CHAPTER - 4

METHODOLOGY AND DATA COLLECTION
4.1 Methodology

The primary purpose of research is to discover principles that have universal application. Research, therefore, needs adequate and accurate data for this purpose. In order to obtain these data, a researcher conducts investigations into a given population. Information thus derived with sufficient accuracy from a sample of the total. Sampling refers to the investigation of a part of the whole population or universe. A sampling procedure is a technique of selecting a sample from a given population. Sample means a small population of a universe selected for study. While using secondary data for research studies, it is essential that these data are properly edited. As the data have been collected by different agencies, there are certain restrictions for using as such for analyses. Hence there exits a need to change the format of the data to suit our own requirements. Computerised packages can be used as tools for accurate analysis of data.

4.1.1 Choosing the sample

Sample for the study is selected by first searching various entries of articles and their citations abstracted in the online database on Toxicology, TOXLINE. As TOXLINE is updating on monthly basis, numerous entries are added to it. As this number was very high, the informetric analysis cannot be concentrated on all these data, the period of study was restricted to 1998 January to December 2003.
4.1.2 Sampling Technique

The method used is random sampling. The articles and their citations abstracted in TOXLINE during 1998 to 2003 was accessed through the Internet and downloaded to a CD-ROM. A total of 2247 articles and 9585 citations were collected. This data is used for informetric analysis.

To study about journals, a rank list of 59 journals is prepared based on their frequency of occurrence in the database. All these journals are having wide subject coverage with almost all areas in Toxicology and are international in scope. From the ranked list 9 core journals are selected for core journal study having 1218 articles and 3411 citations.

4.1.3 Variables for Analysis

The following variables are employed for making detailed analysis:

- Authorship pattern
- Subject dispersion
- Language and types of documents
- Core books
  - Ranking of books
  - Currency of book citations
  - Time lag of citing books
  - Subject dispersion in book citations
  - Core authors
  - Productivity of authors relating to Lotka’s law.
- Core journals
  - Decade-wise distribution of journals.
- Subject wise scatter of journals.
- Fit of rank distribution (Bradford Distribution)
- Country-wise scatter of journals
- Age of journals most cited

- Growth study
  - Rate of citation per article
  - Citing and non-citing articles in different journals
  - Growth of citing and non-citing articles

- Bibliographic coupling
  - Year-wise analysis of bibliographic coupling
  - Strength of bibliographic coupling

- Self-citation
  - Year-wise trend
  - Types of documents
  - Repetitiveness of self-citation
  - Age of self-citation
  - Recency of self-citation with five major categories of authors.
  - Self-citation habit of Indian & foreign authors
  - Categories of self-citation
  - Self-citing articles in different journals

4.2 Data Collection

Several on-line databases on specialised areas of Toxicology are available such as EMIC (Environmental Mutagen Information Center) & ETIC (Environmental Teratology Information Center), CANCERLIT
(CANCER - Related Literature), AGRICOLA (Agricultural On-Line Access), NIOSHTIC (National Institute for Occupational Safety and Health) etc. Among these TOXLINE deals with general topics on Toxicology. TOXLINE is an international, authoritative online database covering almost all the major aspects of toxicology. TOXLINE is provided by National Library of Medicine, (NLM), Washington, can be occurred through Internet at free of cost. Hence the data required for this study was collected from TOXLINE.

4.2.1 TOXLINE (Toxicology Information online)

TOXLINE is the National Library of Medicine's extensive collection of on-line interactive bibliographic database of Toxicology information since 1972 [1]. TOXLINE is covering the biochemical, pharmacological, physiological, toxicological and environmental effects of drugs and other chemicals [2] TOXLINE contains more than 3 million bibliographic citations, almost all with abstracts and/or indexing terms and CAS Registry Numbers. TOXLINE is a merged database composed of subsets from 13 different sources. Files from 11 secondary sources comprise the TOXLINE database. The sources supplied bibliographic records in different formats and data structures. Data from each supplier's format had to be converted into a format suitable for TOXLINE. Most of the citations in TOXLINE refer to journal articles; other publications types include meeting reports, monographs, patents, research projects reports, technical reports and theses. TOXLINE references are drawn from various sources grouped into two major parts - TOXLINE core and TOXLINE special - both of which operate under versatile search engines offering a variety of search and display capabilities [3].

