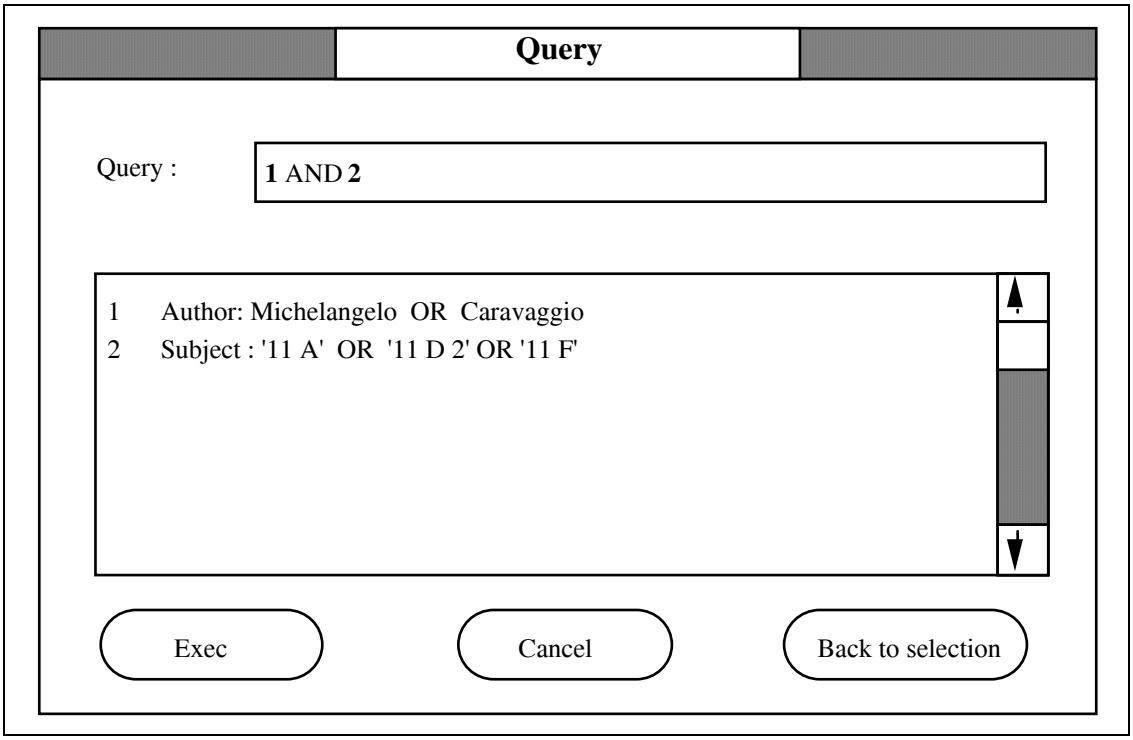
p **The tokens which constitute the query are stored in table.**
The query:

> **(Author : Michelangelo OR Caravaggio)**
> **AND**
> **(Subject : '11 A' OR '11 D 2' OR '11 F')**

is stored as:

Field	Value
Author	**'Michelangelo' OR 'Caravaggio'**
Subject	**'11 A' OR '11 D 2' OR '11 F'**

p **The resulting query may be edited**
(presently only the boolean logic can be modified)



Conclusions

- **Document retrieval requests the exact knowledge of concepts to be found**

- **Proteus makes easier to formulate queries on a document:**

  - **conceptual structure of the document**
  - **automatic check against the available dictionaries and thesauri.**

- **The thesaurus manager gives a snapshot of the structure of the thesaurus in a neighbour of the term of interest, and the user is allowed to navigate in the structure. The navigation on the thesaurus may be accomplished *in parallel* with the formulation of the query on other fields of the document.**

- **The interface is totally independent from the databases, only the thesaurus manager has an intrinsic dependence on the thesaurus structure.**

- **The interface agrees with the SAA standards**

- **While the novice user can receive help in the formulation of the query, the expert user can use some shortcuts in order to speed up the process.**

Future work

Some extensions are under study or development

- **modification of the tokens of the query**
- **implementation of the connection with the remote host according to the CPI standards, using APPC**
- **the query will be saved for future use**
  (establishment of a user profile)
- **multitasking navigation on several thesauri, using appropriate semaphores**
  (but this enhancement seems not to be really useful for the user, who should follow several classification schemas in parallel)
- **use of the Thesaurus manager as a tool for the manipulation of structured thesauri**
  (addition of terms, translation, moving of a set of terms, establishment of new relationships, etc.)
- **parametric representation of the thesaurus structure**
  - **implementation of a single thesaurus manager, which will produce a display that will vary according to the structure of the thesaurus it is processing**
- **implementation in a "client-server" architecture:**
  - **a server will host the support database, and will implement the connection with an external host.**