(connect2)

PAPER SLIPS TO COMPUTERS

Notes on setting up the 'topics' database

Alan Macfarlane

The Slip Index System

Sarah Harrison and I have developed a number of slip indexing systems over the years since 1971. One of them was a large index of slips on numerous topics arranged into a hierarchical index system. 'Classification is based on a preconceived plan: the whole field of interest is *a priori* divided into a number of classes. Nearly all classification systems are hierarchical, i.e. the whole field is divided into a few main classes, each of these is subdivided into subclasses, and this process of splitting-up can be repeated down to the required detail'.[1] In our case, the hierarchical divisions were not worked out all at once, but gradually evolved, the main headings in the first five years, and then lower level ones over the next fifteen years.

The general structure

In this 'topics' database, as I now call it, the top level of the hierarchy had about twenty major headings such as:

Agriculture, Economy, Marriage, Politics, Capitalism

Under each heading would be the next level. For instance, under Capitalism would be 'Causes of Capitalism', 'Consequences of Capitalism', 'Feudalism and Capitalism' and so on.

At the third level of the index, would be further sub-divisions, for instance under 'Feudalism and Capitalism' would be headings such as : 'Was England feudal?', 'Is European feudalism unique?', 'When did feudalism end?.

Roughly there were some two thousand slips[2] per unit of the top level, each of these divided into ten or more sub-headings, and each of these divided into ten or more with an average of twenty slips. In practice, there is enormous variance in this four-level hierarchy.[3]

The content of the slip

The information was written on small slips. There would be a heading, a short quotation or other material in the centre, and the source from which the information came at the bottom.

For example, a general topic or heading might be 'Feudalism and Contract'

---

[1] V.Stibic, *Tools of the Mind* (Amsterdam, 1983), p.97

[2] For reasons of economy of both space and expense, I had reams of foolscap paper cut into blocks of rectangular slips. One ream of 500 sheets produced six thousand slips.

[3] This slip index method was not just used for the 'Topics' database. We later adapted it for use in the study of particular communities, historical and anthropological. The top level then became Person, Place, Date, Source, Subject.

Then the abstracted text or quotation or idea was written in the middle of the slip, for instance:

"The master who taught us that 'the movement of the progressive societies has hitherto been a movement from Status to Contract' was quick to add that feudal society was governed by the law of contract. There is no paradox here..." (referring to Henry Maine)

There would then be the reference to the source of the idea or quotation:

author: Maitland
title: English Law, vol.1
page: 233

The small size of the slips (two and a half inches by three, or about sixty by eighty cm.) forced me to be brief. The quotation above, with about forty words, is about the average; almost always the text would be between ten and fifty words long. The maximum would be about 100 words.

A short history of the creation of a paper slip index.

The topics database is like an archaeological dig. There are sedimentations of stages of my life recorded in it. It may be easier to understand the final database if these are explained briefly.

I started the system as an undergraduate (1960-3), and accumulated a thousand or so slips of quotations, statistics and other 'facts' related to the history papers I was taking at Oxford. So there were quotations from philosophers, or useful quotations from eminent historians.

I then added several thousand more slips when I undertook research for my Oxford D.Phil. (1963-7) on the history of witchcraft in England. This explains why there are quite a large number of records relating to early texts on witchcraft. There are also several thousand from sixteenth and seventeenth century ecclesiastical and other court records, in particular from the early transcripts made by various antiquarians.

When I did a two year's masters degree in anthropology (1967-8)), I added to the historical and anthropological quotations, particularly in the fields of sexual behaviour, on which I was doing a dissertation. I began to index a range of anthropology books as I learnt the discipline and I also began to add thoughts which I picked up from seminars and conversations.

I started to write a book (1967-8) based on the diary of the seventeenth century clergyman Ralph Josselin, and this accounts for several thousand records from early English diaries, autobiographies and letter books. I also began to abstract other historical records.

I then went to Nepal to undertake a doctoral thesis on population and resources, based on anthropological fieldwork (1968-70). Though the two or three thousand slips I made as a way of storing field observations are stored elsewhere, some reflections of the Nepalese phase, and the subsequent book, can be seen in the next layer of the data set.

I then returned and did research on the history of sex, marriage and the family in England as a Senior Research Fellow at King's College, Cambridge (1971-4). During this time I collected some thousands of quotations and 'facts', particularly in the fields of English social historical sources. I was also starting to work on the local history of various English villages, and this led to materials on methodology, and comparative studies from other villages.

As I started to teach anthropology at Cambridge University from 1975 onwards I added in further materials from anthropological monographs, seminars and thoughts. I also added further historical materials while writing my books on *Individualism* (1977), and further materials which would be used in my books on *The Justice and the Mare's Ale* (1978-1980), and *Marriage and Love in England* (1981-1985).

