

SEMINAIRE

E+; 3=!C

SEMINAR

STATISTICAL COMMISSION AND
ECONOMIC COMMISSION FOR EUROPECONFERENCE OF EUROPEAN
STATISTICIANSDistr.
GENERALCES/SEM.38/7 (Summary)
24 April 1998

ENGLISH ONLY

Seminar on integrated statistical information
systems and related matters (ISIS '98)
(Geneva, Switzerland, 27-29 May 1998)

Topic (i): use of communication technologies for external and
internal dissemination of statistical data

SEARCH ENGINE FOR STATISTICAL DATABASES

Submitted by the United Nations Statistical Division

SUMMARY

1. The paper describes a programmatic method for indexing and searching the textual information contained in any of the description columns or metadata tables of a statistical database. Implementation of such a tool is especially relevant when databases with statistics are made available or published on the Internet. It provides an alternative to the more traditional and structured navigation of leading the user through a series of hierarchical selections.
2. Using the relational model, statistical databases are often designed with a single, multidimensional fact table and several reference tables, one for each dimension of the data or fact table. Using the entity/relationship methodology, such databases are usually modeled with "star" schemas. The database model consists of one fact table, often quite large, surrounded by much smaller dimension tables, each one contributing key attributes to the fact table.
3. The idea of the search engine is to prepare an "index" table, containing a row for each word contained in the description columns. The index table will also contain the series code of the series where the word occurs and a unique code of the word.
4. The described technique is applicable to any relational table with a textual or description column such as reference tables and footnotes tables. It is easy to implement and is particularly well suited to Internet and Intranet settings.
5. The approach described here is one strategy which can be combined with others to offer Internet database users an "intelligent basket" of search tools. It is "free form" and text based and has two important strategic advantages:

¹ Prepared by Lubomir Vitkov.

(i) it easily adapts itself to any expansion of the database since it can readily be programmed to search any parts of the site or database whatever its rate of expansion;

(ii) it is easy to combine with various "translation tools" such as key words, synonyms, foreign language versions and so on, such that a search for "voitures" can equally hit on "automobiles" or "motor vehicles".

6. No single Internet search and navigation system for statistics will satisfy all user demands but the present version proposes one strategy, which combines an "open door" policy with behind the scenes programming to steer users in the most fruitful directions.