Development of an Automation Software for Reconciliation of INIS/ETDE Thesauruses

Manoj Singh, Rajiv Gupta, E. R. Prakasan and Vijai Kumar

Library and Information Services Division
Bhabha Atomic Research Centre
Trombay, Mumbai – 400 085.

Abstract

ETDE (Energy Technology and Data Exchange) and INIS (International Nuclear Information System) thesauruses contain nearly twenty thousand descriptors and are not necessarily identical. A project has been undertaken by the international organisations to make a common thesaurus for both INIS and ETDE to facilitate better exchange and retrieval of information between/from these databases.

This paper describes the automation implemented during our participation in the project for Reconcile the Structures of the Word Blocks in the ETDE and INIS Thesauruses, with respect to the descriptors currently in the two thesauruses through a PC based RDBMS Software. The Software THEMERGE was developed in FoxPro 2.5 Relational Database Management Systems. The software handles all possible reconcile recommendation suggested by specialist, printing the recommendation sheet for uploading it later. This has not only widened the scope of flexibility, portability and convertibility of recommendations, but also helped to achieve quicker project completion.

Keywords: INIS, ETDE, Thesaurus, Software, Database

1 INTRODUCTION

A thesaurus is defined as a terminological control device used in translating from the natural language of documents, indexers or users into a more constrained 'system language' (document language, information language). It is also a controlled and dynamic vocabulary of semantically and generically related terms, which covers a specific domain of knowledge.

International Energy Subject thesaurus of Energy Technology Data Exchange (ETDE) and International Nuclear Information System (INIS) thesaurus, contain the standard vocabulary of indexing terms (descriptors), developed and maintained by the respective agencies. INIS and ETDE have some arrangements to share the energy-related inputs to their databases.

India is a participant to the INIS/ETDE thesaurus conciliation project. The word blocks starting with N and O are undertaken by India to make necessary recommendations for INIS/ETDE thesaurus conciliation. For this purpose we developed the software THEMERGE (Thesauruses Merge) in FoxPro Relational Database Management System (RDBMS).

This software can also make some sort of intelligent suggestion while concentrating on the terms of recommendation. The software is very user friendly with Help Key at various levels of software operations. It has also the feature showing the Help in the form of pop down menus at various stages. Basically this software consists of two modules, one for data entry module for recommendations and other for printing the recommended data sheet for uploading later.

ETDE is a consortium of member countries around the globe that share their energy research and technology information through ETDE's database. The Exchange was established in 1987 under the auspices of the International Energy Agency (IEA) and serves all ETDE member countries. The Exchange also collaborates with other IEA as appropriate.

INIS is the world's leading information system on the peaceful uses of nuclear energy. It is operated by the International Atomic Energy Agency (IAEA) in Vienna, Austria. The IAEA is an autonomous organization within the United Nations in collaboration with its Member States and co-operating international organizations.

To facilitate better exchange and retrieval of information between INIS and ETDE databases, there is a need for a common thesaurus. To build a common thesaurus some of the member countries were allotted a part of a thesaurus corresponding to words starting with particular alphabets to make recommendations by the project coordinator.

India has undertaken the thesaurus terms/ word blocks starting with letter N and O. Graphical view of the work assigned to different countries towards the thesauruses reconcile project is given in Fig. 1.

