ACHIEVING HOMOGENEITY IN HETEROGENEOUS KNOWLEDGE BASED SYSTEMS: A CASE STUDY

Sangeeta Deokattey
Manoj Singh
Vijai Kumar
Anil Sagar
Anil Kumar
Scientific Information Resource Divn.,
Bhabha Atomic Research Centre,
Trombay, Mumbai-400 085.
sangeetadeokattey@hotmail.com

Pratibha A. Gokhale
Reader, Dept. of Library and Information Science
University of Mumbai, Mumbai-400 098)

Abstract

The paper attempts to address the problem of interoperability due to semantic differences, in various databases. A single platform was developed to simultaneously search and retrieve data from three bibliographic databases, using the keyword mapping and linking technique. Accelerator Driven Systems, a multidisciplinary micro-subject of global importance, in nuclear energy production and nuclear waste management, was selected for the study.

Keywords:
Accelerator driven systems; DBMS; Information management; Interoperability; Microthesaurus; Nuclear energy; Radioactive waste management

1. Introduction:

The variety and amount of digital information that a user has to confront today, is daunting in terms of different formats, types of indexing and many other variations, found both in structured as well as unstructured databases. If one adds to this the whole gamut of information available on the web, the user may find himself in a state of chaos, searching for the right piece of information, needed by him.

Librarians and information specialists have been devising ways and means of overcoming this problem of searching different sources of digital information, using a single platform [1]. Standardization, in terms of metadata, for either a printed document or an electronic document, has been achieved, to a great extent, through cooperative ventures such as OCLC and Dublin Core. But in the

case of subject indexing, standardization across boundaries remains a distant dream. Integration of heterogeneous sources of digital information requires dealing with controlled vocabulary terms (from a thesaurus), subject headings (from subject heading lists), classification numbers (from various general and special classification schemes), simple lists of keywords or full text terms (as is done in the case of web pages, since they are not normally indexed). A perfect match between a user's query and the text being searched is not possible, as different texts have concepts or terms, which are at different levels of abstraction and representation [2]. And as is well known, all representations are imperfect approximations to reality [3]. Therefore, the best possible approach to solving the problem of semantic heterogeneity is to have some kind of vocabulary mediation or transfer module [4,5], which will facilitate a basic level of interoperability, among different digital resources in a library. This could bring us closer to a one-stop method, for searches in integrated systems and save the time and efforts of the user.

2. Accelerator Driven Systems (ADS): Our Universe is made up of a small number of basic building blocks called elementary particles (electrons, protons, neutrons and other subatomic particles), which are governed by some fundamental forces. Some of these particles are stable and form the normal matter, whereas others live for fractions of a second and then decay to the stable ones. Particle accelerators give high energy to these subatomic particles that are made to collide with different targets. Out of these collisions, many other subatomic particles are generated which pass into particle detectors. From the information gathered in these detectors, physicists determine properties of these particles and their interactions. The higher the energy of the accelerated particles, the more closely the structure of matter can be probed. There are about ten thousand particle accelerators in the world today. More than half of them are used for medical purposes. Apart from basic research, particles are being used in a variety of ways for the benefit of mankind. Cancer therapy, medical and industrial imaging, radiation processing, electronics, measuring instruments, new manufacturing processes and materials etc. Basically, an Accelerator Driven System (ADS) is a hybrid system. It is a combination of an accelerator and a nuclear reactor, used for specific purposes. Accelerator driven systems are being used for energy production, transmutation (breaking up of long-lived radioactive waste particles into short-lived harmless waste), i.e. in high level radioactive waste management and for breeding of nuclear fuel.

3. Materials And Method

Since ADS is still an emerging field, there are no specific tools for Information Storage and Retrieval, on this subject. Information specialists have to rely on the

INIS thesaurus, which is very broad in scope. Three international CDROM databases INIS, INSPEC and CA were selected for the study. Searches were conducted on these three databases, using the following descriptors for query formulation.

