

Desktop Scientometrics

J. Sylvan Katz
Diana Hicks

*ESRC Centre for Science, Technology, Energy and Environment Policy,
Science Policy Research Unit, University of Sussex, Falmer, Brighton, BN1 9RF, UK*

Advanced scientometric tools are moving from the realm of the privileged few with access to mainframe and minicomputers to the desktop of researchers equipped with personal computers. This shift is not only due to the decreasing cost and technological advances in PCs but the ready availability of a powerful multitasking operating system, a versatile text processing language and easy access to the Internet. Furthermore, the latest releases of PC software, such as Microsoft Excel, make it possible to develop graphical user interfaces into complex bibliometric data for a wide spectrum of researchers and policy analysts. Recent developments in digital communication, in particular, tools to access the Internet via the World Wide Web will provide give even greater flexibility to those researchers wishing to make their scientometric data available to a diverse international audience. This paper examines how the BESST project developed a Desktop Scientometric environment using public domain, hardware independent software, prototyped a graphical user interface to provide easy access to UK sectoral level bibliometric data and gives a glimpse into future developments.

1. Introduction

Scientometrics is a unique research area and scientometricians are a unique breed of scientist who endeavour to quantify national and international systems of innovation to help policy shapers weave the political and economic climate required to nurture an R&D community in hopes of deriving long term economic and social benefits. I shall focus on the bibliometric aspects of scientometrics. Bibliometrics, in general, is the art of exploring publication databases in search of indicators that reflect research productivity and quality as well as the interactions between individuals, groups, institutions, sectors and nations.

In 1993 SPRU initiated the *Bibliometric Evaluation of Sectoral Scientific Trends* better known as the BESST Project¹. This project explores the publication trends from six UK institutional sectors and the collaboration patterns between them. The data source is article, note and review publication types indexed in the 1981-1994 Science Citation Index® and extracted from tapes purchased from the Institute for Scientific Information. I shall explain how we used the art of *Desktop Scientometrics* to manipulate, unify, analyse and display publication and collaboration dynamics in the UK science system using a desktop PC

equipped with three public domain software tools: Linux (a Unix operating system), Perl (a powerful text processing language) and emacs (a programmable line editor). Also, I will explain a prototype *Scientometric Graphical Interface* designed to provide researchers and analysts with an easy-to-use tool to explore sectoral level data². Finally, I will explore the expanding boundaries of desktop computing - Internet, World Wide Web, JAVA and distributed databases - and explain how these will fundamentally change the nature of scientometrics and further drive new developments in desktop scientometrics.

Until recently, the manipulation of large bibliographic databases, that is more than a few tens of thousands of publications, was out of reach of most scientometricians for three primary reasons:

1. the high cost of the data;
2. the need to have access to a mainframe or minicomputer with a large complement of main memory and hard disk space; and
3. a limited selection of programming languages suitable for processing text quickly and efficiently.

Recently, data costs have dropped with the availability of publication data on CD-ROM. However, direct access to the CD-ROM data is prohibited and access using the vendor's bibliographic interface can be slow and cumbersome. Furthermore, in some instances the format and amount of the data on CD-ROM differs markedly from that the on tape³. On the other hand, the new generation of personal computers and recent developments in text processing languages can bring scientometrics to most desktops. Let me explain further.

2. The BESST Project

In the Phase I, the BESST Project explored the relationships within the UK scientific community using publication indicators derived from the 1981-1991 *SCI*. More specifically, the objectives were

1. to determine the share of national scientific output in various scientific fields contributed by different institutional sectors (e.g. education, medical, industry, research council, government and non-profit);
2. to map the changes in patterns of inter- and intra-sectoral collaboration in different scientific fields;
3. to investigate changes in the patterns of international collaboration with UK sectors, and;
4. to use the data to investigate policy-relevant questions.

These objectives were met in June 1995 and Phase II was started with the following objectives:

1. extend the database coverage to 1994
2. develop a graphical interface to provide user-friendly access to sectoral level data, and;
3. explore in detail the publication and collaboration activity of UK industry.