From memory, there must have been about 20,000 slips by 1978, and about 30,000 by about 1985. By that date I was becoming increasingly aware that the system was no longer working after some twenty-five years of accumulation.

I was extremely fortunate from 1990 to have the active backing and interest of my friend Gerry Martin, who helped to fund the work of creating a new kind of information retrieval system which would overcome the problems, and which will be described below. He also funded Penny Lang to type in the existing forty thousand slips, and to help to put in another twenty thousand over the next ten years. My handwriting made her task difficult, and some of the strangeness in the current 'topics' database are to be explained by this process of having data typed by someone other than the author of the slips.

It did mean that a new burst of adding materials was possible. These were mainly related to two new projects I had become engaged in. I visited Japan for the first time in 1990 and re-visited it every few years through the 1990's. I read and abstracted a great deal about its history and culture and input some of this into the topics database. Towards the end this overlapped with a series of four books speculating on the demographic, social, political and scientific origins of the modern world.[4] I marked up passages of books and Penny Lang typed in another fifteen thousand or so records. I also, finally, went through a number of encyclopaedias. So the topics database, as at the start of 2005 contains roughly sixty thousand records.

I am leaping ahead, however, for I was becoming increasingly aware from the early 1980's that the slip index system was no longer working. I did not realize why this was. Since many researchers use a similar method at the start of their research, and will face similar problems, it is worth explaining why it failed and how I overcame the problem.

The reasons why paper slip indexing systems collapse.[5]

---

[4] *The Savage Wars of Peace* (1997), *The Riddle of the Modern World* (2001), *The Making of the Modern World* (2002), *The Glass Bathyscaphe* (2002).

[5] Three examples of the collapse of the system may be cited. The anthropologist R.R.Marrett used a similar system, but the best article he ever wrote was when he was accidentally parted from the increasingly unwieldy monster (Marett, *Jerseyman at Oxford*, pp. 117,156). I have temporarily in my possession some of the index books with perforated slips which Sir J.G.Frazer filled in order to write *The Golden Bough*. Many have not been detached for filing. The great historian Lord Acton assembled

As long as I kept to about 20,000 slips and did not want to change the classification, the hand-indexing within a hierarchical system was manageable. After about 1980, however, I began to be aware that, as with everything else, the law of diminishing marginal returns applied.

One of the reasons for this concerns the basic fact of size. There is a general law of indexing that the larger the body of other slips into which a slip has to be put, the slower it becomes to file away a specific slip. Thus, if I have one drawer with 2000 slips, I can probably put a slip straight into it in a few seconds. By 1980, after about twenty years of indexing, I had about 17 drawers with about 35,000 slips.

It had thus become necessary when filing the new index slips to sort them first into major headings, then sub-headings, and only then to approach a drawer to look at the specific headings. It might take up to a minute to put away each slip - and it was tedious and complex work.

The difficulty did not just arise from the fact that there were more and more places where it could be filed because there were more slips. It also arose from the nature of hierarchical classification systems.

As I put away index slips, I was trying to remember the layer upon layer of previous classifications I had established. Did I have a category for attitudes towards menstruation, or should I make a new one? Should I put it under 'Gender', 'Life Cycle', 'Sex', 'Body' or what? The difficulties can be imagined.

When I refer to the layers of classification, this brings up a related difficulty, namely that over a period of twenty years my understandings and hence classifications shifted. As my interests changed, old distinctions and divisions meant less, and new one's were created. One effect of this, was that to file something under a system whose principles were laid down 20 years ago, I had to try to think back into the classifications of that time - another cause of friction and difficulty.

This also affected the finding of information as time passed. A slip with the same words written on it might have been put in all sorts of places depending on the time in which it was filed away. In order to find it again, or to know under which subjects to look for information on a particular topic, I had to remember when I might have indexed something, the classification system at that time, and then make the search.

The overall classification itself became less meaningful, yet, being hierarchical, while it was possible to tinker with it, it was impossible to change it fundamentally. As Stibic points out, "The principal disadvantage of a hierarchical classification system is its rigidity. It is difficult to adapt the existing divisions, and it is practically impossible to change its basic structure". [6]

This problem was related to another. By definition, all 'slips' contain a *relation* between at least two things, and probably many more. Hence each slip needs to go under *at least two* headings. In fact, it should probably go into three or four places in the filing system.

---

boxes of cards, which are now in the Cambridge University Library. As he read more and more and abstracted many thousands of references, he wrote less and less. Much of his best writing was done earlier in his career (see Butterfield, *Man on his Past*, 63; David Mathew, *Lord Acton and His Times*, p.104). I have also been told that the German philosopher Walter Benjamin created a very large collection of cards on various subjects. Whether this was related to his failure to publish his long-planned great work on nineteenth century Paris I do not yet know.