ATW
Accelerator Transmutation of Waste
Accelerator driven*
Accelerator driven systems
ADS
ADEP
Accelerator driven energy production
ADSS
Accelerator driven subcritical systems
ADTT
Accelerator driven transmutation technology*

(contd. on next page)

AAA
Advanced accelerator applications

After eliminating duplicate and spurious records, a total of 2336 unique records were retrieved. Fig.1 shows the percentage distribution of records in the three databases and Table 1 gives the exact breakup. Fig.2 gives a representation of the growth of publications on the subject over the years.

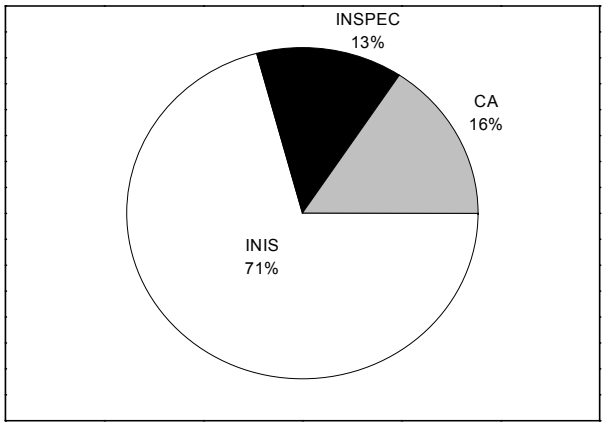

Fig1: Database-wise breakup of publications on ADS

Table 1: Distribution of records in the three selected CDROM databases

Database Name	Year	Records retrieved	Records deleted	Final records
INIS	1970-2003	1904	250	1654
INSPEC	1969-2003	808	493	315
CA	1966-2003	809	442	367
Grand total	> 30 years	3521	1185	2336

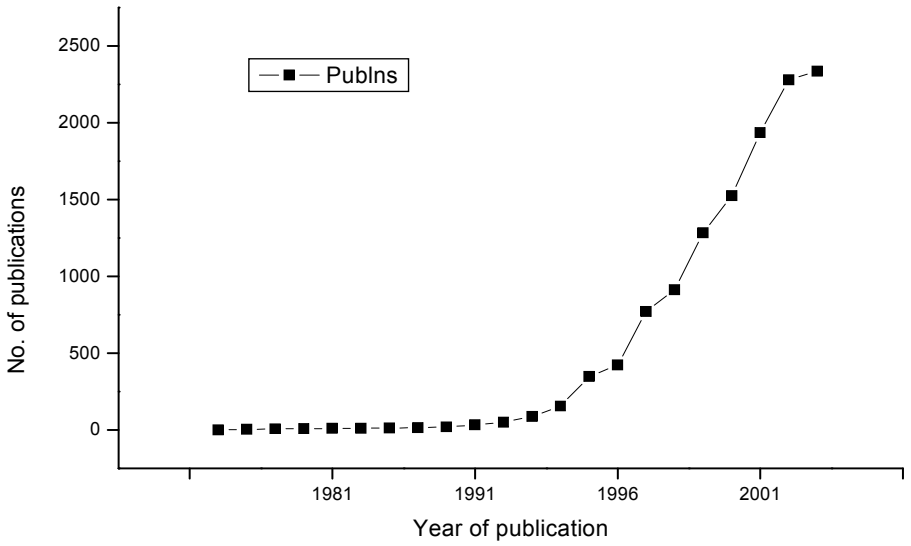

Fig.2: Cumulative growth of publications in the area of Accelerator Driven Systems

All the 2336 bibliographic records were downloaded onto the PC. The keywords or descriptors assigned to each document were downloaded separately, database wise and sorted alphabetically. Thus three files of descriptors/keywords were created as Excel worksheets and the frequency of occurrence of each descriptor in the records was also counted. Table 2 shows the breakup of descriptors in the three databases.

Table2: Distribution of descriptors in the selected databases

Database Name	No. of unique descriptors	No. of orphan descriptors
INIS	1872	102
INSPEC	312	28
CA	599	11

3.1 Creation of a separate file of keywords that were not a part of the controlled vocabulary of the three selected databases:

The use of accelerators for ADS is an expensive proposition. Therefore a lot of national and international collaborative R&D programs and projects are being carried out worldwide. Dissemination of information about the status of these projects is important for the scientific community [6]. To address this problem, appropriate keywords from the abstracts of 1654 INIS records were selected [7,8]. All the INIS abstracts were downloaded separately and saved as a text file. A stop-word file of 820 words was created and a small program in PERL was written and run to eliminate all the stop-words from the abstracts. A minimum threshold level of 2 was maintained in the selection of the keywords. 302 such keywords were identified.

