SSARSC

Scientific Society of Advanced Research and Social Change
SSARSC International Journal of Library Information Network and Knowledge
Volume 1 Issue 1, February 2016, ISSN: 2455-5207

Problems and Prospects in Thesaurus Construction: A Case Study of Nanotechnology Thesaurus

Abhishek Sharma
Scientist, CSIR-NPL
sharmaa@nplindia.org

Thesaurus construction involves various activities like collection and control of terms, semantic association of terms, organization of terms etc. Each step has its particular associated problems. Over the years, significant efforts have been made to highlight the problems and to suggest alternative solutions for the problems. Recommendations and suggestions appeared in the direction embedded in standards and guidelines that are developed for thesaurus compilers. This paper is a contribution in the direction, where the researcher's points of views are listed in the form of issues and concerns associated with Nanotechnology thesaurus construction. Further, some of the decisions that were taken during the course of development of Nanotechnology thesaurus are also addressed in the paper. The work will support other users and researchers to perform further research in the area.

Keywords: Controlled Vocabulary, Information Retrieval System, Thesaurus, Nanotechnology

1. INTRODUCTION

Communication involves meaningful exchange of thoughts between the participants using various means like gesture, speech, writing, etc. It is continuous and goes on throughout our lifetime. Developments in information technology have affected the process of communication to a great extent. Internet has brought in a paradigmatic change in communication. It has enabled instant global transmission of thoughts. Interactive exchange is the hallmark of the Internet. Collaborative studies and research have been facilitated due to the developments in Internet and web technology. These developments have led to increasing generation of information.

Indexing helps to retrieve information from various access points, as per the needs of users. It has assumed a more important role with the exponential growth of information. Conventionally, information retrieval has been done using classification and cataloguing schemes. Library catalogue has been considered as the forerunner of information retrieval systems that helped the user in getting his information. It also informs the user about the library holdings.

Earlier, library and information practitioners were involved in information retrieval practices. But with the advent of IT, retrieval has been made easy and user friendly that end user can do it himself/herself. Disintermediate searching is preferred and more popular now. Web search engines play an important role in information retrieval. Indexing terms useful in deriving search expressions can be labelled either by using natural language of authors or language that makes use of controlled vocabulary tools. Choosing from among the two is always a controversial issue. Reporting the drawbacks associated with the use of natural language in information retrieval, Svenonius (1986) highlighted that the usage of controlled vocabulary helps in overcoming the difficulties appeared with the use of natural language i.e. homonyms and synonyms.

Among the existing controlled vocabulary tools thesaurus is considered as prominent in knowledge organisation and retrieval systems. The observation has been unanimously accepted by a number of scholars like (Gopinath, 1985; Kumbhar, 2005; Pinto, 2008).

2. STATEMENT OF THE PROBLEM

Nanoscience is the study of properties, characteristics, phenomenon, and manipulation of materials at nanoscale level. Nanotechnology is the applied part of Nanoscience. At this scale i.e. 10^{-9} , properties of matter are entirely different from that at large scale. From the view point of Nanotechnology, separate domains converge under this umbrella i.e. with an understanding of Chemistry and Physics of matter at nano level. Nanotechnology opens up breath taking opportunities in Biology, Data Storage, Engineering, Environmental Science, Materials Science, Medicine, etc. It has huge potential, and is receiving substantial financial support from governments. Due to its vast scope and motivational promotion from government, it is attracting intellectuals from distinct disciplines.

Research in Nanotechnology has generated a vast amount of literature. It is important that access to the literature is available uninhibited to the researchers and other users. A survey of the literature reveals that there is no thesaurus in the field. There exist some glossaries in Nanotechnology like Glossary of Nanotechnology Terms developed by The Institute of Nanotechnology, Scotland, ASTM Standard Terminology relating to Nanotechnology, ISO Nanotechnology vocabularies. A glance through the IET Inspec Thesaurus shows around 20 terms in Nanotechnology. LC List of Subject Heading, 32nd edition, 2010 has around 10 terms of Nanotechnology. These two sources are general in nature, hence a special controlled vocabulary is required in the area. Thus, to provide semantic knowledge structure in the field, a Nanotechnology thesaurus was developed by the researcher. The study was undertaken with an aim to present researcher's point of views for issues and concerns associated with Nanotechnology thesaurus construction.