4.2.1.1 TOXLINE Core

TOXLINE Core covers much of the standard journal literature in Toxicology. TOXLINE Core is a separately searchable subset of the much larger MEDLINE database of biomedical literature and is accessible through the PubMed system by selecting 'Toxicology' as a subset limit.
TOXLINE Core takes advantage of Pub Med’s powerful search and retrieval engine to offer features such as MeSH term Selection, document delivery, saving search queries via Cubby, and linking out. A search link to TOXLINE Core is also available from the TOXLINE interface on TOXNET, as described under TOXLINE special below.

4.2.1.2 TOXLINE Special

As the name implies, TOXLINE Special is a component of TOXLINE, available on the TOXNET system at http://toxnet.nlm.nih.gov/.

TOXLINE Special complements TOXLINE Core with references from an assortment of specialized journals and other sources. TOXNET allows to facilitate the search query between TOXLINE Special and TOXLINE Core. An array of special features on TOXNET such as relevancy ranking, flexible sorting and downloading options adds to TOXLINE Special's search effectiveness. The constituent component of TOXLINE Special are listed below:

- Special journal and other research literature:
  (a) Development and Reproductive Toxicology (DART)
  (b) International Labour Office (CIS)
  (c) Swedish National Chemicals Inspectorate (RISKLINE)

- Technical Reports and Research Projects
  (a) Federal Research in Progress (FEDRIP)
  (b) Toxic Substances Control Act Test Submissions (TSCATS)
  (c) Toxicology Document and Data Depository (NTIS)
  (d) Toxicology Research Projects (CRISP)

- Archival Collection
(a) Aneuploidy (ANEUPL)

(b) Environmental Mutagen Information Center File (EMIC)

(c) Environmental Teratology Information Center File (ETIC)

(d) Epidemiology Information System (EPIDEM)

(e) Hazardous Materials Technical Center (HMTC)

(f) International Pharmaceutical Abstracts (IPA)

(g) National Institute for Occupational Safety & Health (NIOSH)

(h) Pesticides Abstracts (PESTAB)

(i) Poisonous Plants Bibliography (PPBIB)

(j) Toxicological Aspects of Environmental Health (BIOSIS)

4.2.2 Accessibility

 Previously, TOXLINE had been accessible in its entirety via the Elhill and TOXNET SYSTEMS. With the implementation of TOXLINE's current configuration, the file became segmented into TOXLINE core and TOXLINE special. TOXLINE core now essentially replaces TOXLINE'S former TOXBIB sub-file. TOXLINE'S BIOSIS and IPA have become static components and are no longer being updated [4], [5].

 More information about TOXLINE can be had at

 Specialized Information Services

 National Library of Medicine

 8600, Rockville Pike

 Bethesda, MD 20894

 Telephone (301) 496-1131
4.2.3 Retrieval of Information

Three different successive retrieval systems were used for the TOXLINE database which required reformatting of the data. Algorithms for generating terms for inverted file search methods were tested. Special characters peculiar to the scientific literature were evaluated during search term generation. Developing search term algorithms for chemical names in the scientific literature required techniques different from those used for non-scientific literature. Problems with replication of bibliographic records from multiple secondary sources are described.

4.3 Computer programmes used

Although the collected data is in a standard format, the statistical analysis through SPSS is impossible. So the data is first converted to CDS/ISIS format and then analysed using SPSS. The data include a total of 2247 articles and 9585 citations. The major programmes used are:

4.3.1 CDS/ISIS

CDS/ISIS is an application software developed by the Computer Division of UNESCO in 1985 for the PC environment using MS-DOS. CDS/ISIS is a powerful software for database creation and information retrieval. The present version being 3.07 can be used in a networked environment. The software is distributed free by UNESCO for non-profit organisations in developing countries. In India, NISSAT, Department of Scientific and Industrial Research, New Delhi is the nodal centre for distribution.
> **Capabilities**

CDS/ISIS allows

- define databases containing the required data elements.
- enter new records into a given database.
- modify, correct or delete existing records.
- automatically build and maintain first access files to each database.
- retrieve records by their contents, through a sophisticated search language.
- sort the record in any sequence required.
- display the records or portions thereof according to out requirements.
- print partial or full catalogues, and/or indexes from any given database.

