CHAPTER 2

REVIEW OF RELATED LITERATURE

2.1  INTRODUCTION

The term “information” is considered as the currency of information world. In the present scholarly society, it is information which allows one to have an edge over others and significantly permits us to successfully adjust with the external existence conditions. In national development, information performs similar role as distinct forces like air force, army, and navy do for national defence and power. In the former case, instead of tanks, airplanes, and submarines various information carriers in the form of books, periodicals, etc. and people with distinct requisite skills to organise and disseminate information are needed (Spade, 2012). Thus, considering information as power, Lesko (n.d) states that “those who control information are the most powerful people on the planet”. Consequently, timely delivery of desired information is essential. Increasing demand for information forces mankind to develop various information organisation and retrieval tools, and methods like cataloguing, classification, controlled vocabularies, indexing, search engines etc. Among the existing information organisation and retrieval tools, controlled vocabularies stand out as prevailing tools. They perform an important role in information organisation and retrieval activities. Controlled vocabularies are often essential for practitioners in making use of consistent terms to represent the same concept. In the retrieval operations, controlled vocabularies guide users to select correct terms for their searches (Harpring, 2010). Further, role of controlled vocabularies extended to web world where search engines are considered as popular information retrieval tools. Hedden (2008) states that search results obtained with search engines can be improved by using controlled vocabularies for assigning keywords for documents to be indexed on the web.

Several studies have been conducted that highlight the role of controlled vocabularies in different eras. These include, the impact of controlled vocabulary tools for organising information retrieval in conventional world (Lancaster, 1972; Svenonius, 2003), and their significance in modern digital environment (Aitchison, Gilchrist & Bawden, 2000;
The review of literature in the following pages deals with two promising controlled vocabulary tools i.e. thesaurus and ontology in organising information retrieval. The review begins with a general overview of the controlled vocabulary. This is followed by discussion on thesaurus, its development, standardisation efforts for its construction, its role in full text era. Further, the chapter focuses on a move from current web to semantic web, and application of ontology as a tool for semantic web. Finally, the review highlights the role of thesaurus as a tool for ontology building.

2.2 CONTROLLED VOCABULARY

Communication which is considered as carriers of thoughts and ideas required language as a medium of interaction. Though, it used to take place even before the development of language, usually through gestures. With the development of language it became easier for people to interact. Language involves terminology and rules for their arrangement. Lewis and Jones (1996) categorise language into natural language and artificial language. The former refers to our language, whereas, the artificial language referred to describe the language designed for a specific purpose such as for classification and retrieval operations (Deschatelets, 1986). Using either a natural language or artificial language as indexing/retrieval language has always been a matter of debate among the scholars. Supporters of controlled vocabulary which comes under the domain of artificial language argued that for effectiveness of information storage and retrieval system, vocabulary control is essential (Borst, 2012; French, Powelly, Gey & Perelmanz, 2001; Kuang-Hwei & Lee-Smeltzer, 2000; Lancaster, 1972; Pasche, Gobeill, Vishnyakova, Ruch & Lovis, 2013) and many more. Whereas, followers of free-text, a mode that provide autonomy to user to perform full text searching for his/her word(s), favoured usage of natural language in machine readable full text environment (Alizadeh, Fattahi & Davarpanah, 2010; Frid, Logounova, Michailov, Nusinzon & Zeltser, 1997; Khokane & Atique, 2013; Strzalkowski, Lin, Perez-Carballo & Wang, n.d.). Stating the strengths and weaknesses of both the approaches,
Aitchison, Gilchrist and Bawden (2000) reported that in a manual system effectiveness of controlled vocabulary is undoubted, whereas, features of natural language like currency of terms, exhaustivity of coverage, ensures significant recall in full text era. Thus, stating that a system with free text arrangement is better than one without it is debatable, as the database like INSPEC still makes use of controlled vocabulary to improve precision and recall in machine readable post coordination system (Aitchison, Gilchrist & Bawden, 2000). 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.

Example of Homonym

- Ray: a beam of light
  Ray: a type of fish
- Shear: strain
  Shear: very steep
- Sight: vision
  Site: place
- High: elevated
  Hi: a greeting

Example of Synonym

- blank, empty
- cracks, fractures
- look, see

Works of many scholars revolves around controlled vocabulary on various aspects like designing a controlled vocabulary in a domain, complementary role of controlled vocabularies in correspondence to free text searching, converting a controlled vocabulary into ontology, analysis of search statements drawn with the help of controlled vocabulary etc. (Cimino, Hripcsak, Johnson & Clayton, 1989; Fidel, 1991; Qin & Paling, 2001;
Further, the term ‘controlled vocabulary’ appears frequently in texts such that Google Scholar reflected 1,180,000 records in response to the query made on October 24, 2015 with the key term ‘controlled vocabulary’.