Presently, we have met objectives 1 and 2 and we have almost completed 3. Before I explain how we manipulated and analysed the data, let me ask you to think about the magnitude of the tasks we faced. In general, we wished to explore fourteen years of *SCI* data containing the names of about 6000 unique UK institutions (about 50,000 variants) each classified into one of six institutional sectors. Furthermore each publication was classified into one of seventeen science fields and foreign collaborators into one of eight international regions.

Recall, we want to examine publication time trends by sector in each scientific field and to explore changing patterns in intra-sectoral, inter-sectoral and international collaboration. Now, let me ask “how many tables and graphs does it take to reasonably portray publication and collaboration dynamics in the UK science system?” Remember, that each table can be normalised in a number of different ways. For example we may want to express the publication trends by sector in a science field as a percent of UK publications or simply as a percentage of publications in the field. Or we may wish to express the number of industry-education collaborative papers (i.e. papers that list institutions in the education and industry sectors) as a percentage of industry's papers or percentage of education's domestic collaborations.

There are thousands of such tables and each one tells a story. However, there are only two of us, Diana and myself, currently exploring the BESST database and between us we can only tell a few of them. We want to make our data available to other researchers so they can tell their own stories or to analysts who wish to answer specific questions about the UK science base. We decided to prototype an easy-to-use graphical interface into the BESST sectoral data complete with on-line documentation using Excel 5.0. I shall describe the graphical interface and database which allows the user to explore about 2,000 graphs and tables before describing the details of the methodology we used to produce the database. Our long term objective is to make the tables and graphs accessible over the net from an interactive World Wide Web server that interrogates an Oracle database.

In more detail the BESST Database was built in Excel 5.0 and is composed of a *database* workbook, a *templates* workbook and a *graphical interface* written in Visual Basic for Applications and supplied as an Excel Add-In . The database is read-only and the templates workbook which contains table layout template but no data is password protected. This has been done to maintain the the integrity of the workbooks. The Add-In can not be opened because it contains the compiled version of the software for the graphical interface. The graphical interface is supplied with a complete on-line help system.

4. Desktop Scientometric Tool Kit

Now, I shall explain in how we unified, manipulated and analysed the SCI data. First, a little background. Before the project was initiated SPRU had already purchased the 1981-1989 SCI on tape from ISI. We obtained funds to update the data through 1991 in Phase I and to 1994 in Phase II, pay the salaries of 1.5 person for 2 years and purchase a PC.

In summary, we needed to extract about 500,000 article, note and review SCI publication types from 600 MB of tape data, isolate the names of institutions listing a UK address, manually unify them to a set of standard names and link the standard names back to the original dataset. However, we faced a dilemma. About the time the project began the University of Sussex computing services started to experience a tremendous increase demand on its facilities so large computing resources were scarce. Furthermore, considerable experience had shown us that although Microsoft products running under Windows was adequate for analysing paper, citation and collaboration counts it was not suitable for manipulating large text database. So we decided to built a *Desktop Scientometric ToolKit* on a personal computer that consisted of three main components: a multitasking operating system, a text processing language and a programmable line editor. I shall briefly describe each of these in turn.

Linux Operating System

Linux⁴ is an independent implementation of a POSIX compliant 386 based Unix operating system freely distributed under the GNU Public License⁵. Furthermore, much of the GNU project Unix freeware has been bundled in with the Linux distribution. We gambled on Linux since it contained most of the standard string manipulation tools need to manage large datasets. We purchased a 66 Mhz 486 with 32 MB RAM, a 1 GB SCSI hard disk with tape drive and installed Linux. This is a slow and small machine in compared to today's 150 MHz Pentium. Recently we installed a new version of Linux from CD-ROM in less than a day. The documentation is better than older versions but it is still requires that the user have some hardware and Unix system administration knowledge.

