Knowledge Classifications, Bibliographic Classifications and the Internet

Abstract: The advent of the Internet has intensified problems of classification and nomenclature in a range of disciplines that have been the concern of librarians and scientists for over a century. Biology and Medicine are selected as examples, and the work of specialists both in scientific systematics and taxonomy and in bibliographic classification is examined in order to suggest some solutions to current difficulties in information retrieval.

1. Introduction

Much effort is being expended at the present time on the creation of «search-engines» so as to bring a measure of order to the chaos that we know as the Internet. In the process, there is considerable reinvention of the wheel. It is disappointing that librarians who, after all, have been concerned with knowledge organization since at least the time of Callimachus, have not been able to make their skills and achievements more attractive to those who are involved in seeking solutions to the problems of knowledge organization. I would like to suggest that attempts to date have not been very successful because of several factors, the chief being:

The very natural wish for users to search on words, which results in an over-heavy reliance upon keywords rather than a structure as the basis for conducting a search;

There are too many different people working in isolation on various methods of arrangement and knowledge organization, and biologists, doctors, information scientists, computer scientists and others, all engaged in seeking similar solutions, work in total ignorance of what the rest are doing - there is very little cross-fertilization in professional journals and other sources where such research is reported;

Librarians and other information workers do not make sufficient noise about what they have done and can do, and the organization of knowledge is no longer considered to be one of the fundamental skills of our profession.

2. Definitions of classification and standard sequences

What do we understand by the term classification? It does not mean the same thing in all contexts. There are two sorts of classification - the scientific search for an order, frequently thought of as the activity of biologists, but actually practised by many scientists and social scientists in their search for creating a «perfect» order. This type of classification, by its very nature, must remain constantly fluid and subject to change, especially at the edges of what seem clearly distinct classes. The second, more pragmatic, type of classification seeks to create an order for the specific purpose of easing scholarly communication, enabling the interaction of ideas, and simply arranging objects, which may be books, specimens in a museum, items in a bibliography or any other collection of miscellaneous material.

It is this second kind of classification, or what a biologist would call a «standard sequence», to distinguish it from the perfect, unattainable ideal, that librarians and information workers have ever since the end of the last century when libraries adopted the practice of...
open-access, been concerned with and made their special expertise. Today, we are, in a sense, back at the end of the last century, but on a much grander and more intangible scale. We are faced with a mass of miscellaneous information, much of whose origin is unclear. There is, more often than not, no indication of for whom it is intended, and we have no precise idea of the quantity or quality of what is available. So, it is comparable to a closed-access collection, but without the assistance of an author/title catalogue.

3. Requirements of a retrieval tool

In order to access the information that is available both electronically and in all the other formats with which we are familiar, and to select sensibly from it that which is relevant, in a manner that gives some measure of confidence that the selection is the right one, there are two major elements that need to be taken into account. The first is a systematic arrangement that enables concepts to be displayed in such a way that their relationships to one another are clear, and provides a pattern that acts as a map to guide the uninitiated round the subject. The second is the terminology that is used to label the concepts so displayed. Both of these are activities on which many people are expending a great deal of energy and money, in trying to find solutions. I would like to identify some of these problems in relation to the natural sciences and medicine, in particular, and try to suggest some ways in which better use might be made of all these separate endeavours.

First of all, what is meant when we speak of classifications, systematic displays, linear sequences (and this list alone demonstrates the problem of terminology in communication - they all are part of the same process)? The first two may not be in linear form, but are frequently so ordered for pragmatic purposes, often through the use of a notation which imposes a seeming linearity on what is often a tree, or even a bush structure. Classification takes place in all levels of society so as to co-ordinate information in a way that makes it easily retrievable, for a variety of purposes, many of them nothing whatsoever to do with the bibliographical. First of all, there are the various classifications produced at international level, such as the International Statistical Classification of Diseases and related health problems of the World Health Organization, or the classification of occupations used by the International Labour Office for its statistical publications, or the Merck index of drugs. These are respectively, used by doctors for recording the incidence of diseases, and reporting as required under national or international law, or by those responsible for collecting census data and similar information, or are widely used by chemists and pharmacists as a base of reference. How many library classifications take note of these, much less actually incorporate them into their tables as a means of expansion? Jack Mills’s revision of the Bliss Bibliographic Classification pays some attention to the International Statistical Classification of Diseases and related health problems, but only as a source for vocabulary, together with a number of other sources. The Universal Decimal Classification suggests the use of the ILO list for expansion in its Sociology tables, but does not give any assistance as to how the user might find it, or actually copy relevant sections into the scheme. These classifications are mirrored at national level by such listings as the so-called «Reed codes» used by the National Health Service in the United Kingdom for the classification of diseases, or the British Central Statistical Office’s Classification of occupations, like the ILO list, for the purposes of the census and other demographic statistical compilations.