**Figure 2.1**

**Scholarly Output from Google Scholar**

Similarly, to chalk the popularity of term among the scholars, topic wise search was performed on the database ‘Web of Science’ on October 24, 2015. Continuous increase in the number of publications as reflected in Fig. 2.2 revealed that day-by-day, the concept ‘controlled vocabulary’ is catching the interest of scholars.
Further, reporting the popularity of controlled vocabulary among various category of users Harpring (2010) highlighted that indexers use controlled vocabularies for assigning subject terms and also contribute to them, information seekers consult controlled vocabularies to improve their search results, and online publishers post them on their system.

The concept ‘controlled vocabulary’ is an umbrella term that covers range of information organisation and retrieval tools. Stating the form of controlled vocabularies, Jaasklainen, Moschner and Wackerow (2009) explained that they may appear in the form of short lists or complex structures having hierarchical relationships.

On the same lines, two systems where controlled vocabulary can be used to index and retrieve documents are highlighted by Lancaster (1972). One way is the pre-coordination of terms in which indexer coordinates the concepts together in advance. The coordinations so developed are fixed and rigid. Best example that suits this category is book subject indexes. On the other hand, the second system which is free for post-coordination and offers flexibility to combine concepts in any way is discussed. Further, he supported faceted classification scheme to achieve control over the drawbacks associated with the use of natural language. With the use of faceted classification schemes same class number
to nearly synonyms terms can be provided and thus, the problem of synonyms and homonyms can be prevented. As explained by the author, the functions of controlled vocabulary are also performed by the list of subject heading, as in it the terms which are considered to be synonymous are linked by ‘see’ and ‘see also’ references. The contributor also presumed thesaurus as widely accepted controlled vocabulary tool that lists descriptors and display various relationships between the terms to represent a semantic structure for a domain.

Another work that to summarise the tools that comes under controlled vocabulary is by Harpring (2010). According to the author, subject heading lists, simple controlled lists, synonym ring list, taxonomies, folksonomies, thesaurus, and ontology are the most common types of controlled vocabularies.

Morshed (2012) has also classified the controlled vocabulary tools under general controlled vocabulary, subject specific controlled vocabulary, Library of Congress authority list.

2.3 THESAURUS: A CONTROLLED VOCABULARY TOOL

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 (Gopinath, 1985; Kumbhar, 2005; Pinto, 2008). Various experts have discussed the concept ‘thesaurus’ with distinct perspective. Aitchison and Clarke (2004) cited Shorter Oxford Dictionary to state that the word ‘thesaurus’ first appeared in English in 1736. In the perspective of information retrieval, thesaurus was first used in the year 1957 (Vickery, 1960). Among the existing controlled vocabulary tools, thesaurus emerged as an aid to indexers that adds their vocabulary (Roberts, 1984). Bernier (1968) views thesaurus as an external guide which directs the user from known concepts to those required. Further, expanding the scope of thesaurus from users’ point of view, Chen, Lynch, Basu and Dorbin (1993) viewed thesaurus as a concept space, which represents various domain concepts along with their semantic relations and thus significantly enhance their understanding for a domain. Hence, for effective information
processing and retrieval, thesaurus serves as a helping tool for both practitioners and information seekers.

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.

Further, scholarly work on various aspects of thesaurus like different approaches for thesaurus construction, standards for thesaurus building, use of thesaurus in full text environment etc. also keep on appearing.

2.3.1 Approaches for Thesaurus Construction

Anderson and Perez-Carballo (2001) categorised the work of thesaurus construction into the following two group:

- Manual construction
- Computational/ automatic construction

In current scenario, thesauri are usually constructed on automatic ground. Whereas, manual practices to construct a thesaurus were observed during the older days (Dagan, 2000).

For construction of controlled vocabulary Lancaster (1972) suggested the following steps:
**Table 2.1**

Comparison between Manual and Automated Approach for Thesaurus Construction

<table>
<thead>
<tr>
<th>Step</th>
<th>Explanation</th>
<th>Manual System</th>
<th>Automated System</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identification of subject area to be covered</td>
<td>Defining the subject field, establishing the subject boundaries in which a thesaurus has to be constructed</td>
<td>Manual approach where intellectual efforts are required</td>
<td>Manual approach where intellectual efforts are required</td>
</tr>
<tr>
<td>Selection of suitable terms to describe the area</td>
<td>Collection of adequate number of domain concepts to be incorporated in thesaurus</td>
<td>Manual selection</td>
<td>Computational selection and preparation of ranked list using statistical techniques</td>
</tr>
<tr>
<td>Making decision about the exact form of terms to appear in the vocabulary</td>
<td>Deciding acceptable grammatical forms, singular &amp; plural forms, variant spellings, abbreviations, and acronyms</td>
<td>Thesaurus construction and development standards are referred</td>
<td>Thesaurus construction and development standards are referred</td>
</tr>
<tr>
<td>Organisation of terms in useful pattern</td>
<td>Establishing meaningful relationships between the domain concepts</td>
<td>Manual approach where intellectual efforts are required</td>
<td>Based on statistical and syntactical ground</td>
</tr>
<tr>
<td>Display of terms in helpful manner</td>
<td>Arrangement of terms in alphabetical order</td>
<td>Manual technique such as Card system are employed</td>
<td>Performed with the help of computer</td>
</tr>
</tbody>
</table>
Thesaurus construction was earlier done manually but now automatic methods are used for it. On semantic ground, manual procedures are preferred in contrast to automatic methods. This is due to the fact that automatic thesaurus are usually developed with algorithms which are incapable of establishing semantic relationships between the concepts. Denoting the notable example of human efforts towards thesaurus construction, Nakayama, Hara and Nishio (2007) highlighted WordNet, which is a prominent work of large number of contributors. Similarly, emphasising the manual approach for thesaurus construction Chen and Thiel (2004) stated that manually developed thesauri have better semantic knowledge structure. However, manual thesaurus construction and maintenance is an annoying, resource demanding and time consuming work. Thus, the associated limitations of manual approach and attractive feature of automated mode in terms of economy of men, money and time has influenced practitioners and researchers to adopt automated approach for thesaurus construction.