3. OBJECTIVES AND SIGNIFICANCE OF THE STUDY

The research work was undertaken to develop a semantic knowledge map of Nanotechnology concepts which were previously not related in literature. The present study highlights some of the problems that appeared during the development of Nanotechnology thesaurus and to further discuss the decisions taken by the researcher for those concerns. The paper will help the users in building their understanding for issues pertaining to the development of thesaurus and will serve as guideline for researchers to perform further research in the area.

4. REVIEW OF RELATED LITERATURE

Due to the overwhelming significance of thesaurus in information processing and retrieval operations, thesaurus construction on various subjects has always been the focus of researchers. To offer a semantic knowledge structure and to help the academia working in the field of Laser Science, Singh (1993) designed a thesaurus for Laser Technology. Study by Neelameghan and Raghavan (2007) reported the experiences of contributors in developing Tamil-English thesaurus for Classical Tamil studies. Work of Nishikawa et al (2010) described the construction of a pathological thesaurus. Likewise, since their founding, thesauri from various learned bodies like IET Inspec thesaurus, Joint Thesaurus (ETDE/INIS), NASA thesaurus etc. are revised in series of editions.

Regarding the issues involved in thesaurus construction, there exists vast literature in the field. Miller (1997) has discussed general problems involved in thesaurus construction. In the same line, Nielsen (2004) presented a set of readings that deals with the issues and problems related with the thesaurus construction. Discussing about the various approaches for thesaurus construction Ghose and Dhawle (1977) highlighted the problems that appeared during thesaurus construction. Stating the usefulness of web based dictionaries in thesaurus construction, Nakayama, Hara and Nishio (2007) discussed the problems associated with the use of Natural Language Processing (NLP).

5. NANOTECHNOLOGY THESAURUS

Nanotechnology thesaurus with around 2500 terms was developed by the researcher as one of the objectives of his research work. The study was undertaken to provide better understanding of the semantic relationships between the concepts covered in literature.

For example: Sample Entries in Nanotechnology Thesaurus

aberration correction
SN aberration corrected scanning transmission electron microscope will enable imaging and analysis of nano structures
RT aspheric lenses

abrasive blasting
SN generation of specific surface features at the nano scopic scale by using abrasive blasting
BT nano mechanics
NT sand blasting
RT nano abrasion
nano topography

abrasive flow deburring
USE **abrasive flow machining**

abrasive flow machining
SN abrasive flow machining is a technique developed for nano finishing of parts even with complicated geometry
UF abrasive flow deburring
BT nano abrasion machining
RT surface texture

Figure 1

6. ISSUES AND CONCERNS

This section covers the overview of steps involved in thesaurus construction. Factors that were considered and decisions taken during the development of Nanotechnology thesaurus are also highlighted in this section.

6.1 Approaches of Thesaurus Construction

There exists two approaches for thesaurus construction, one is manual and the other is automatic. Both the approaches have their pros and cons. Emphasizing the manual approach for thesaurus construction, Chen and Thiel (2004) stated that manually developed thesaurus have better semantic knowledge structure. This is due to the involvement of statistical and syntactical methods, namely co-occurrence statistics in automatic methods which are incapable in supporting the semantic relational structure (Mandala, Tokunaga & Tanaka, 2000). However, manual thesaurus construction and maintenance is an annoying, resource demanding and time consuming work.

In contrast to manual approach, automated algorithmic approach offers pace, flexibility and ease in developing a thesaurus.

6.1.1 Factors Considered and Decision Taken

Considering the complexities and limitations of both the approaches, a manual intellectual operation in addition to computer applications has been followed for the development of Nanotechnology thesaurus. This approach has been supported and recognised by scholars, as it helps in overcoming the limitations of both the methods by complimenting each other (Granada, Vieira & Strube de Lime, 2012; Kohlhof, Schijvenaars & Diwersy, 2009; Liebeskind, Dagan & Schler, 2013).

6.2 Domain Selection

Defining precisely the fields to be covered in a thesaurus is the preliminary stage. Making choice from among the alternatives requires clarity for following points:

a. Demand of the work
b. Comfort and interest of the creator

6.2.1 Factors Considered and Decision Taken

Initially, various research areas in Physical Sciences were examined carefully. Analysis revealed that Nanotechnology which is relatively a new field of study is interdisciplinary in nature has its impact in almost every research area, whether it is Energy Harvesting, Materials Science, Radio & Atmospheric Sciences, Time & Frequency, Metrology, Quantum Phenomenon. Stakeholders around the globe are researching, giving new concepts and coining terminology in the field. The terminology needs to be explained, popularised and tools created to achieve consistency in its usage. Standard vocabularies are needed for the purpose. Substantial efforts have been made in this direction by different organisations. American Society for Testing and Materials (ASTM) and International Organisation for Standardisation (ISO) are the leading agencies working in this direction at international level. CSIR-NPL, the nodal organisation of ISO Nanotechnology Programme in India is working in this direction at national level. Though extensive efforts are being made to conceptualise the standard terminology for Nanotechnology, yet the literature available in the field revealed that there is not much work available that describes the semantic relationships between the Nanotechnology concepts particularly in Physical Sciences. Therefore, the outputs of the study in form of Nanotechnology thesaurus would help the readers and researchers to lay hands on the relevant literature and in furthering research in the field.