4. Appropriate timing of major changes to standard sequences

These official or «quasi-official» classifications are the fruit of much research by
scientists or social scientists who have eventually arrived at a compromise solution or what a biologist might call a standard sequence. Behind them, as with the biological classifications for plants or animals, lie many years of patient research and much argument, and this does not cease simply because the compromise classification has been completed. Knowledge did not stand still in 1876, but ever since that date it has been poured into Dewey's framework for millions of library users, and in the sciences exactly the same procedures are taking place. In addition to all this explicit classification activity, a great many people concerned with computerized systems are at present busying themselves in devising systems and coming up with solutions which they think are innovative, but which we have actually known for years. How do we find out what others are doing, and how do we impart our knowledge to others seeking solutions to similar problems? To what extent is it desirable that a standard arrangement should be followed for the organization of the same concepts in different contexts? For how long can we, if we adopt them, regard these standard sequences as standard? The biological sciences provide us with useful object lessons; for instance, an article on provisional classifications versus standard avian sequences written by two noted systematic biologists (Mayr & Bock, 1994) begins thus:

‘When reviewing the manuscript of an excellent new ornithological text, one of us was startled to read the comment «... the Yellow-throated Longclaw (Macrollonyx croceus), a member of the pipit family (Passeridae)...» Every ornithologist knows that the pipit family is the Motacillidae and that the Passeridae (Old World sparrows), if recognized as a separate family, is composed of sparrows of the genera Passer, Petronia, Carpospiza and Montifringilla. After recovering from our shock, we realized that the authors of this text were following the newly proposed classification by Sibley and Ahlquist».

This classification published in 1990, turns the well-established Wetmore-Peters classification (Wetmore, 1930), which has been the recognized standard sequence since the 1930s, on its head. The authors of the article in question express concern that beginner students in avian biology, for whom the textbook in question was written, would be baffled by such placings as this, when all the standard field guides, checklists, etc. would place the bird in question with the Motacillidae. Such puzzling occurrences led David Lack (Anonymous, 1967) to propose an internationally agreed world list of birds, and subsequent discussions revealed what I think is also a problem for us. The biologists had great difficulty in distinguishing a standard sequence which acts as a common means of communication between scholars, from the work of specialists in systematics, who aim to improve the understanding of evolutionary relationships between genera and species and which will be subject to constant fluctuation. The most difficult decision both for scientists and for those of us concerned with information retrieval in general is when to alter a standard sequence to reflect such findings. To be useful, any classification/standard sequence has to have an element of stability, and an examination of some of our library classifications shows the desperate efforts we make to compromise and find a happy medium.

5. Problems of general versus specialist context

There are, of course, many differences between biologists working in one, rather narrow, field (though they might claim otherwise) and the information worker who has to tackle the whole of knowledge, and has to get to grips with the problem of the same concept recurring in different disciplines and contexts. Staying with biology, if we accept a "standard sequence" and use it as the basis for our structure of knowledge, to what lengths should we go? We can organize Palaeontology, Botany and Zoology according to parallel divisions, with an arrangement that our users will recognize. Both the Dewey Decimal Classification and the

Advances in Knowledge Organization, Vol.6(1998)
Universal Decimal Classification do exactly this, leaving spaces in the notations where necessary for genera and species which are either only extinct or only living, and so covering such eventualities as occurred when the coelocanth was identified as a living species. So that there is actually one sequence, in either case, notated initially with 56 or 58/59, depending upon whether the organism is extinct or living.

But, what do we do when we come to Medicine or to Veterinary Science - do we arrange diseases caused by bacteria or viruses in the same order as we have arranged those organisms in Bacteriology or Virology? Or do we list the diseases in an order more readily recognized by the medical profession, and link them to the symbol we have selected to signify the bug concerned? Do we arrange crops or animals in Agriculture in the same order as the scientist would like them? Farmers are normally more interested in an arrangement that reflects seasons, crop rotation, and other agricultural considerations, and their activities to produce foodstuffs, or in an arrangement of animals that conforms to farming practices - which means that horses, which perform something of the same sort of function as tractors on a farm, would be placed a long way away from cattle or poultry, which are reared for their milk or eggs respectively as well as for meat, though they are far apart from one another in a taxonomic arrangement. Animals and plants are not necessarily viewed in the same way in Biology and in Agriculture or Farming. Animals, for instance dogs, cattle or horses, have many breeds that man has created and continues to create, rather than the handful of species that occurs in nature, and the same is true of plants. A scientist will not talk about pests or weeds, but the farmer or horticulturist will. And in certain cases the same organism may be regarded as both a pest and something desirable - deer ruin trees, but may be reared for venison, clover is not welcome on our lawns but we feed animals on it. Some creatures attract a great deal of literature for reasons other than their own intrinsic essence. An example is the tsetse fly - largely of interest because of the spread of sleeping-sickness.