In contrast to manual approach, construction of automatic thesaurus involves wide range of statistical and syntactical methods, namely co-occurrence statistics, and grammatical relations (Mandala, Tokunaga & Tanaka, 2000). Over the years significant amount of research has been conducted on automatic approaches for thesaurus construction. Crouch (1990) proposed term clustering algorithm for automatic thesaurus construction. The suggested approach collects the documents and is capable of generating thesaurus classes. The algorithm is based on term discrimination principle. Study conducted by Guntzer, Juttner, Seegmuller and Sarre (1989) describes thesaurus generating system (TEGEN) as a component of information retrieval system. The program is capable of automatically capturing the concepts from user’s search pattern. These tapped concepts represent the domain knowledge to be represented in thesaurus. Further, users’ feedbacks are also invited in case of any ambiguity. Work of Park and Choi (1996) presented another approach for automatic thesaurus construction. Based on statistical approach, they proposed Bayesian network for encoding relationship between terms. A probabilistic graphical model, which is the core of Bayesian network served as the base to represent relationship between terms. Further, based on the proposed methodology for automatic thesaurus construction, an average performance gain of 27.1% is listed by the contributors. Chen and Thiel (2004) proposed the Layer-Seeds algorithm for automatic
the effectiveness of algorithm, authors observed that thesaurus built by using Layer-seeds methods can be effectively used for automatic query expansion by using MS-relationship, whereas, TS-relationship route can be employed for interactive query expansion. Contribution by Yang and Power (2008) introduced another statistical approach for automatic thesaurus construction. Authors claimed that significantly higher precision can be obtained by using syntactically conditioned co-occurrences or grammatical relations approach in computing distributional similarity.

Albeit, automated algorithmic approach for thesaurus construction offers pace, flexibility and ease in deriving domain concepts from author’s text, yet significantly recognising the best semantically related terms in an automated approach of thesaurus construction is the foremost problem (Granada, Vieira & Strube de Lima, 2012).

Due to the relational complexities, semantic ambiguities associated with the automated approach, there appears a demand for certain manual intellectual operations in addition to computational algorithms in thesaurus construction. Thus, considering the complexities and limitations associated with both the approaches, complementarity of manual and automatic approaches is recommended. Kohlhof, Schijvenaars and Diwersy (2009) called this mixed approach as hybrid technique or semi-automatic approach for thesaurus construction. Number of information retrieval workers has supported this approach.

Work of Liebeskind, Dagan and Schler (2013) focuses the semi-automatic approach for building cross-period thesaurus. The work helps the users to make search for modern terms, listed as target terms in thesaurus, and obtain earlier period terms as related terms to target terms. Like automatic approach, candidate terms are initially extracted automatically by using algorithm scheme in combination with statistical techniques. Afterwards, to add a flavour of semantic to a list obtained through automatic technique, all the candidate terms are examined manually by domain experts. Intellectuals are responsible for taking a decision on whether the candidate is related to the target term, and whether it can be used as a query expansion term for target term.
Paper by Kohlhof, Schijvennaars and Diwersy (2009) describes a hybrid method that combines the statistical, linguistic and intellectual efforts of subject experts for semiautomatic construction of domain thesaurus.

2.3.2 Role of Computer in Thesaurus Construction

Development of quality thesaurus involves intellectual efforts besides clerical operations. The chief intellectual jobs involves recognising main domain concepts, selecting suitable descriptors, representing relationships between the concepts, and helpful arrangement of descriptors in structured manner (Raizada, Ramachandran, Satyanarayana & Pal, 1975). Conventional manual way of constructing a thesaurus has been considered as time taking and cumbersome task. Utility of computer technology in thesaurus building operations can offer relief to thesaurus compiler by reducing human editing and maintenance sub-processes. As stated by Lancaster (1985) computer can be utilised in the following areas of thesaurus construction:

- Gathering of indexing terms on the basis of frequency statistics.
- Sorting and counting to produce co-occurrence tables.
- Making a consistent alphabetical display.
- Representing thesaurus in number of formats.
- Addition, deletion and alteration of terms at any point of time.
- Offer non-linear searching, if a thesaurus is in machine readable form.