In addition to this, Physics and Information Science background of the researcher is one of the motivational factor towards the development of Nanotechnology thesaurus.

6.3 Data Collection and Data Cleaning

Here the aim is to collect as many candidate terms that could adequately describe the domain. Further, terms were evaluated to remove all the irrelevant terms.

6.3.1 Factors Considered and Decision Taken

Resources that were found to be rich in Nanotechnology content were selected for term collection in consultation with subject experts. Following categories of resources were considered for term collection:

a. **Printed Resources**: Encyclopedia; Standards Terminology for Nanotechnology by ASTM (American Society for Testing and Materials), BSI (British Standard

Institution), ISO (International Organisation for Standardisation); Conference Proceedings and Handbooks.

b. **Electronic Resources**: Three electronic resources namely Web of Science, Scopus, and INSPEC were accessed in addition to other web resources.

c. **Committee Approach**: All the candidate terms were submitted to subject experts for their opinion. During the process, they were asked to delete the irrelevant concepts and add new concepts.

For data cleaning, all the terms were examined carefully and the terms that were found to be outside the scope were deleted from the list. Here manual approach, as well as fundamentals of MS Excel was applied to get the terms in desired format, i.e. single term in single cell of MS Excel Worksheet.

6.4 Preferred Term Selection

Preferred term in a thesaurus is a focal point where related information about the concept is placed. In contrast, non-preferred term is used to direct the user to the preferred term. There has been a debate between the scholars- one who favour the word frequency as a base for inclusion of a term in controlled vocabulary (Hulme, 1950; Dabney, 2007) and the other limits the thought by saying that the terms with very high frequency are considered to be too general to be useful in describing the subject matter. Similarly, those that appeared very infrequently may represent concepts that have little relevance for the subject (Lancaster, 1972; Ghose & Dhawle, 1977). Thus, adoption of any approach demands proper justification.

6.4.1 Factors Considered and Decision Taken

Examining the views of various scholars in the present study, make us to mark all the terms as preferred terms that appeared more than once. As in the present work, omitting frequently as well as infrequently terms resulted in neglecting core concepts that actually represents the need of users.

Table1

Frequency of Terms in Literature

Terms	Frequency
nanotechnology	366
nanometrology	344
atomic force microscopy	266
calibration	157
nanostructured materials	139
nanoprobes	6
quantum dot	5
fullerene	4
nanograting	3
nanosensors	2

In addition to this, new terms were also admitted as member to Nanotechnology thesaurus, as per the advice/interest of subject experts who would be the people to be served with this knowledge structure.

6.5 Defining Relationships between the Concepts

Establishing semantic relationship between concepts can be described as a practice of assigning meaningful association between two or more entities. The basic relationships described by Mazzocchi, Tiberi, De Santis and Plini (2007) and National Archives of Australia (2003) in the guidelines for common wealth countries are Hierarchical, Associative and Equivalence. Among the three classes of thesaurus relationships, defining associative relationship is considered as the most difficult task (Aitchison, Gilchrist and Bawden, 2000; Chowdhury, 1999).

6.5.1 Factors Considered and Decision Taken

Associative relationship is used to denote the relationship between terms that are neither hierarchical nor equivalence. To define the categories of terms that can be incorporated under

this class, Chowdhury (1999) cited BS5723. For the present work following categories were defined:

- ***A Process and its Instrument***

Eg: imaging RT spectroscopy

- ***Action and its Product***

Eg: emission spectroscopy RT emission spectra

- ***Concept and its Properties***

Eg: nano structures RT thermoelectric property

- ***A Technique similar to other Technique***

Eg: nuclear magnetic resonance RT ferromagnetic resonance

The task of creating semantic relationships between concepts is an intellectual work are usually performed by a group of subject experts and Information Science professionals. In the present study following three approaches were applied for semantic association of concepts:

- Physical Science and Information Science background of the researcher;
- Subject experts approach; and
- Review of related literature

6.6 Standards and Guidelines

Standards and guidelines for thesaurus construction are indispensable documents for thesaurus developers. These documents exist for both monolingual and multilingual thesaurus.