Perl Language

Most bibliometricians that work with the *ISI* tape data write their text processing software either using the built-in string manipulation functions in their relational database or using conventional programming languages like Fortran, Pascal and C. Database languages although ideal for simple text processing tasks can be difficult to use for certain mathematical problems. For example, it is difficult to write a routine using a relational database's SQL to produce co-occurrence matrices. Conventional languages are ideal for handle mathematical processes but have poor string handling ability.

However, in the late 1980s Larry Wall at the NASA Jet Propulsion Laboratory developed a script writing language for system administrators called Perl⁶, the Practical Extraction and Report Language. This language was created specifically for manipulating large files and makes extensive use of the string routines that are inherent in the Unix environment. Perl is now considered to be a 'must know' language for text manipulators and Web site maintainers and developers. Perl is a 'glue' language, that is it can be used as a stand alone programming language or to glue the power of programming languages like C or Fortran to a database like Oracle or Microsoft Access. It is also distributed free under GNU Public Licence.

Perl has a rich collection of text handling, pattern matching and array and database manipulation functions. Without getting too technical let me illustrate its versatility for bibliometric analysis. I will briefly describe two functions that were invaluable for constructing the BESST database.

Splitting and joining strings is a common text processing function. Assume we have a line of text, perhaps an *SCI* corporate address, that is composed of five primary fields each separated by a slash (/). For example, consider the corporate address from an *SCI* record that has been assigned to the variable \$ISI_line (note the \$ which commonly used as the first character in a Perl variable):

```
$ISI_line = "QUEENS UNIV BELFAST,CTR MED BIOL,DEPT PHARM,97  
LISBURN RD/BELFAST BT9 7BL/ANTRIM/NORTH  
IRELAND/"
```

SPLIT, a Perl function, divides a string into an array of strings delimited by a pattern of characters and has the form

```
@array = SPLIT(/pattern/, expr)
```

Thus, the Perl command

```
@ISI_field = SPLIT('/', $ISI_line)
```

splits the string stored in expression \$ISI_line into pieces separated by a slash (/) and stores the result in the array @ISI_field as follows:

```
$ISI_field[0] = "QUEENS UNIV BELFAST,CTR MED BIOL,DEPT  
PHARM,97 LISBURN RD"  
$ISI_field[1] = "BELFAST BT9 7BL"  
$ISI_field[2] = "ANTRIM"  
$ISI_field[3] = "NORTH IRELAND"
```

JOIN , the inverse of split, can be used to join an array of strings into one string . It is a laborious and time-consuming task to write general functions like these using conventional programming languages. Fortunately, Perl has a rich set of functions of this type.

Associative arrays is the most valuable data structures in Perl. Unlike a conventional array which is indexed by a number, an associative array is indexed by a string. For example an associative array might look like the following:

```

$array{"UNIV SUSSEX"} = "UNIV SUSSEX"
$array{"IMPERIA COLL"} = "UNIV LONDON,IMPERIAL COLL"
$array{"SUSSEX UNIV"} = "UNIV SUSSEX"

```

The index into the second element of this associative array is the string "IMPERIA COLL" which contains the string "UNIV LONDON,IMPERIAL COLL". The latest release of Perl allows arrays with multiple associative indices. An associative array is useful for holding a thesaurus of variant and standard institutional names in order to produce unified bibliometric data.