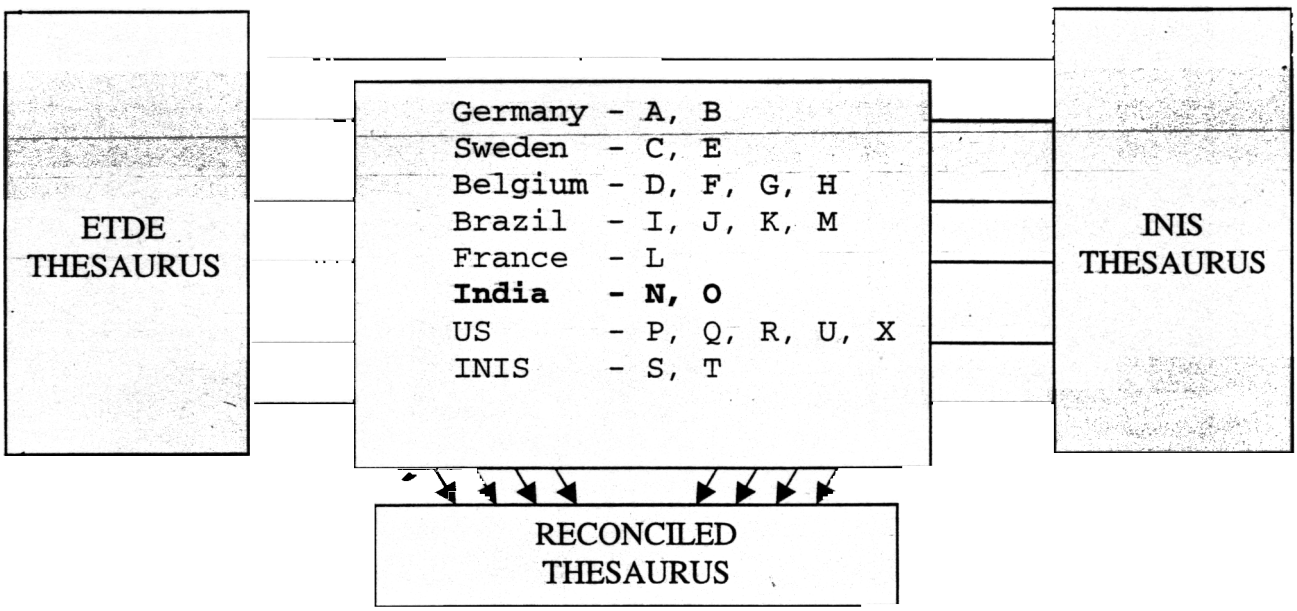

Fig. 1

2. Data Flow Cycle

`The flow of data during the project is shown in Fig. 2. Data initially starts from the FTP server form FIZ-Karlsruhe, Germany in HTML format then converted into DBF form for the software THEMERGE and finally into text form of recommendation data sheet at our end.