Factors Considered and Decision Taken

Present work deals with the development of monolingual thesaurus, therefore, as suggested by Aitchison, Gilchrist and Bawden (2000), the US Standard ANSI/ NISO Z39.19-2005 (Guidelines for the construction, format, and management of monolingual controlled vocabularies) was consulted. In addition to this, guidelines covered in a book by Aitchison, Gilchrist and Bawden (2000) were also followed. Further, to represent the structure of entries, IET Inspec Thesaurus was approached. Nanotechnology standard terminology defined by American Society for Testing (ASTM), British Standards Institution (BSI), International Organisation for Standardisation (ISO) are also covered in the current work.

6.7 Scope Note

Scope note clarifies the intended meaning of term in a thesaurus. It is added to limit the scope of the term.

6.7.1 Factors Considered and Decision Taken

Usually, scope note for a preferred term is added to define its scope. In the present work, scope not has been given for concepts in view of their application in Nanotechnology, as available in literature and also for ambiguous terms. Core terms identified by experts have been defined. These include fanciful terms and those that are relatively recent. Terms that are obvious and are from traditional discipline of Science with ‘nano’ prefix have not been given a scope note.

Eg:

- **abrasive blasting**

  Scope Note: generation of specific surface features at the nano scopic scale by using abrasive blasting
- **buckypaper**

  Scope Note: buckypaper is a thin nano material composed of cylindrical carbon nano tubes
- **quasi particles**

  Scope Note: quasi particle is a disturbance, in a medium, that behaves as a particle
- **repeatability**

Scope Note: repeatability is the closeness of agreement between successive measurements carried out under the same conditions

7. CONCLUSION

A well-constructed thesaurus serve as a tool for representation of knowledge. It offers a semantic knowledge structure to support representation and retrieval of meaningful and precise information. To relate the Nanotechnology concepts that were previously not related in literature and to offer semantic knowledge structure to researchers, a Nanotechnology thesaurus was developed as a part of research work of the author. Some associated issues of thesaurus construction are analysed and presented in the paper. The analysis is based on the research experience of the author. Some methodical approaches to overcome the issues are highlighted in the study. It is felt that the various points discussed in this paper would benefit all the stakeholders connected with knowledge representation tools.

REFERENCES

Aitchison, J., Gilchrist, A. & Bawden, D. (2000). *Thesaurus construction and use: A practical manua*l (4th ed.). London: Europa.

Chen, L. & Thiel, U. (2004). Language modeling for effective construction of domain specific thesauri. In F. Meziane & E. Metais (Eds.), *Lecture Notes in Computer Science: Vol. 3136. Natural Language Processing and Information Systems: Proceedings of the 9th International Conference on Applications of Natural Language to Information Systems, NLDB 2004*, June 23-25, 2004 (pp. 242-253). Berlin, Heidelberg: Springer-Verlag. Retrieved November 07, 2014, from http://books.google.co.in/books?id=7FD94nXAOboC&printsec=frontcover#v=onepage&q&f=false

Chowdhury, G.G. (1999). *Introduction to modern information retrieval*. London: Library Association. Dabney.

Dabney, D. (2007). The universe of thinkable thoughts: Literary warrant and West's key number system. *Law Library Journal, 99*(2), 229-247. Retrieved from http://www.aallnet.org/main-menu/Publications/llj/LLJ-Archives/Vol-99/pub_llj_v99n02/2007-14.pdf

Ghose, A. & Dhawle, A.S. (1977). Problems of thesaurus construction. *Journal of the American Society for Information Science, 28*(4), 211-217.