But, plants or animals that are not traditionally associated with farming may well be farmed in the future. In recent years - no doubt due to a European Union directive or grant - English fields have been full of bright yellow oil-seed rape, and blue linseed in May and June, crops which twenty years ago were never seen, and nowadays frozen pheasants may be bought in supermarkets, just like chicken. So, to "divide like" as Dewey put it, may have some attractions in providing for the unforeseen future requirements and is certainly one method of attempting to "keep pace with knowledge". Again, the same items will be needed in Cooking, so should we use an arrangement based on what is in origin a taxonomic classification, for the organization of recipes? Do we base our recipe books for ethnic cooking on a set of subdivisions originally devised for the organization of Anthropology? And if we look too deeply into this, we shall soon find that we need a hotch potch of different subdivisions, some derived from language and others from geography - the majority of people in the West Indies speak English, and their origins may be from Africa, India, China, Britain, or a mixture - where do we turn for a logical subdivision for Caribbean cooking? We need to collocate regional cooking, which will group Italian cooking - another geographical subdivision, Jewish cooking - a religious and ethnic subdivision, Chinese cooking - a mixture of the ethnic and geographical, and so on? How far should we go? This is a very difficult question to answer. To achieve a manageable and hospitable system and at the same time to satisfy the particular needs of individual groups is well nigh impossible.

6. Standard arrangements taking Biology as sample field of study

It may, however, be useful to be able to draw on a well-recognized arrangement to accommodate some of these needs. The biological sciences provide a useful starting place to

Advances in Knowledge Organization, Vol.6(1998)
look for precedents. The classes which comprehend Biology, Botany and Zoology in the major
general schemes of classification were originally based on sequences that were standard at the
time of their creation, at the end of the last century. Both DDC and UDC have, in the recent
past produced major revisions. Parallel with these changes in these library classifications, and
the DDC revision was 20 years in the making, there have been great upheavals in the thinking
of scientists, largely as the result of the use of DNA as a basis for establishing relationships,
rather than earlier evolutionary theories. The outline arrangement of mammals in five different
classifications will help to demonstrate some of the problems. These classifications are taken
from three recognized zoological works of reference, *Walkers Mammals of the world* (Nowak,
1991), Corbet’s *World list of mammalian species* (Corbet & Hill, 1991) and *Mammal species
of the world* (Homacki, 1993) and two bibliographic classifications, the DDC and the UDC.

The first, and most striking thing about the two library classifications compared to the
zoologists’ standard sequences is the position of Primates. In both DDC and UDC they are
placed at the end, possibly in order to create a harmonious transition from the end of class
5/500 and the start of 6/600, where Human Biology is placed, among the preclinical elements
of Medical Sciences. Alternatively, this may simply be a survival from the 19th century
arrangement on which both were originally based, which placed Primates last, with Man as
the superior being - an example of the maintenance of an arrangement in a one-time standard
sequence retained, despite modern thinking. Another oddity is the rabbits (Lagomorpha)
which DDC places very early on, while Corbet has them almost at the end, but all with the
exception of DDC collocate rodents with rabbits! Two of the three zoological listings separate
Sirenia and Cetacea, whereas both DDC and UDC, together with Corbet, collocate them.
MSW splits up the marsupials into five individual orders, and places them near the end of the
sequence rather than near the beginning, as the others do. Another difference between the
*Dewey Decimal Classification* and the rest is that DDC regularly interpolates terms which are
unscientific or provides notations for species on which there is popular literature. To go into a
more specific level, the arrangement for birds gives examples of this.