Emacs Programmable Line-Editor

In order to assign each UK institution listed on a paper to an institutional sector we had to unify institutional names to a set of standardised institutional names. This is a laborious manual process. Using a well-defined set of rules⁷ and simple Perl programs, UK corporate names were extracted from the SCI data. We needed a tool to use for manually assigning each of the approximate 50,000 variant names a standardise name. For example, the following list some of the variant names that we found for the *Bethlem Royal Hospital & Maudsley Hospital* and *University of London, Imperial College*:

```

# LONDON
% BETHLEM ROYAL HOSP & MAUDSLEY HOSP:S
  BETHLEHEM ROYAL HOSP
  BETHLEM MAUDSLEY HOSP
  MAUDSLEY & BETHLEM ROYAL HOSP
  BETHLEM ROYAL & MAUDSLEY HOSP
  MAUDSLEY HOSP & INST PSYCHIAT
  BETHLEM ROYAL HOSP & MAUDSLEY HOSP

% UNIV LONDON,IMPERIAL COLL:U
  IMP COLL
  IMPERIA COLL
  IMPERIAL COLL
  IMPERIAL COLL CTR ENVIRONM TECHNOL
  IMPERIAL COLL SCI & MED
  IMPERIAL COLL SCI & TECHNOL & MED
  IMPERIAL COLL SCI MED & TECHNOL
  UNIV LONDON,IMPERIAL COLL
  UNIV LONDON,IMPERIAL COLL SCI TECHNOL & MED
  UNIV LONDON IMPERIAL COLL SCI & TECHNOL
  IMPERIAL COLL SCI TECHNOL & MED
  IMPERIAL COLL SCI TECHNOL & MED
  UNIV LONDON IMPERIAL COLL SCI & TECHNOL

```

At first glance one might consider using a word processor to manipulate the list of names but experience has demonstrated that although word processors are ideal for manipulating text in documents, line editors are better suited for manipulating lists. Emacs, also free under GNU Public Licence, is a programmable line editor⁸. The programmability was useful because we could tailor it to our specific needs. For example, the city name is prefaced by a # symbol and each standard institutional name with a % symbol. We programmed Emacs to recognise these symbols and automatically display the city name in one color, the standard name in a second color and the variant names in a third color. Simple, things like this greatly assist the data cleaner to discriminate between various elements in the list.

Essentially, our tool kit is complete. Using Linux, Perl and emacs on a personal computer we were able to manipulate, unify and analyse 500,000 publications (about 8% of the SCI data) published over fourteen years so that we could derive the BESST sectoral level data.