Similarly, Surace (1970) mentioned that addition of computer capabilities in thesaurus building task has significantly altered the information storage and retrieval system by reducing the human efforts involved in editing and maintenance of thesaurus. The author further added that unlike traditional system where cards system was used to count the frequency of terms and making alphabetical arrangement, computers can perform the tasks accurately and speedily. Further, explaining the role of computer in thesaurus construction Devasason (n.d.) highlighted that application of computer adds ease in updating, editing, evaluating and printing of thesaurus.
2.3.3 Standardisation Efforts in Thesaurus Construction

Standards are essential to ensure quality, safety and efficiency. They help in international collaboration. Standards contribute significantly to terminology work. Role of standardisation in terminology domain increases with the accessibility of information via Internet. Standardisation efforts have been carried out both at the organisational and document level.

2.3.3.1 Standardisation Efforts at Organisational Level

Thesaurus is considered as a premier tool for information processing and retrieval. The tool was developed long back and since then it is a subject of research for both national and international organisations. Number of organisations have played significant role in establishing standards for thesaurus construction. Brief introduction for few of them is covered as under:

(i) National Information Standards Organisation (NISO)

NISO, a non-profit association accredited by the American National Standards Institute (ANSI) has been established to identify, develop, maintain, and publish technical standards related to information management in the emerging digital world. NISO standards have a wide range of spectrum and are suitable for all information-related needs such as retrieval, re-purposing, storage, metadata, and preservation.

NISO was founded in the year 1939 and since then it has been receiving support from the groups it serves. Throughout the year, NISO offers innovative programmes on standards and organises exploratory conferences/workshops on important themes.

Overall 55 standards are listed under NISO standards project on various topics like abstracting, indexing, and standard numbering schemes for documents but the one which is related with the controlled vocabulary is “ANSI/NISO Z39.19-2005 (R2010) Guidelines for the Construction, Format, and Management of
Monolingual Controlled Vocabularies” (National Information Standards Organisation, 2015a).

Moreover, standards listed under NISO projects are openly available in public domain (Milstead, 1998).

(ii) **International Organisation for Standardisation (ISO)**

ISO is an independent, non-governmental organisation. It has been considered as world’s largest developers of standards and till date more than 19500 international standards are listed under various domains like quality management, environmental management, food safety management, social responsibility, risk management, information security, language codes, sustainable events, energy management etc. All standards of ISO are considered as voluntary international standards for global use (International Organisation for Standardisation, n.d.).

As stated by Clarke and Zeng (2012), ISO 25964 is the international standard published by ISO for thesauri construction.

(iii) **International Information Centre for Terminology (Infoterm)**

Infoterm was established by UNESCO in the year 1971. Later, in the year 1996, it was registered as an independent international scientific non-profit association. Its basic aim is to support international cooperation in the field of terminology. To highlight its other objectives, following tasks are listed under Infoterm activities:

- Cooperation in the field of terminology
- Terminology policies
- Legal issues related to IPR
- Access to information and knowledge
- Terminology standardisation
Regarding its contribution, a series of publications on terminology have been brought out in collaboration with its partners (Infoterm, 2014).

(iv) **International Society for Knowledge Organisation (ISKO)**

ISKO has played a vital role in international cooperation on vocabulary control. It is an international society in the field of knowledge organisation. It was founded in 1989 with mission to develop conceptual work in the field of knowledge organisation. As an interdisciplinary society, ISKO comprised of more than 600 members around the globe from distinct fields like Computer Science, Information Science, Linguistics, Medical Informatics, Philosophy etc. To promote cooperation at international level, major activities of ISKO involve (International Society for Knowledge Organisation, 2014):

- Organising international conferences in every two years and national and regional conference on special topics;
- Publications of leading journal “Knowledge Organisation” (formerly known as International Classification);
- Bringing out “ISKO News”, a newsletter which is now a part of Knowledge Organisation; and
- Production of the series “Advances in Knowledge Organisation (AKO)”.

It contributes its bit in standardisation in vocabulary control by publishing articles in its publications and organising seminars/conferences on related themes.

### 2.3.3.2 Standardisation Efforts at Document Level

Existence of standards and guidelines for information retrieval thesauri was noticed during 1959 to 1993 (Krooks & Lancaster, 1993). They serve as essential document for thesaurus compilers. Though most of the elementary problems were recognised and answered by 1967 (Krooks & Lancaster, 1993), still the role of standards and guidelines can’t be denied in the present scenario. They help the practitioners in making decisions at various stages of thesaurus construction. Ample efforts have been made in this direction. A brief introduction for few of them is covered here as under:

The standard emerged with the overwhelming efforts of the working groups of NISO. It is meant for monolingual thesaurus and offers guidelines for display, construction, evaluation, maintenance and management of controlled vocabulary tools. The development of proposed standards involves rigorous peer review from NISO voting members. Other interested parties are also invited to participate in reviewing process. With the completion of peer review process, standard is submitted to the American National Standards Institute (ANSI) for its final approval. Once approved and verified, NISO standards get the status of American National Standards (National Information Standards Organisation, 2010).