[6] Stibic, *Tools*, p.98

The quotation from Maitland could have been indexed under any of the following:

Maitland's ideas
Maine's ideas
feudalism - nature of
contract - relation to feudalism
status - destruction of by feudalism
progress - evolutionary views of
modernity - what caused it

and no doubt many others.

It has been arbitrarily assigned to one of these (feudalism) and tends to be lost to all the others.

This necessity to choose only one, or occasionally two, places where material is filed, adds to the problems of finding the information again when I came to work on a topic. In other words, the law of diminishing marginal returns also applied to information retrieval. This is partly the problem of more slips to search through. There is a tension because, in theory, the larger the data base (i.e. the more slips), the more exciting and interesting the searches should prove. This was often true and perhaps reached a peak when the number of slips had reached about 30,000.

Yet it also began to be obvious that it often takes a very long time to find relevant information. As the hay-stack got larger, it naturally became more difficult to find the few relevant needles. What was relatively simple with one thousand became progressively more difficult.

Yet all of this ran counter to another aim, namely the desire to encourage overlap, change classifications, relate things which were normally kept apart by the ways in which we divide the world up. Without pursuing this goal, how was I going to break out of the mould of previous classifications?

The result of all these un-analysed pressures was that after assembling about 40,000 or so slips, I began to lose heart. It was usually quicker to remember an author and go to the book - to rely on human memory. I might then use the more usual method of indexing a book at the back. As long as I could remember the author, I could find the material. If I did not have the book, I abstracted it and did the same.

Some other weaknesses of paper slips

One danger inherent in paper indexes is the amount of effort they take to add to and maintain. That means that more and more of the worker's energies go into the creation of the tools for research, and the less time there is to actually do the research and the writing.[7]

A second is the dangers from fire, and to a lesser extent burglary or moving house, to a large set of paper files. There is only one copy available, and it is very easy for it to be destroyed, whereas with a computer system it is easy to keep copies in a number of places.

A third difficulty is that as the paper indexes increase in size, they take up more and more space around the researcher. It becomes difficult to keep them within arm's reach, which is essential for quick research, and they crowd out other reference materials. Even when made on very small and very thin paper, as in my index, they take a lot of space.

[7] I was reminded of both of these first points by Dr. Brian Harrison, who kindly commented on the whole of this description.

Finally, much research and writing is now done on the move. A term at another university, a trip to work at a holiday home in the south of Europe, a week-end away in the mountains, or even working between home and a room in a University, means that the researcher is away from the paper indexes at the very time that the materials are likely to be most useful. The large boxes are too precious to carry around. Again, only through drastic compression into a portable system can they be used really efficiently.

A way round the problem: from paper slips to computer databases

During the last thirty years, the data gathering methods in the arts and social sciences have developed rapidly. In particular the development of photography, film, tape-recording and other gathering devices, as well as, recently, the use of new input devices (scanning), laptop computers and the internet have increased the amount of material that can be gathered and stored. Yet the equally important activity of analysing, absorbing and re-classifying the materials, which should absorb two thirds of our effort and time, has received less attention.

Let us break this down into the various stages. The first is the filing away of what we had previously envisaged as written record slips, and now become separate computer 'records'. With the development of modern computing this problem is largely solved.

This is despite the fact that the mathematical rule which explains why it becomes impossible for a hand slip index to expand much beyond about fifty thousand slips still operates. This rule, when applied to computers, states that the larger the set of records existing in a database, the longer it takes the computer to index/store a new item. The mathematics of this rule is as follows:[8] "a database of any kind will gradually slow down as more data is added", because "to place a record in an index, an efficient program can do no better on average than a time proportional to the natural logarithm of the number of items already in the index." This means, for instance, that if there is only one index item in the database, it will take 0.001 seconds to insert one item. If there was no slowing down effect, one thousand items should be inserted in 1 second. In fact, because of the above law, it takes 7 seconds.

The reason why we feel that with the advent of the computer the problem has been solved, even though the number of records increases, is because of the increasing speed of computers. As Moore's law has predicted, the power of computers has been doubling every eighteen months for more than thirty years. Their central processing power is growing so fast that it easily keeps pace with, indeed exceeds, the needs of most social scientists.

A second reason for relief from the problem is that the strain is transferred to the machine. Though it may take a number of minutes in 'real time' for the computer to add a batch of records to a database, the human being only takes a few seconds to set up the process and can then go and do more interesting things. Thus the problem of adding records is, at least temporarily, solved.

The same relaxation of the constraints is true of the other operation - finding the data again. While it is true that the larger the data set, the longer, in principle, it takes to find any particular item, this difficulty is overcome by two features of computers.