BESST Project Unification and Analysis Scheme

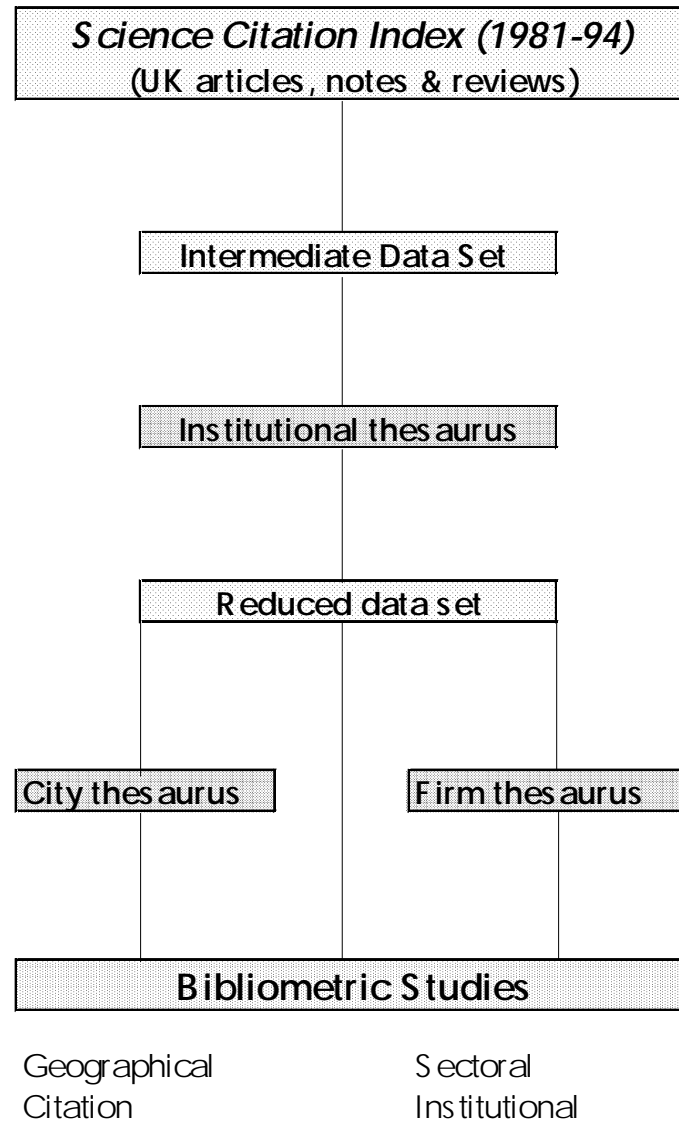


Figure 1

5. Overview

Figure 1 gives a schematic overview of the general processes we followed using our tool kit to produce the underlying master database from which we derived our sectoral data. Basically, we extracted the corporate addresses from SCI articles, notes and reviews to produce an intermediate data set. From this we extracted UK institutional names and manually unified them to produce a thesaurus of standard names. Finally, the thesaurus was used to produce a reduced database for bibliometric studies.

The reduced bibliometric database was designed for easy use and to be imported into a relational database. Each publication record in the database begins with a unique six character *ISI* article identifier which links it back to the original *SCI* data set. In addition, the number of citations from the publication year to 1994 has been incorporated to facilitate citation analysis. The format of a unified publication record is as follows:

ISI code:Journal name:number institutions:number authors:tape year:source
year:publication type:volume:start page

ISI code:Annual citation counts from current year to 1994

ISI code:Country list, number of institutional address in each country

ISI code:Unified UK Institutional name:Sector:City:Postcode

The following example listed two publications extracted from the 1981 data set:

NF9SEO:CARIES RES:3:1:81:81:N:0015:0070

NF9SEO:0:5:7:5:3:6:2:3:2:4:0:2:1:1

NF9SEO:UK,1

NF9SEO:UNIV LEEDS:U:LEEDS:LS2 9JT

NK1K1A:BR J HAEM:11:5:81:81:A:0047:0133

NK1K1A:2:12:17:11:9:13:10:9:10:4:6:5:3:4

NK1K1A:UK,4:USA,1

NK1K1A:ROYAL FREE HOSP:H:LONDON:NW3 2QG

NK1K1A:UNIV COLL & MIDDLESEX HOSP:H:LONDON:WC1E 6JJ

NK1K1A:WESTERN INFIRM:H:GLASGOW:G11 6NT

NK1K1A:UNIV COLL & MIDDLESEX HOSP:H:LONDON:N18 1QX

One year of bibliometric data in this reduced form occupies about 4 megabytes (or less than 1 megabyte in compressed format). Although there is a fair amount of redundancy in this data structure it can be eliminated when imported into a relational database.

6. The Expanding Boundaries of Desktop Scientometrics

The Internet, the World Wide Web, Java and distributed databases are expanding our desktop (Figure 2). A few of months ago we copied the BESST database and our ToolKit to a Cray Superserver⁹ at the University of Manchester. Everything we develop on one

machine runs smoothly on the other. For example, a Perl program developed by Nick Winters, my researcher assistant, running on a new Pentium can produce a series of snapshots of collaborative relationships between 40 UK industrial sectors and 160 top UK research institutions by examining joint institutional authorship on the 40,000 papers that a UK participated in between 1981 and 1994. Taking as input the reduced BESST database it produces as output fourteen 200 x 200 co-occurrence matrices that capture the dynamic evolution of research co-operation between industrial sectors and key research institutions and between the industrial sectors themselves. This takes less than 30 minutes on the Pentium. On the Cray Superserver we can produce snapshots of the co-operative research activity between 6,000 UK institutions in any one or all of seventeen science fields in a couple of hours.

Also we have established a new relationship with a super computing group at the University of Southampton. They have an 23 node machine, each node composed of an RS6000 with 128 Mbytes DRAM. This machine has 4 Gbyte of hard disk and operates at 250 Mflops/s performance (double precision, peak). It runs a version of Unix optimised for a parallel machine and it has DB2 IBM's new parallel relational database. We hope to use this machine in co-operation with some Southampton researchers to explore the dynamics embedded in citation networks.

Finally, we are adding two new tools to our ToolKit: a relational database and an interactive WWW database server. To do this we purchased a 64 bit DEC AlphaStation with 256 MB RAM and 2 x 5 GB fast wide SCSI hard drives. We are installing the Oracle relational database with Web server access. This machine will provide greater speed and better utilities for explore the BESST data and allow us to prototype a Web-based graphical user interfaces.