Thus, a total of four EXCEL files of keywords were created. The INIS file was considered as the master file, as users in BARC are familiar with the INIS database and search strategies.

3.2 Intellectual mapping and linking of the descriptors/keywords:

This was a manual procedure. As a first step towards mapping and interlinking, descriptors from the INIS file were checked and matched against each other, to establish broader, narrower and related term relationships amongst them, based on the INIS thesaurus. A total of around 20,000 such linkages were formed among the 1872 INIS descriptors and an interlinked file of INIS descriptors was created. In the next step, the second INSPEC file of descriptors was matched and linked to the INIS descriptors, to create a second file of interlinked INIS and INSPEC descriptors. The third inter-linked file of INIS and CA descriptors was created, by matching each descriptor from the CA records with the INIS descriptors. A similar procedure was followed for the fourth file of keywords from INIS abstracts. Thus four interlinked files of descriptors/keywords on ADS were created. Subject experts were consulted for the verification of subject linkages. Certain keywords could not be matched to other descriptors and were thus retained as orphan keywords (orphan keywords could be separated and if necessary, eventually deleted).

3.3 Analysis of the selected descriptors or terms

In the course of manual checking and interlinking of the four EXCEL files, certain characteristics of the descriptors came to light. These can be seen in Table 3. None of the keywords/descriptors from CA were single-word terms. All of them were compound multi-word concepts. Therefore, more than one descriptor from INIS was assigned to each multi-word descriptor from CA.

Table 3: Characteristics of descriptors in the three databases

Types of terms	INIS-INSPEC	INIS-CA
Identical terms*	91	Nil
Partially identical terms**	45	Nil
Semantically equivalent but morphologically different terms***	62	Nil
Terms not in common	28	Nil

* e.g. corrosion protection; direct reactions; fuel element failure etc.
** e.g. accelerator driven transmutation and accelerator based transmutation; compression strength and compressive strength; radioactive wastes and radioactive waste etc.
*** e.g. physical radiation effects and biological effects of ionizing radiation; targets and nuclear bombardment targets; waste transportation and nuclear materials transportation

3.4 Tools and setup for software development

3.4.1 Input:

Downloading of text files from INIS and INSPEC databases (for INIS and INSPEC bibliographic tag in top to bottom field structure)
Downloading of text files from CA (non structured file descriptors)

3.4.2 Offline data conversion:

For INIS and INSPEC databases

Conversion of each record of top-to-bottom tag fields of downloaded data in Microsoft Excel using Active Server Pages scripting.
Save Microsoft Excel as tab delimited text file
Creation of database, table etc. for Microsoft SQL server

Import of tab delimited text file into INIS records table
PHP script with SQL server to extract keywords for each record into INIS-keywords table in relation to INIS table

For Chemical Abstracts database

Pattern recognition of bibliographic fields in downloaded text file data
Delimiting each record with unique identification
Creation of database, table etc. for Microsoft SQL server
PHP scripting to convert recognized pattern into CA SQL table
PHP script with SQL server to extract keywords for each record into CA-keywords SQL table in relation to CA SQL table

For Keywords from INIS abstracts

Saving of Microsoft Excel My-Keys file as tab delimited text file
Creation of database, table etc. for Microsoft SQL server
PHP script with SQL server to extract abbreviated keywords for each record into MY-keywords SQL table in relation to INIS SQL table

3.4.3 Processing

PHP script for displaying alphabetical list of INIS main descriptors with its broader, narrower and related term IDs.
Selection of INIS main descriptor will display all possible term linkages.
Selection option for database searching with respect to INIS, INSPEC, CA and Keywords from INIS abstracts
Based upon selection criteria, parameters passed to another PHP script for data searching and processing.