Firstly, their actual processing power is increasing so fast, that, as with adding material, it is far exceeding the social scientist's or historian's requirements. Secondly, there is the growing sophistication of database management systems and information retrieval software. This again is an important but largely overlooked feature of information storage and retrieval.

---

[8] *PC Magazine*, July 1992, p.275

From hierarchical and relational systems to probabilistic retrieval.

When I started to use computers seriously in the early 1970's the data was still held in a hierarchical structure similar to the hand index described above. This had many of the unsatisfactory features which I have described. The new development was in moving towards flat databases based on Boolean algebra, often known as relational databases. So with Charles Jardine and Tim King we set up the first large relational database system in Cambridge, which was working by the late 1970's.

This was a great advance, but it still did not quite solve my problem. The limitations of relational databases are easy to demonstrate. If I use conventional relational searching methods and put in a query of the and/or/not type which they accept, it is like firing a bullet through the data.

For instance, in my topics database, if I ask for 'status **or** contract **or** feudal' I get over one thousand answers. The particularly interesting ones, like the one cited above, would probably, on average, come about half way through. I would have to look through several hundred answers before finding the particular quote I was looking for.

On the other hand, if I tried to compel a more interesting answer by asking 'status **and** contract **and** feudal', in other words records with all three terms in them, the answer would have been found more quickly, but all sorts of other interesting materials would not emerge. Each time I narrow the query down with a boolean query, I lose interesting material, but to leave it 'open' makes searching very tedious.

Relational databases, which keep information in fields and columns and then look at intersections between these, are very useful for commercial and other applications. They may be excellent in order to find out how many people in Glasgow drink a certain whisky, or how many tins of baked beans are left in stock in Sainsbury's, they may be excellent. And for certain types of statistical work in sociology, they are very helpful. But they do not find unexpected things. They require the user to know what is there, and then helps him or her to find it.

So in the middle of the 1980's, based on the work of Keith van Rijsbergen and Martin Porter, we helped to develop a new database system based on probabilistic mathematics. The early system was called *Muscat* (Museum Cataloguing System), and we have recently helped the system to be re-written the system from scratch for use on the web in a system named 'Bamboo' developed by Richard Boulton.[9]

This system allows me to search for any shape or pattern of words I like in order to find what I am looking for. It ranks the 'finds' in order of probable usefulness, using a powerful algorithm to determine which words within a particular database have a higher retrieval value. It 'suffix strips' or stems words so that 'marriage', 'marriages', 'married' will all be found. And it can be combined with relational or boolean searching. It is much more like firing a shot-gun. A lot of pellets scatter out and the best answers, then the next best, are found.

In the Maitland quotation used above, I could have found the material by relational methods. I could have put in the query by name, title, or major heading 'feudalism and contract'. But more interestingly, I could be working, say, on the problem of the peculiarity of western capitalism and put in a query 'to what extent did feudalism represent a movement from status to contract'.

[9] Richard Boulton is a member of Lemur Consulting of Cambridge.

If I search on the words 'status contract feudal' in my current database of 60,000 records, the system finds the above very quickly. What is more, I can browse through a number of other interesting records on the same theme, for instance a discussion by Granet of the same phenomenon happening in China.

The real delight is that since I am now pursuing problems and associations, I ignore the artificial divisions created in the hand indexing system and find material of an unexpected kind.

I had always sensed that half-forgotten treasures were in my material, gathered over a long period. But like a squirrel's nuts buried in autumn, their location was half forgotten. I knew, for instance, that if I was trying to consider the use of animal analogies in politics that there might be material under all sorts of fields - under general headings of 'animals', 'metaphors and similes', 'politics' etc. But also in other places - say under a section headed 'nature and culture' or 'Hobbes' or 'Greek thought'. Only a small part of what was relevant would be found under the specific heading under which, twenty years ago, I would have decided to put it.

Now, thanks to a brainless, senseless, statistical, machine, and a new kind of highly flexible search system, all sorts of unexpected connections can be made to refresh the mind and stimulate the imagination.

Stibic wrote that 'If there were a uniform classification system, and all the available information were organized by it, then the whole "knowledge of mankind" could be stored in a structured form, any one piece of information would occur only one and in one place, interrelated with other elements of knowledge via a multi-access and cross-connected system'.[10] In some ways he was writing a description of what we now have.

It is not based on a uniform classification system, which is impossible to achieve, but on a much more powerful and flexible system. The computer indexes every item in every possible way, so that it can easily be found, though only, as Stibic writes, only held once. Putting it in another way, as the material is put into the system, the database anticipates every possible search users may make. And the web now makes this a 'multi-access' possibility. It is something which we have glimpsed with Google, but can be tailored to particular research uses.

[10] Stibic, *Tools*, p.262