(ii) **ISO 25964: Information and Documentation - Thesauri and Interoperability with other Vocabularies**

ISO 25964 is an international standard issued by the International Organisation for Standardisation. The drafting of ISO 25964 is commenced by WG8, a Working Group with its members from 15 countries. This standard is maintained by its Secretariat in National Information Standards Organisation (National Information Standards Organisation, 2015b).

This standard is the revised, updated version of its precursor i.e. ISO 2788, jointly produced by UNESCO and ISO in 1972-73 (United Nations Educational, Scientific and Cultural Organisation, 1976) and ISO 5964. Key difference between the two generations is of the move from conventional printed thesaurus world to a networked scenario that focuses on interoperability.

It is available in following two parts and can be purchased individually either from ISO or from any of its member nations.
Part 1: Thesauri for Information Retrieval - it cover guidelines entirely for the thesaurus (monolingual or multilingual) on various aspects like thesaurus construction and its content, applying facet analysis to thesaurus etc.

Part 2: Interoperability with other Vocabularies - it attempts to describe taxonomies, classification schemes, subject heading lists, synonym rings, ontologies, authority lists and file plans. It focuses mainly on how to establish interoperability between each of the type of the controlled vocabulary (Clarke, 2012).


These standards are absorbed by distinct organisations/ societies for their publications in form of manuals like Rare Books and Manuscript Section (RBMS) of the Association of College and Research Libraries (ACRL), a division of the American Library Association (ALA) produced a manual “RBMS Manual / Thesaurus Construction and Maintenance Guidelines” in accordance with NISO Z39.13 (RBMS manual, 2014).

Similarly, the work “Thesaurus Construction Guidelines: An Introduction to Thesauri and Guidelines on their Construction” resulted from a collaborative effort of the Digital Repository Ireland (DRI) and the National Library of Ireland (NLI). Guidelines were created in consultation with ISO 25964 (Ryan, 2014).

Further, stating the usefulness of thesaurus guidelines Spiteri (1998) mentioned that use of published guidelines is not a compulsion for thesaurus creators. Their use is suggested only because of the advantages associated with them i.e. their use helps in - variety reduction, and ensuring quality.
2.3.4 Thesaurus Relationships

Scholarly works for thesaurus define three basic types of terms relationship i.e. equivalence, hierarchical and associative (Aitchison, Gilchrist & Bawden, 2000; Kumbhar, 2005; Lancaster, 1972; Will, 2012). Different conventions are used to represent distinct relationships.

(i) Notation ‘U or USE’ and ‘UF or USED FOR’ are suggested for equivalence term category. Both the conventions are reciprocal to one another. Further, for equivalence relationship Aitchison, Gilchrist and Bawden (2000) suggested following basic types:

- **Synonyms**: two or more interchangeable terms whose meaning is considered to be same.
  Eg: cracks/ fractures

- **Lexical Variants**: two or more terms/ form of terms that represent the same concept.
  Eg: nano fiber/ nano fibre

- **Quasi Synonyms**: quasi refers to words which are considered to be similar in context of controlled vocabularies but are not exactly similar. Antonyms may be covered under this category.
  Eg: desorption/ absorption

(ii) In addition to equivalence relationship, there is another relationship which is denoted with the help of BT (broad term) and NT (narrow term) designators. This category is used to present hierarchical relationship between preferred terms in tree like structure.

(iii) Besides, equivalence and hierarchical relationship, there is another type of relationship i.e. associative relationship. Aitchison, Gilchrist and Bawden (2000) mentioned that defining associative relationship between terms is a complex task. Chowdhury (1999) cited BS 5723 to list various guidelines that can be used in establishing associative relationship between the terms. Further, notation RT is suggested for this type of relationship.
2.3.5 Thesaurus in Digital Era

Full text era refers to the environment where full information in form of books, images, journal articles etc. are available for online viewing, downloading and printing. Usually, free text searching is performed in digital environment that causes many problems in retrieving relevant information. This is because when the same information is represented with different terminology then it becomes difficult for computers as well as for mankind to reach on same conclusion. Thus, to support agreement between two agents, it is essential that there should be a common understanding and agreement for the usage of terminology (Spear, 2006). This objective of effective retrieval of information in web world could be achieved with the use of thesauri. Much has been written for use of thesaurus in Internet world. According to Aitchinson, Gilchrist and Bawden (2000, p.1) thesaurus is “a vocabulary of controlled indexing language, formally organised so that a priori relationship between concepts are made explicit, to be used in information retrieval systems, ranging from the card catalogue to the Internet”. The last word ‘Internet’ in this definition specifies the use of thesaurus in digital era. Study conducted by Shiri and Revie (2006) described the role of thesaurus enhanced system in query expansion behaviour of users on web. The study further revealed that thesaurus offered additional terms to users of which they were not aware at the beginning of the formulation of their search queries. In the study by Bhat (2013), author reviewed literature on knowledge organisation systems in digital environment. Findings of the study concluded that use of thesauri in digital environment offered improved access. Highlighting the application of thesaurus in web world, Shiri and Revie (2000) stated that the semantic structure of thesauri plays a vital role in organising and retrieving information resources available on web.