> **File structure**

CDS/ISIS has got two major files:

- **Master file**

  Master file contains all the records of a given database, each record consisting of a set of variable length fields. Individual records are identified by a unique number automatically assigned by CDS/ISIS and SUPERDOC.

- **Inverted file**

  Inverted file provides additional ways of accessing the Master file. Contains all terms which may be used as access points during the
retrieval for a given database as well as the list of corresponding postings. The inverted file is essentially an index to the contents of the Master file.

4.3.2 SPSS (Statistical Package for Social Sciences)

SPSS is designed especially for the analysis of Social Sciences data. SPSS is a package of programmes covering a wide range of analysis. These programmes include frequency distribution, univariate and multivariate analysis variance, regression analysis, correlation analysis, factor analysis, cluster analysis, reliability analysis, long linear models etc. The sub commands direct precisely what analysis had to be performed. The GROUPS sub-command specifies which groups are to be discriminated between and the VARIABLES. Sub command specifies which variables are to be in operation.

The mainframe, PC and windows versions of SPSS are available. Internet, Discussion Group for SPSS Users has been formed to share and interact each other's experiences. The special characteristics of SPSS package in data analysis are:

- SPSS package has the provision to generate the summary of statistics including mean, median, mode, standard error, standard deviation, variance, range, skewness, kurtosis, minimum and maximum;

- SPSS package allows univariate distributions as well as bivariate distributions. The association between two variables can be done by bivariate analysis. The three major functions of multivariate analysis i.e. regression and multiple regression are also achieved through equations and subprograms. The bivariate tables and statistics for nominal and ordinal measures can be done with CROSSTABS procedure. The CORERELATIONS subprogram provides interval measure;
• From the data analysis any type of graphic representation can be derived from the graphic portion of the package. The Windows version provides a wider range of choices in graphic modules. The information on the sub programmes and their uses are provided by SPSS User's Guide. SPSS Windows Ver.6 is a well integrated Windows program that adheres to many typical windows conventions and commands. The typical windows commands File, Edit, Help, New, Open, Save, Exit, etc. are used in the package.

4.3.3 Standard Format

The basic format of data retrieved from TOXLINE is shown below. The articles and the citations are the basic unit of study. Articles include original research papers, short communications, pre-prints, review articles, technical notes etc. For each article and citation the following data elements were entered in the computer.

1. Name of the author/s
2. Title of the article
3. Source (Name of the journal)
4. Volume and Page Number
5. Publication Year
6. Title Abbreviation
7. Key words
8. Publication type
9. Total number of citations

A sample entry downloaded from the computer is given below.

1. Name of the author/s Lee PA; Freeman S
2. Title of the article  

Allergic contact dermatitis due to para-tertiary-butylicatechol in a resin operator.

3. Source  

Australas Journal of Dermatology


Vol. 40 & P 49 - 50

5. Publication Year  

1999

6. Title abbreviation  

Aus J Dermato

7. Key words  

Dermatitis, Occupational - Diagnosis, Allergic Contact, Synthetic - Adverse Effects.

8. Publication type  

Journal Article

9. Total number of citations  

12

A total of 9585 entries were prepared. Majority of the articles and citations were in English language. Articles in foreign languages having their abstracts and titles in English were also found. Such articles were also taken for study.

4.4 Organization of Data

The data of citations was sorted to get author index, keyword index and cumulated total citations. From the authors, individual, double, and more than two authors articles citations are counted for studying author collaboration (Table 5.1.1.2, 5.1.1.3, 5.1.1.4, 5.1.1.5, 5.1.1.6, 5.1.1.7.). Keyword is used for preparing subject wise scatter of citations (Table 5.2.3). Language-wise and publication type index is used for the study of language and type of document cited (Table 5.3.2, 5.3.3, 5.3.4). The correlation and correlation coefficient of book and journal citations was done (Table 5.3.5).
The total number of book citation studied is 1174. The data collected was sorted by the name of the book, author, year and number of occurrence of these elements in the data. The study was focussed on ranking of books, core books (Table 5.4.2), currency of book citations (Table 5.4.3), subject dispersion in book citations (Table 5.4.5). From the author index, number of authors in each rank, ranking of authors (Table 5.5.2, 5.5.3), the productivity of authors is done using Chi-square test in relation to Lotka's law and K - S statistical test (Table 5.5.4, 5.5.5.), author contribution vs number of authors, values of observed and expected number of authors (Table 5.5.6,5.5.7) are presented.