3.4.4 Output

Based on parameters & database selection, display of matched bibliographic data

For INIS and INSPEC

Sorting of bibliographic data with publication year for matched main INIS descriptors
Sorting of bibliographic data with publication year for matched broader, narrower and related terms

For CA

Sorting of bibliographic data with publication year for matched CA keywords with INIS descriptors

Keywords from INIS abstracts

Sorting of bibliographic data with publication year for matched Keywords from INIS abstracts

3.4.5 Client/Server Requirements

3.4.5.1 Server Side

Microsoft Windows 2000 with IIS 5 or higher on server class Pentium based machine
Server class hardware configuration
Microsoft Excel
Microsoft SQL server with Enterprise manager
PHP 5

3.4.5.2 Client Side

Desktop Pentium with Windows 98/2000 OS
Web browser preferable Internet Explorer

4. Creation Of A Dynamic Microthesaurus On ADS

This method automatically generates a dynamic web-based microthesaurus on ADS, which can be easily updated and maintained (provided the choice of the database and query formulation are comprehensive). On the basis of frequency of occurrence of the terms, new descriptors assigned to future ADS documents, in any of the three databases, can be added separately and linked to each other. New and potentially useful concepts, which are not part of the controlled vocabulary of the databases, can also be added to the 4th keywords from INIS abstracts file and similarly interlinked. Any number of such files can be created and linked to form a comprehensive, homogeneous, web-based search and retrieval tool.

To initiate a search through the user interface, the user selects a descriptor from INIS (all INIS descriptors are alphabetically displayed). Appropriate records from INIS, INSPEC and CA, which contain the selected descriptor (and its broader, narrower and related terms) are simultaneously searched and displayed in the search results. When a user selects an INIS descriptor for search, he/she can also see at a glance, all the other terms from the three databases, which are related to it. If the user wishes, he/she has the option of narrowing down the search, by selecting any other terms from the displayed list and retrieve records. Otherwise, by default, all the terms from all databases, related to the selected descriptor, are included in the search.

5. Conclusion

In the present web-based information scenario, with hundreds of databases and millions of records, there is very little scope for standardization of terminology. Data conversion from one system to another is a tedious and time-consuming process. The best option to make all these resources available to the users on a single platform, is through the development of a dynamic thesaurus or more

appropriately a dynamic dictionary of synonyms, in specialized subject areas of interest to the users.

6. Acknowledgements
The authors wish to acknowledge with thanks the guidance and advice received from Dr. R.C. Sethi, Head, Accelerator & Pulse Power Divn., Shri P.K. Nema, SO/H, Nuclear Physics Divn. and Dr. S.B. Degweker, SO/G, Theoretical Physics Divn., BARC. Thanks are also due to Mrs. A.A. Kadam, SO/D, Computer Divn., BARC for PERL programming.

7. References:

1. Chan, Lois Mai & Zeng, Marcia Lei: Ensuring interoperability among subject vocabularies and knowledge organization schemes: a methodological analysis. *68th IFLA Council and General Conference, Aug.18-24, 2002* (Classification and Indexing Group).
2. Qin, J.: Semantic similarities between a keyword database and a controlled vocabulary database; an investigation in the antibiotic resistance literature. *JASIS*, 51(3), 2000, 166-80.
3. Binwal, J.C. and Lalhmachhuana: Knowledge representation; concept, techniques and the analytico-synthetic paradigm. *Knowledge Organization*, 28(1), 2001, 5-16.
4. Krause, J.: Current research information as part of digital libraries and the heterogeneity problem;integrated searches in the context of databases with different content analyses. http://www.uni-koblenz.de/~krause/
5. Krause, J., Plumer, J. and Schwanzl, R.: *Content analysis, retrieval and metadata; Effective networking for Mathematics, Physics and Social Sciences* (Personal communication).
6. Hurd, J.M.: The transformation of scientific communication; a model for 2020. *JASIS*, 51(4), 2000, 1279-83.
7. Hjorland, B.: Towards a theory of aboutness, subject, topicality, theme, domain, field, content… and relevance. *JASIS,* 52(9), 2001, 774-78.
8. Hartley, J. and Kostoff, R.N.: How useful are keywords in scientific journals? *Journal of Information Science*, 29(5), 2003, 433-38.

**