Data Flow Cycle Of The Project

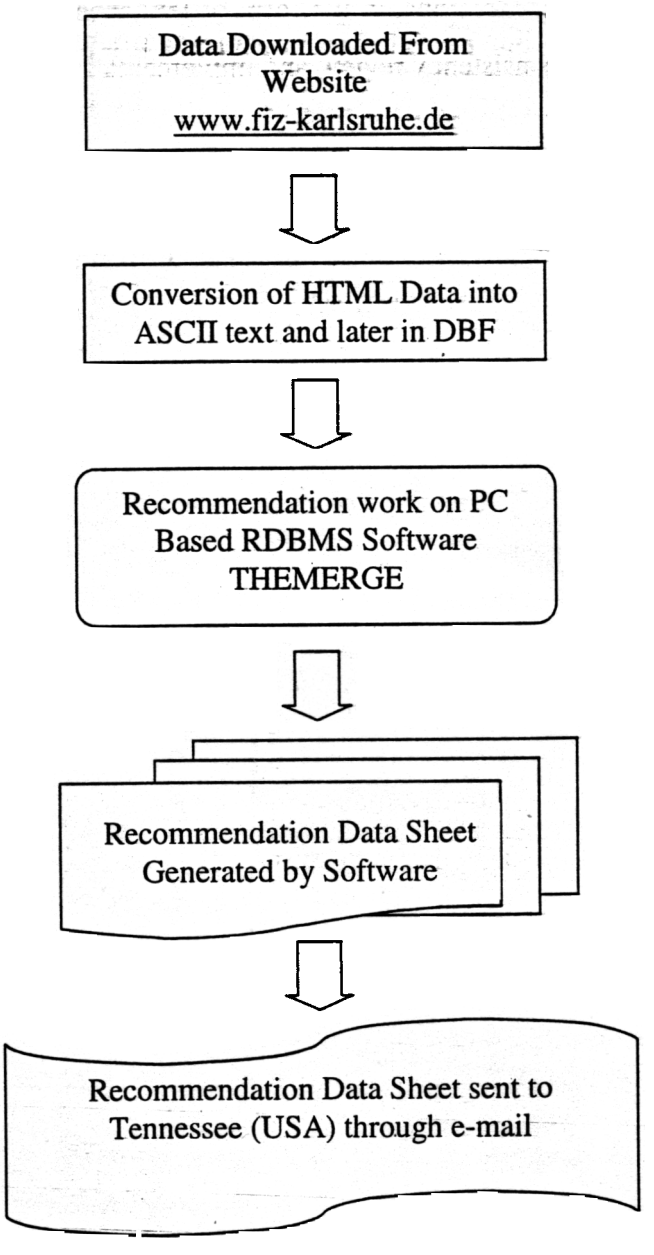

Fig. 2

The FIZ-Karlsruhe, Germany has compiled both the thesauruses for the difference in valid terms using INIS thesaurus as base and prepared 26 HTML files for each of the alphabet.

The HTML files downloaded for the word block N and O have been converted into database files *MASTN.DBF & MASTO.DBF* using offline procedures *BIG.PRG*

& MAJMIN.PRG. Once the database is created, the specialist can operate the software for inputting his recommendation as per guidelines for the valid difference in terms.

Results of the recommendations, in the form of text sheet, will be sent to the ETDE Operating Agent for consistency review and implementation.

3 Hardware & Software Requirements

- Pentium with 16 MB RAM with 3.5 inch floppy drive
- Color or Monochrome monitor
- 1 MB disk space for software installation
- DOS 6.22 or WINDOWS 95/98
- FoxPro 2.5 RDBMS

4 Data Preparartion

Various steps for data preparation involved before using it for the software are listed below.

Step I: Downloading of HTML data files form FTP server located at Germany.

Step II: Conversion of HTML data files into ASCII text files.

Step III: Creation of database structure and appending ACSII data into the database file.

Step IV: Codification of records for the purpose of data retrieval by codes. All the main descriptors are uniquely codified and each of its word blocks is codified with respect of it.

Step V: Modification of each record for ensuring error free data for the software operations.

5 OPERATIONAL ASPECT OF SOFTWARE

5.1 Software Installation

The software THEMERGE consists of 22 files which include programe, database and executable files. During the execution of the program two text files for word blocks starting with letter N & O are generated.

All the files of the software THEMERGE are installed in the same directory where FoxPro has been installed on the PC.

5.2 *OPERATIONS: Data Entry Module*

Execute the FoxPro RDBMS and from the command prompt of the FoxPro, run THEMERGE program file. Select the module for *Recommendation data Entry* from the main screen of the software and select the letter N or O for recommendation data entry.

The software takes care for last record updated and will prompt for next automatically. One can select another term for work either by Help Key or inputting the main descriptor code.

The specialist can choose one of the recommendations for main descriptor from the active pop down menu. Once the selection of recommendation for main descriptor is over, the software filters all-possible valid terms for recommendation for the selected main descriptor in a pop down menu.

Now select an entry from the word block pop down menu, the software examines the entry and suggests some hints for its recommendation. Next one can save or abort the recommendation into the database file. If opted for saving the recommendation, it will elbow to the term within the pop down menu. This elbowing of recommendation will give a quick look of the saved recommendation against each term in the pop down menu.

The work for another term of the main descriptor can be continued or aborted. If aborted at this stage, then whenever one starts working, the software takes care of the incomplete word block of the main descriptor. It is not mandatory to complete the word block of a descriptor in one stretch but the software will track all incomplete terms of the descriptor automatically.

One can change the recommendations at any level of the entry by using its code. The unique codification of each entry enables to edit any recommendation at any stage which is another, flexible feature of the software. The flow chart of this module is shown in Fig.3.