The total number of journal citation studied is 6922. The data collected was sorted by the name of the journal, author and year which gave the number of occurrence of these elements in the data. This study is more concentrated on number of authors, number of journals, subject scatter, age of citation, self-citation, bibliographic coupling etc.

On the basis of citations received, a rank list and rank distribution of 59 most cited journals is prepared (Table 5.6.2,5.6.3). Using this index Bradford's theory of distribution was tested with the core journal data (Table 5.6.4). An alphabetical list of most cited journals was prepared (Table 5.6.5) with year of publication, country of publication and periodicity to study the year-wise, country-wise distribution and publication periodicity. The data elements were taken from 'Ulrich's International Periodical Directory' and 'World List of Scientific Serials'. Using this data, most cited journals by year of origin (Table 5.6.6), decade-wise distribution of journals (Table 5.6.7), subject-wise scatter of most cited journals (Table 5.6.8), distribution of citation by subject of most cited journals (Table 5.6.9) were prepared. Graphs were plotted from these data showing Bradford distribution (Fig. 5.6.1, 5.6.2, 5.6.3). The country wise distribution of most cited journals, age of journals (Table. 5.6.13, 5.6.14) were also presented. Taking the first 9 journals from the core journal rank list, the growth of Toxicology literature and citations was analysed. The above 9 journals
have a total of 1218 articles and 3411 citations. By using the above data, rate of citation per article (Table 5.7.2), percentage of contribution of citation in different journals (Table 5.7.3), rate of citation per article during different years (Table 5.7.4), citing and non-citing articles in different journals (Table 5.7.5), growth of citing and non-citing articles (Table 5.7.6.) were compiled. Bar diagram was prepared from the data showing citations per article of individual journals (Fig.5.7.1). Graphs were plotted from these data showing citation per article (Fig.5.7.2) and comparison of citing and non-citing articles (Fig.5.7.3). Relative growth rate and doubling time for citations and publications (Table 5.7.7, 5.7.9) were compiled. Graphs were prepared to study the growth pattern (Fig 5.7.4, 5.7.5, 5.7.6, 5.7.7).

To study bibliographic coupling, the above 9 core journals are taken. From the above data, bibliographic coupling in Toxicology journals (Table 5.8.2), bibliographic coupling during different years (Table 5.8.3) and strength of bibliographic coupling (Table 5.8.4) were prepared.

To study self-citation, the 9 core journals are taken. The self-citation in different journals (Table 5.9.1), year wise trend of self-citation (Table 5.9.2), types of documents self-cited (Table 5.9.3), repetitiveness of self-citations (Table 5.9.4), age difference of self-citations (Table 5.9.5), recency of self-citation within five major categories of authors (Table 5.9.6), self-citation habit of Indian and foreign authors (Table 5.9.7), rate of self-citations in different categories of self-citing articles (Table 5.9.8), categories of self-citations (Table 5.9.9), self-citation habits of single and two authored categories in different journals (Table 5.9.10), self-citation habits of three and more than three authored categories in different journals (Table 5.9.11), year-wise distribution of categories of self-citations (Table 5.9.12), sequence of cited authors and self-citations (Table 5.9.13), self-citing articles in different journals (Table 5.9.14) and year-wise self-citing articles (Table 5.9.15) were compiled. Graphs were plotted showing different aspects of self-citation (Fig 5.9.1, 5.9.2, 5.9.3, 5.9.5). Bar diagram
were drawn to show the percentage of self-citing articles in different journals (Fig. 5.9.4).

The data required for the study was collected and organized by using the above methods. Through the analysis of the collected data, 55 tables and 25 graphs were prepared. This formed the basis for informetric analysis to test the hypothesis and objectives and formulate major findings and conclusions.
References