As you can see from our desktop we have access to an arsenal of computers and so far our ToolKit functions on most of the machines we use. The boundaries of Desktop Scientometrics have expanded.

Scientometric Desktop

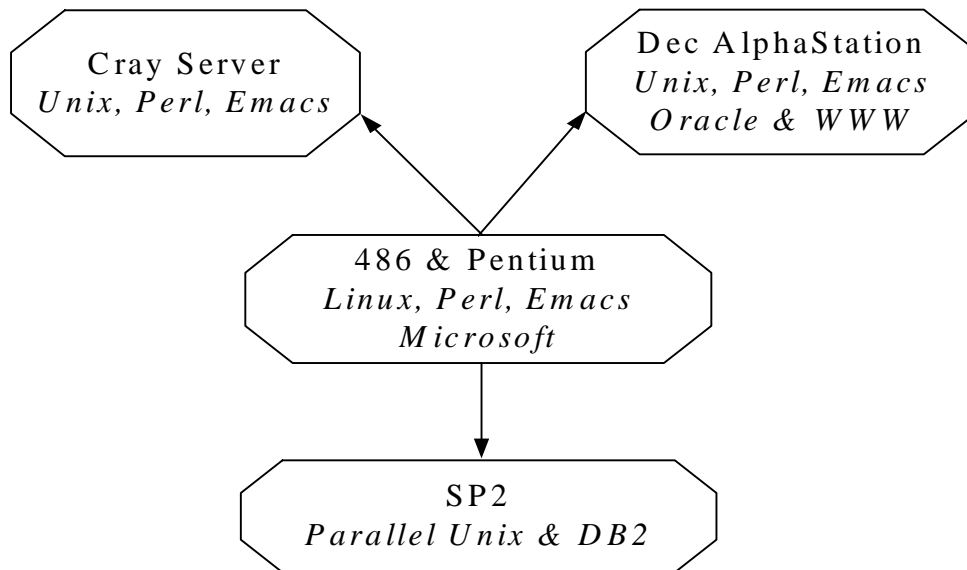


Figure 2

7. The Future of Desktop Scientometrics

“And what”, you may ask “are our aspirations?”

We dream of two things. First, we dream of building an interactive graphical interface to assist social scientist and policy analysts from all disciplines to probe the dynamic signature of national systems of innovation as reflected in publication records. This dream is rich with interesting research problems. For example, how does one visualise research co-operation dynamics between 6000 institutions? We are not sure but we have some ideas for techniques for visualising the collaboration dynamics between six UK institutional sectors. Exploring collaboration networks is not as difficult a problem as examining citation networks but even this may challenge the frontiers of data visualisation and data mining techniques. We realise that this is an ambitious dream but we believe it is achievable. To accomplish this dream we will need help from people skilled in the computing and mathematical methods.

Our second dream is grander. We dream of expanding the BESST database to explore as many national systems of innovation system as possible starting with the European Union. A project of this nature would require pan European co-operation involving the whole of the European bibliometric community. Can you image the possibilities? We could explore the various European national systems of innovation in a comparative manner. We could

examine the dynamic patterns of interaction between these systems as the European Community evolved through the 1980s and 1990s. However, there are many complex issues that would surround a project of this kind such as Intellectual Property Rights, quality control and data security. We believe that with a good measure of cooperative effort, a reasonable level of funding and a lot of hard work from the talented resources in the European bibliometric community a EuroBESST could be constructed within three to five years. Ah, but that is just a dream and Desktop Scientometrics is a virtual reality!

Acknowledgments

The authors are grateful to the Economic and Social Research Council, Cabinet Office - Office of Science and Technology, Department of Trade and Industry, Medical Research Council, Science and Engineering and Physical Sciences Research Council and Department of Health for support. Also, the authors would like to acknowledge Nigel Ling for his valuable contribution in preparing the 1981-1991 thesauruses and Nick Winters and Jane Calvert for their assistance in preparing the 1992-94 data.