Flow Chart For Data Entry Module

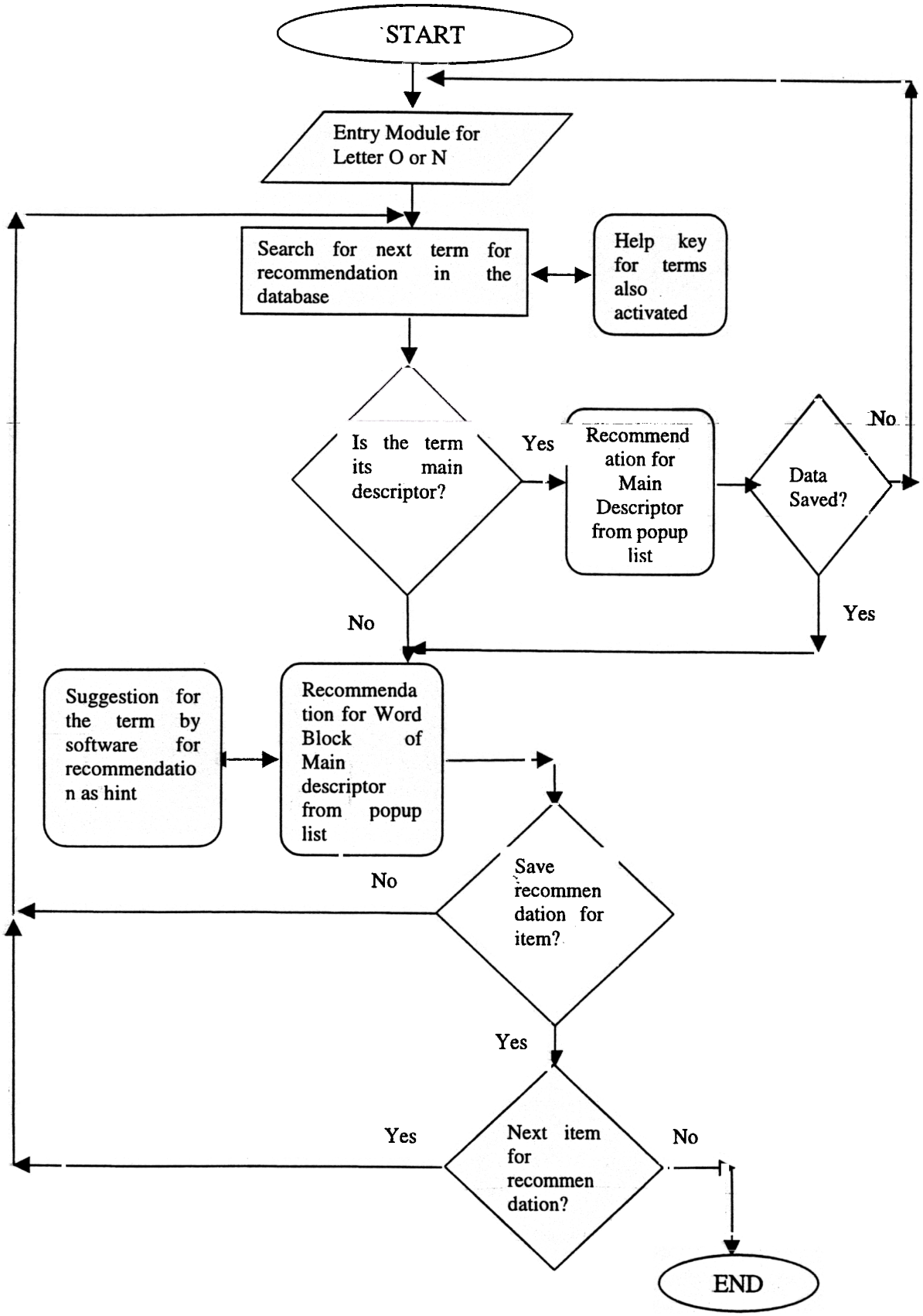

Fig. 3

5.3 OPERATIONS: Print Module

Once the data entry work for recommendation is over, the software can generate the recommendation data sheet by using FoxPro report Menu feature.

For this one has to select second option from the main screen of the software *Recommendation Sheet Print* module and then choose one of the letter works.