Moreover, it has been observed that thesaurus is a part in number of electronic platforms like Inspec thesaurus played a vital role for INSPEC database. With this powerful search aid, specialists can perform searches on complex topics (Institute of Engineering and Technology, 2015). Further, for SQL Server 2014, Microsoft itself is using SQL Server thesaurus to define synonyms for specific language (“Configure and manage ….”, 2015).
2.4 INTERNET

The history of information retrieval systems does not begin with Internet (Sriram, Pravallika, Neelima & Rahmathulla, 2014). It began with manual approaches like catalogue cards, indexes etc. which rely on linear organisation of information. File accessing was usually performed through File Transfer Protocol (FTP), where files were arranged in hierarchical order on FTP server. As more servers were added, task of keeping-up the central list became difficult (Seymour, Frantsvog & Kumar, 2011). Thus, the first real tool that appeared for searching on Internet was “Archie”, which appeared in 1990. This program downloaded all the files on FTP server and created a searchable database of file names (Kuyoro, Okolie, Kanu & Awodele, 2012). Further, in the next successive year i.e. in 1991, another program known as ‘Gopher’ was launched. The text based Gopher, offered menus of files and helped users in finding their information (King, 2004). After “Archie”, several other Internet search engines like W3Catalog, Aliweb, Jump Station, WebCrawler etc. also emerged (Levene, 2010).

Explaining Internet search engines, Seymour, Frantsvog and Kumar (2011) mentioned that search engines are software programs that help in searching the linked documents available on World Wide Web based on users’ query terms. Linking of documents on World Wide Web is known as hypertext, in which automatic links are created between documents on web. This hyperlinking of web documents is possible with the help of Hypertext Markup Language (HTML) (King, 2004). The main focus of current web is to index the documents and creating hyperlinks between them. Due to the missing semantics in current web environment, existing search engines appeared to be inadequate for the purpose of effective retrieval of relevant information. The situation led to the development of semantic web, the concept proposed by Tim Berners-Lee (Greenberg, Sutton & Campbell, 2003).

2.5 SEMANTIC WEB

The Web is a techno-social system through which interlinked and hypertext documents are accessed through Internet (Aghaei, Nematbakhsh & Farsani, 2012). It was created with the aim that anybody from anywhere and at any time could retrieve his/her
information. Sir Tim Berners-Lee was the one who created world’s first website in 1991 (“World’s first website---”, 2012). Since 1991, there appeared an immense progress in the manner in which information is published on web. On the basis of characteristics and features added, scholars defined various generations of the Web (Aghaei, Nematbakhsh & Farsani, 2012; Choudhury, 2014; Meilender, Lieber, Palomares & Jay, 2012; Patel, 2013).

(i) **Web 1.0**

This was the first generation of web where few developers designed web pages for large number of audience. This era was considered as read-only phase. HTML, HTTP and URI were considered as the core protocols of this stage. Web pages that were designed under this category were static in nature and had no connection with machine compatibility. Web master could only manage the content of the web pages.

(ii) **Web 2.0**

The term ‘web 2.0’ was coined by O’Reilly (Choudhury, 2014). It provided enhancement over web 1.0 through participation, collaboration, and distributed practices of people. Introduction of technologies like blogs, RSS feeds, social networking softwares, wikis, etc. add freedom of writing to users. Thus, this phase is considered as read-write phase. In other words, web 2.0 offers more participation and less control.

(iii) **Web 3.0**

The phrase web 3.0 was coined by John Markoff of the New York Times in 2006 (Spivack, 2006). It is considered as the third generation of web and is also termed as semantic web. Here, data structure is defined and interlinked in order to facilitate effective retrieval, reuse, and integration of various applications. Its features include representing meanings, hyperlinking content in a manner such that in addition to mankind, machines could also interpret data. Thus, with the introduction of semantic web technologies, representation of knowledge became more structured and meaningful.
Discussing the concept, ‘semantics’, Svenonius (2003) mentioned that semantics denotes the study of meaning. In the context of controlled vocabularies, the author has further categorised the concept as:

- **Category Semantics**: It refers to the meaningful grouping of the concepts on the basis of classes or facets which are totally exhaustive and mutually exclusive. Some examples of the category include –
  - Kaiser’s three facets (1911) (concretes, processes, and country/locality)
  - Ranganathan’s five fundamental categories (1933) (personality, matter, energy, space, and time)
  - Vickery’s eight explicit facets (1959) (i.e. kinds of soil, structure, constituents, properties, processes, operations, laboratory techniques, and general) in his Soil Science classification work.