Notes and References

- 1 J S Katz, D M Hicks, M Sharp & B R Martin, *The Changing Shape of British Science*, STEEP report, SPRU, 1995.
- 2 A demonstration of the BESST database and graphical interface was presented at the conference. A copy of the database and interface can be obtained from the authors.
- 3 For example, the unique article numbers that are supplied on *ISI* tape version of *SCI* are not available on the CD-ROM version. Thus, there is no easy way to link a subset of papers derived from CD-ROM to the original CD-ROM articles.
- 4 Linux runs on an IBM PC compatible with an ISA or EISA bus and a 386 or higher processor, Apple Macintosh, Amiga, MIPS, Sun workstations and DEC AlphaStations. The Linux kernel was written by Linus Torvalds from Finland and by other volunteers. Most of the programs running under Linux are generic UNIX freeware, many of them from the GNU project.

Linux is available from the following anonymous FTP sites:

<u>Site</u>	<u>numeric address</u>	<u>Linux directory</u>
tsx-11.mit.edu	18.172.1.2	/pub/linux
sunsite.unc.edu	152.2.22.81	/pub/Linux
nic.funet.fi	128.214.6.100	/pub/OS/Linux

Linux is also distributed on physical media, including floppies, CD-ROM and tape, by several commercial vendors. Please read the distribution HOWTO available for FTP at sunsite.unc.edu as /pub/Linux/docs/HOWTO/distribution-HOWTO. The typical cost is \$50-100 (US) and includes various compilers, editors, X windows, TCPIP and numerous other utilities.

- 5 Project GNU is organized as part of the Free Software Foundation, Inc. The Free Software Foundation has the following goals: 1) to create GNU as a full development/operating system. 2) to distribute GNU and other useful software with source code and permission to copy and redistribute.
- 6 PERL was written by Larry Wall at the NASA Jet Propulsion Laboratories. It usually bundled in with most Linux distributions and is available for anonymous FTP from

<u>Site</u>	<u>numeric address</u>
ftp.uu.net	137.39.1.9
archive.cis.ohio-state.edu	128.39.1.9
jpl-devvax.jpl.nasa.gov	128.149.1.143

There are two excellent reference books: Wall L and Schwartz RL, *Programming Perl*, O'Reilly & Associates, Inc (Sebastopol, USA), 1991, and Schawartz RL and Wall L, *Learning Perl*, O'Reilly & Associates, Inc (Sebastopol, USA), 1993.

- 7 The corporate address list was generated using the following criteria:
 - a. only corporate addresses from article, note and review publication types containing the keywords *England, Scotland, Wales* or *North Ireland* in the country field were selected.
 - b. the county field was not used
 - c. the city name was extracted from the city-postcode field where possible.
 - d. in general only the first, or the first and second sub-field from the first field was used. However, if certain keywords such as *MRC, AFRC, NERC*, or *SERC* (or a variation of a keyword) appeared *anywhere* within the first field, all four sub-fields were used when possible.

- 8 In 1975, Richard Stallman developed the first Emacs, an extensible, customizable real-time display editor & computing environment. GNU Emacs is his second implementation. It offers true Lisp--smoothly integrated into the editor--for writing extensions & provides an interface to the X -Window System. It runs on Unix, MS-DOS, & Windows NT. In addition to its powerful native command set, Emacs has extensions which emulate the editors vi & EDT (Digital's VMS editor). Emacs has many other features which make it a full computing support environment. Source for the `GNU Emacs Manual' & a reference card comes with the software. Sources for the `GNU Emacs Lisp Reference Manual' & `Programming in Emacs Lisp: An Introduction' are distributed in separate packages. Emacs can be ftped from many international sites, including unix.hensa.ac.uk and is included in many Linux distributions.

- 9 This is a powerful multiprocessor machine running 12 SuperSPARC processors each with 2 megabytes cache. It has 768 megabyte of memory and 139 gigabytes of fast access disc space.