Once the letter is selected, the next screen prompts for the first and last descriptor of the word block to be printed. Provision is also made for selecting first and last descriptor from Help menu for printing.

After selecting starting and ending number of main descriptor, the software processes on the selected range of descriptor and generates an output file *RECN.TXT* or *RECO.TXT* as recommendation sheet as shown in Fig. 9. Each recommended term is indicated with an arrow symbol as per the desire of project coordinator. The flow chart of this module is shown in Fig. 4

6 Statistical Data

The total descriptors for letter N are 370 containing a total of 10,410 terms Similarly total descriptors for letter O are 227 containing a total of 6488 terms. Corresponding database files *MASTN.DBF* and *MASTO.DBF*, thus, contain 10,780 and 6,715 records respectively. The software handles these two database files with a total number of records 17,495.

Flow Chart for Print Module

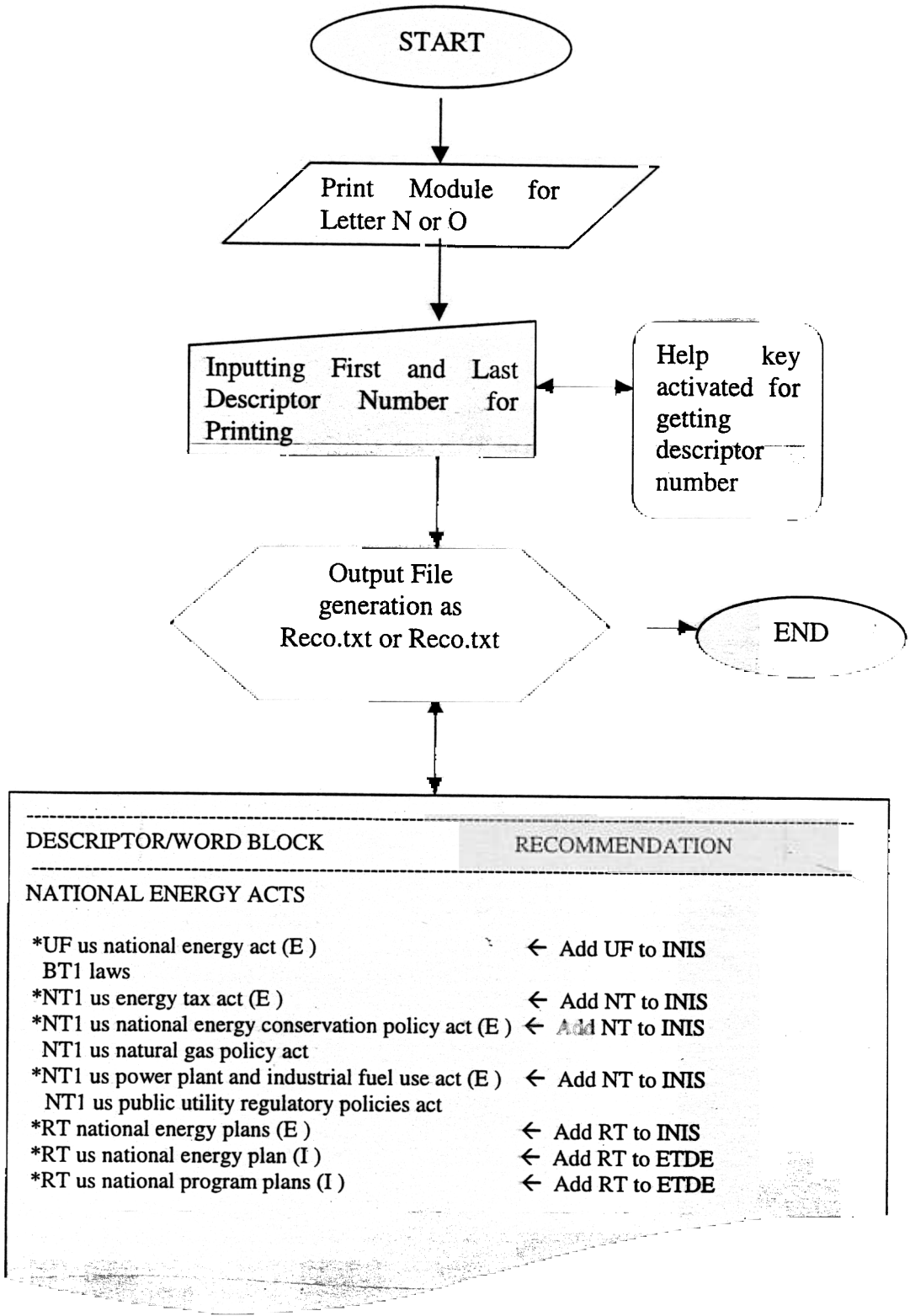

Fig. 4

Table 1 and 2 shows the different type of valid terms processed by the software for letter N and O respectively and bar chart view in table 3.

	NT1	NT2	NT3	NT4	RT	BT1
Letter N	230	3087	670	10	469	199
Letter O	160	311	289	118	447	100

Table 1