- **Referential Semantics**: It denotes the categorisation on the basis of reference structure followed in controlled vocabularies. This is to achieve homonym and synonym free knowledge structure by establishing one-one relationships between concepts and their respective referents.

- **Relational Semantics**: It indicates the tagging on the basis of distinct relationships like equivalence, hierarchical and associative associations between the terms.

The concept ‘semantic web’ was envisioned by Tim Berners-Lee with an idea to design a web where documents are linked meaningfully (Kuck, 2004). It is considered as an extension of the current web where in addition to human understanding, web documents are marked with meta-information such that it can be processed with the help of machines. It is not just for displaying the content, but for integrating and reusing available information across various applications. To achieve semantic web environment, ontologies serve as the central enabling technology (Davies, Fensel & Hermelen, 2003).

### 2.6 ONTOLOGY AS TOOL FOR SEMANTIC WEB

Any metadata schema that represents the conceptual and/or physical description of a resource that seems to be relevant for users could be referred to as ontology. Set of
elements in Dublin Core represent ontology (Jacob, 2003). Ontologies provide a conceptual structure in machine processing language. They offer formal conceptual structures in a manner such that semantic web technologies depend on them. Ontologies offer a shared knowledge structure of common understanding for a discipline that can be communicated between humans and machines (Ding & Foo, 2002). In semantic web field, ontology building is receiving research momentum. It is an emerging technology that has potential to improve information organisation and management system. It has gained immense popularity and has been adopted widely by different communities.

Smith and Ceusters (2006) proposed ontology of biomedical informatics to overcome the difficulties appearing with different naming conventions for genes, proteins, and other molecular structures and to offer a standard terminology for the domain. Similarly, Biasiotti and Tiscornia (2010) represented legal concepts in ontological framework to support understanding, re-use and sharing of knowledge in legal domain. Some other studies reflected the development of ontology in various other disciplines. Hastings, Magka, Batchelor, Duan, Stevens, Ennis and Steinbeck (2012) presented developments of ontology in Chemistry. A study by Gokhale, Deokattey and Bhanumurthy (2011) reported methodology for constructing sample domain ontology on energy amplifiers. As a source for keywords, INIS database was employed in the study. Thomas, Pappu and Baker (2011) reported the designing and development of Nanoparticle Ontology (NPO), which is executed in Ontology Web Language (OWL). Esbjorn-Hargens (2010) proposed an ontology on climate change.

Apart from the development of ontology in distinct disciplines, usage of the concept in several disciplines has been reported. Discussing the enthusiasm of ontology among the scholars belonging to distinct disciplines, study by Patkar (2011) discussed ontology in various perspectives like Philosophy, Computer Science, and Library & Information Science.

The concept ontology is gaining recognition across distinct disciplines, in the last few years, efforts have been made towards developing tools and ideas for ontology modelling. Though there is no specific method for ontology design, modelling of ontology depends on the application in view and extensions projected for the near future (Patkar, 2011).
Malik, Prakash and Rizvi (2010) cited Denny to elaborate the following basic steps for ontology modelling:

- Attaining deep insight of the discipline in which domain ontology is to be built.
- Identification of concepts that represent the domain selected for ontology.
- Creation of hierarchical structure by creating classes, and their respective subclasses.
- Defining properties to establish relationships between the concepts.
- Verification of constructed ontology for consistency.
- Deployment of ontology for transfer and sharing of information in machine readable environment.

Several tools have been developed in the recent past for convenience in creating ontologies with formal languages. Some of these can be downloaded freely from the Internet like Apollo, IHMC Cmap, OWLGrEd, Protege, SWOOP etc. Though there are a number of ontology editors, identifying and selecting the one that works better for a particular project is always a challenging issue. Literature that compares various ontology editors on distinct parameters appeared from time to time. Study by Kamidi, Malik, Prakash and Rizvi (2008) presented a comparative study of features of various ontology editors and focused on Protege for designing ontology. Analysis done by Khondoker and Mueller (2010) revealed that a number of users support Protege because of online help available through mailing list. A comparative study by Kapoor and Sharma (2010) was conducted to compare four ontology editors namely Protege 3.4, Apollo, IsaViz and SWOOP on parameters like interoperability, openness, easiness to update and maintain, and market status and penetration. The study concluded that each software application provides different functionalities, therefore, selecting the better alternative depends on the project goal and the ease with which it can be maintained and expressed easily. Survey conducted by Abburu and Babu (2013) presented a comprehensive analysis of ontology development tools. Many features like mode of availability, collaboration, OWL editors, extensibility etc. were compared in the study.
In addition to ontology editors, several other semantic web technologies have been considered to implement ontologies. These technologies are referred to as ontological or knowledge representation languages. Usually, ontology languages use XML as base for their syntax. According to World Wide Web Consortium (W3C) recommendations, RDF and OWL are considered as semantic web standards for ontologies (“World wide web---”, 2004).