Gopinath, M.A. (1985). Equivalence relations in information retrieval thesaurus. *SRELS Journal of Information Management, 22*(1), 57-63. Retrieved from http://library.isical.ac.in/jspui/bitstream/10263/1105/1/LSWASTD-22-1-1985-P57-63.pdf

Granada, R.L., Vieira, R. & Strube de Lima, V.L. (2012). Evaluating co-occurrence order for automatic thesaurus construction. *IEEE Thirteenth International Conference on Information Reuse and Integration*, August 8-10, 2012, Las Vegas, NV (pp. 474 - 484). Retrieved from IEEE Xplore. doi: 10.1109/IRI.2012.6303046 Hulme, 1950

Hulme, E.W. (1950). *Principles of book classification*. London: Association of Assistant Librarians. Retrieved July 10, 2014, from http://www.iva.dk/bh/Core%20Concepts%20in%20LIS/Hulme_444-449.pdf

Kohlhof, I., Schijvenaars, B. & Diwersy, M. (2009). Semi-automatic construction of domain-specific thesauri. In R. Aly, C. Hauff, I.D. Hamer, D. Hiemstra, T. Huibers & D. de Jong (Eds.), *Proceedings of the 9th Dutch-Belgian Information Retrieval Workshop*, February 2-3, 2009, Enschede, The Netherlands (pp. 64-70). Enschede: Centre for Telematics and Information Technology. Retrieved from http://doc.utwente.nl/65379/1/dir2009proceedings.pdf

Kumbhar, R. (2005). Speciator based faceted depth classification's application in thesaurus construction. *Annals of Library and Information Studies, 52*(1), 15-24. Retrieved from http://nopr.niscair.res.in/bitstream/123456789/3982/1/ALIS%2052(1)%2015-24.pdf

Lancaster, F.W. (1972). *Vocabulary control for information retrieval*. Washington, DC: Information Resource Press.

Liebeskind, C., Dagan, I. & Schler, J. (2013). *Semi-automatic construction of cross-period thesaurus*. Paper presented at the Seventh Workshop of the Association for Computational Linguistics on Language Technology for Cultural Heritage, Social Sciences, and Humanities, August 08, 2013, Sofia, Bulgaria (pp. 29-35). Retrieved November 11, 2014, from http://www.aclweb.org/anthology/W13-2704

Mandala, R., Tokunaga, T. & Tanaka, H. (2000). Query expansion using heterogeneous thesauri. *Information Processing & Management, 36*(3), 361-378. doi: 10.1016/S0306-4573(99)00068-0

Mazzocchi, F., Tiberi, M., De Santis, B. & Plini, P. (2007). Relational semantics in thesauri: Some remarks at theoretical and practical levels. *Knowledge Organisation*, *34*(4), 197-214. Miller (1997)

Miller, U. (1997). Thesaurus construction: Problems and their roots. *Information Processing & Management, 33*(4), 481-493. doi: 10.1016/S0306-4573(97)00009-5

Nakayama, K., Hara, T. & Nishio, S. (2007). A thesaurus construction method from large scale web dictionaries. *Twenty First International Conference on Advanced Information Networking and Applications*, May 21-23, 2007, Niagara Falls, Canada (pp.932-939). Retrieved from IEEE Xplore. doi: 10.1109/AINA.2007.23

National Archives of Australia. (2003). *Developing a functional thesaurus: Guidelines for Commonwealth agencies*. Canberra: Author. Retrieved from http://www.naa.gov.au/Images/developing-a-thesaurus_tcm16-47228.pdf

Neelameghan, A. & Raghavan, K.S. (2007). *Online bilingual thesaurus for subjects in the humanities: A case study*. Paper presented at the International Conference on Semantic Web & Digital Libraries, July 21-23, 2007, Indian Statistical Institute, Bangalore. Retrieved from http://drtc.isibang.ac.in/ldl/bitstream/handle/1849/394/065_p24_raghavan.pdf

Nielsen, M.L. (2004). Thesaurus construction: Key issues and selected readings. *Cataloging & Classification Quarterly, 37*(3-4), 57-74. doi: 10.1300/J104v37n03_05

Nishikawa, S., Yamashita, T., Imai, T., Yoshida, M., Sakuratani, Y., Yamada, J., … Hayashi, M. (2010). Thesaurus for histopathological findings in publically available reports of repeated-dose oral toxicity studies in rats for 156 chemicals. *The Journal of Toxicological Sciences, 35*(3), 295-298. Retrieved from http://www.jtoxsci.org/manuscript/35_295_manuscript.pdf

Pinto, M. (2008). A user view of the factors affecting quality of thesauri in social science databases. *Library & Information Science Research, 30,* 216-221. doi: 10.1016/j.lisr.2007.12.003

Singh, J.P. (1993). *Development of a model information retrieval system for laser technology* (Unpublished doctoral thesis). University of Delhi, Delhi.

Svenonius, E. (1986). Unanswered questions in the design of controlled vocabularies. *Journal of the American Society for Information Science, 37*(5), 331-340. Retrieved from http://polaris.gseis.ucla.edu/gleazer/462_readings/Svenonius_1986.pdf