	UF	SF	USE	AND	OR	SEE
Letter N	282	27	170	23	04	12
Letter O	96	08	85	04	02	03

Table 2

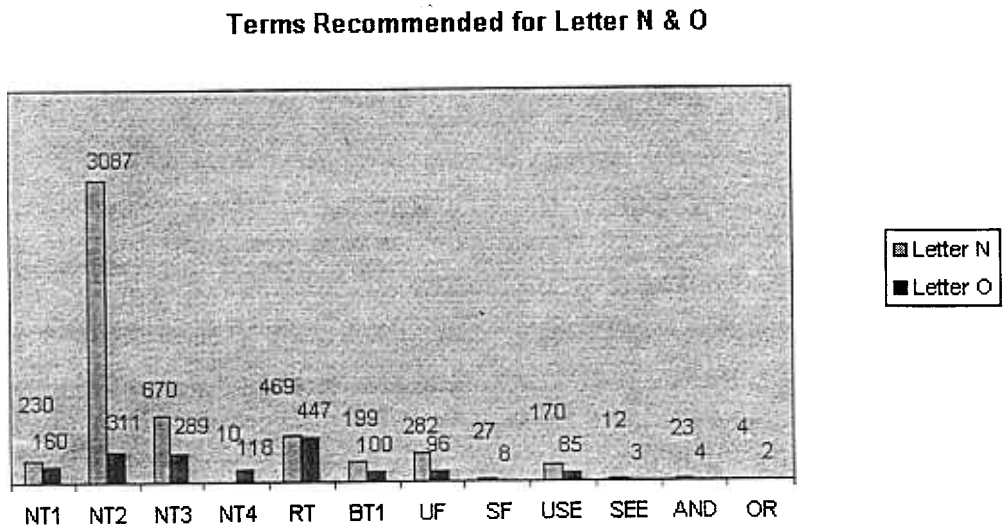

Table 3

7 CONCLUSION

The software THEMERGE developed by us gave enough flexibility to retrieve, edit and make recommendations for the thesaurus reconciliation project in a systematic and phased manner. The portability of data is also tested by converted data sheet into standard documentation format such as PDF, HTML and other formats of data exchange. It is planned to upgrade the software for the usage in client server platform and also to utilise it for future reconciliation projects.

8 ACKNOWLEDGEMENT

We are grateful Dr. M.K.V Nair, Scientific Officer (E), Library & Information Services Division, BARC for taking keen interest throughout the course of this work.

9 REFERENCES

1. Carl Townsend, "Mastering dBase IV Programming"
2. Daniel Martin, "Advance Database Techniques"
3. Ed. K.L.Clark & S.A.Tarnlund, "Logic Programming" (APIC studies in Data processing No. 16)
4. Ed. Won Kim, David S. Reiner & Don S. Batory, "Query Processing"
5. "INIS: Thesaurus", IAEA-INIS-13(Rev. 37),Jan-1998
6. "International Energy Subject Thesaurus", ETDE/PUB—2 (rev.2) (DE97009551)
7. IAEA Web Site http://www.iaea.or.at
8. ETDE Web Site http://www.etde.org