2.7 THESAURUS AS AN AID FOR ONTOLOGY CONSTRUCTION

Thesaurus is similar to ontology in a sense that both the systems are used to denote domain concepts and relationships between the concepts. But when we talk about semantic rich and machine processing environment, thesaurus is found to be less expressive. Thesaurus provides conceptual framework and vocabulary for a discipline. Whereas, ontology offers additional and refined semantics besides what a thesaurus provides.

Number of studies highlighted a clear demarcation between the two. Contribution by Kless, Milton, Kazemierczak and Lindenthal (2014) examined the structural similarities and differences between a thesaurus and ontology in OWL syntax. The study suggested that both, thesauri and ontologies should be treated as two distinct conceptual models. They further highlighted that for a thesaurus to be considered as ontology, structural and content changes need to be incorporated, and vice versa. Similarly, Gokhale, Deokattey and Bhanumurthy (2011) in their work presented a well-defined comparison between a thesaurus and ontology.

Though, thesaurus and ontology differ from each other in formal structure and level of semantics, but still because of the intrinsic feature of thesaurus to represent the knowledge structure, it can serve as base for ontology development. Literature in the field revealed that ontology may be created by using the concepts listed in thesaurus and by refining the relationship between the concepts (Albert & Steiner, 2005; Hilera, Pages, Martinez, Gutierrez & de-Marcos, 2010).

Chun, GuoJian and GuangDa (2008) in their study reported conversion of a Chinese Agricultural Thesaurus (CAT) to the agricultural ontology. Thus, thesaurus serves as
domain knowledge. The paper further briefed the efforts made in the direction during the successive year i.e. in 2004 when crop part of CAT was converted into RDF style ontology. Again in 2006 attempt was made to convert entire parts of CAT to OWL style ontology by using a new program. Further, based on the OWL agricultural ontology, an intelligent search engine was developed that can add equivalent class automatically while searching for a concept. This system can also expand the search concept or give more relative search concept based on the associative relations of the thesaurus. In a similar direction, reengineering of the existing knowledge resources such as the thesaurus was suggested by Weicong, Jia, Zheng, Cao, Yang and Yu (2009). Based on specific domain thesauri, this paper discussed the method of reengineering the thesauri to ontology and the method to obtain the correct semantic relations. A traffic thesaurus was selected as knowledge base resource and Protege was selected for the development tool along with OWL as the ontology language. To reduce the workload of ontology construction Li and Li (2012) proposed transformation of thesaurus into domain ontology.

2.8 CONCLUSION

Analysis of literature on thesaurus and ontology revealed that both the controlled vocabulary tools played a vital role in information organisation and retrieval activities of professionals and scholars disbursed around the globe. Many studies depicted the requirement of controlled vocabulary tools in the networked environment where due to the expanding nature of web in terms of literature appearing from all streams, it becomes difficult to properly manage and retrieve the desired content.

Further, strength of natural language system over controlled vocabulary tools in Internet world like currency of language, coverage, high recall etc. made scholars to raise their voice in support of natural language systems. On the other hand, some important factors like lack of precision due to exhaustivity, absence of association between the terms, lack of terminological standardisation due to the use of synonyms and homonyms, forces educationalists to define the new role of controlled vocabularies in full text era.

Moreover, research in interdisciplinary areas like Nanotechnology is among the rapidly growing fields in Science wherein there is no single approach for research. Scholars
working in different disciplines are using their own terminology to represent the same concept. Thus, it becomes difficult to reach the same conclusion. Some relief has been achieved with the terminology standardisation initiated by organisations like ASTM, ISO etc. However, no study have been traced by the researcher where efforts have been made to define the semantic relationships between the Nanotechnology concepts with their scope limited to Physical Sciences. So, by developing Nanotechnology thesaurus and using it for ontology, the present study aimed at representing and expressing semantic knowledge map for Nanotechnology concepts, especially for Physical Sciences.
REFERENCES


human indexing. *Information Processing and Management*, 37(2), 231-254. doi:10.1016/S0306-4573(00)00026-1


Levene, M. (2010). *An introduction to search engines and web navigation*. Hoboken, New Jersey: John Wiley. Retrieved from https://www.google.co.in/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&ved=0CQFjAA&url=http%3A%2F%2Fwww.google.co.in%2Furl%3Fsa%3Dt%26rct%3Dj%26q%3D%26esrc%3D%26source%3Dweb%26cd%3D1%26ved%3D0CQFjAA&url=http%3A%2F%2Fluwj.googlecode.com%2Fsvn%2Ftrunk%2FEbook%2FAn%2520Introduction%2520to%2520Search%2520Engines%2520and%2520Web%2520Navigation%25202nd.pdf%3Fei%3DAhjTVKPlOeG5mwWe4oKgBQ&usg=AFQjCNHmbD-OtAwXfwoHETXQQnTfsDZC1lwv sig2=hl5vDMmYU3l89cHOGlaLhQ