Here, UDC divides into 2 big divisions - Non-passerines (598.22/.27) and Passerines
(598.28/.29). Passerines are then split into Deutero-Oscines (598.281) and Oscines
(598.282/.299), then broad “popularly” recognized groupings: Old World insect eaters and
related families (598.283/.291); Crows, Birds of Paradise and related families (598.292/.294); New World insect eaters and related families (598.295/.296). Then every order and selected
families within each order are specified, with commonly found species spelt out in “including
notes”. DDC, on the other hand, mixes the scientific with the popular approach, dividing into
two major categories, 598.3-598.4 Water birds and 598.5-598.9 Land birds. This use of «lay»
terminology can create problems with headings like «Birds of prey» which include several
different families and therefore cut across the zoological classification. On the other hand, a
classification intended for general use must provide such headings, in accordance with literary
warrant. There is also the decision to make as to how far down one goes, and whether to
enumerate individual species when there is literary warrant for doing so, as DDC does, e.g. for
Peacocks or Ring-necked Pheasants, or as both DDC and UDC do for Salmon or Trout.

7. Necessity of pragmatic approach for retrieval tools

The need to cater for a range of approaches is something that schemes devised for a
library or bibliographical function, such as DDC or UDC, have to contend with. In this, they
are quite different from the more theoretical classifications devised for a specific purpose,
normally within a restricted disciplinary range. The same problem exists on the Internet,
where again the unpredictable has to be catered for by whatever form of search-engine is used.
Similarly, both bibliographic classifications and «search-engines» have to cater for the layman - something that is not normally the province of a scientific classification, though it may be required of a standard sequence. The Dewey Decimal Classification provides many instances of this in its terminology, and indeed its division into Land and Water birds is not a scientific one - though it does have literary warrant. The Internet presents the same problems, since it is an incoherent assemblage of material, derived from a variety of sources, and with no clear objective in mind.

One of the leading experts on biological classification, Ernst Mayr states that "the difference between the classification of organisms and of inanimate objects [by which he means books] is total.” He admits that there are certain common features, such as being based on the possession of certain common attributes and on principles of hierarchy. But he thinks that that is where the similarity ceases because, he states, that the library of his own institution (Harvard) uses four principles of classification, (1) year of publication; (2) subject matter; (3) size and (4) author's name. It is disappointing to note that one of the leading biological classificationists, as soon as he leaves his own discipline, confuses factors that relate to the management of an individual collection rather than its subject matter. "Classification” by size is simply a management practice indulged in by very large libraries, to save space, and year of publication is another such measure, while the author's name relates to a catalogue entry or possibly an author mark in a library which uses cutters for shelf arrangement. He claims, erroneously, that in scientific classifications, but not in library classifications, only one of the identified elements, subject, represents a true classification (Mayr, 1997). The same author (Mayr & Bock, 1994) when writing about ornithology argues strongly in favour of a standard sequence of families, in order to ease communication in all fields (not just systematics). He suggests that this would remove the need for endless upheavals in data storage and indexing systems and the layout of avifaunas, checklists and fieldguides. This would seem to be an acceptable point of view for both scientists and information specialists, provided that it is universally understood that the standard order is not intended to reflect the continuous changes of mind concerning taxonomic relationships.

Zoological linear sequences and the bibliographical classifications are both aiming at a pragmatic order which is intended to satisfy the perceived requirements of the majority of users. This is a quite different function from the detailed classification undertaken by taxonomists which is concerned with finer details that will never be satisfactorily perfected. This latter is a scholarly pursuit that is related to but by no means identical to the purposes either of organizing information for collection in museums or works of reference, or for the organization of information in any form of store. The two can exist happily side by side, provided that it is clearly recognized and accepted that a compromise is needed in order to achieve some sort of satisfactory order for reference purposes.

8. Possible ways forward

The classification of diseases is similar. As already noted, there exist standard classifications, not universally liked by the medical profession and different ones for international use from national, but these are not normally incorporated into bibliographic classifications. One way forward might be for the editors of schemes to be more explicit about the sources that they have used as the basis for their arrangements. Published versions of bibliographic classifications are not in the habit of including in the tables a note to this effect. It is not regarded as in any way a state secret, and the literature resulting from the DDC revision (New, 1996, New & Trotter, 1996) gives the basis clearly, as does the introductory matter in the revision for the UDC in Extensions and corrections. Precise information relating
to the sources used would provide guidance for institutions wishing to go into greater detail than the scheme provides, or simply to choose the correct place for a species not listed in the tables. This practice could usefully be extended to all parts of the classifications where well recognized standards exist, and proper citations would assist in locating the sources if institutions wished to do so. It would give the specialist user the confidence of having a recognizable sequence reflecting that to which he is accustomed in other contexts such as museums, field guides, floras and the like. It would also make for a great saving in the expensive business of revision to incorporate some of this existing work into our schemes. Serious consideration might also be given to making wider use of such classifications, such as taking the selected zoological classification for animals down to a certain clearly defined level, and then expanding as needed. In Agriculture, for example, specific breeds, can be indicated either alphabetically, or by some other acceptable arrangement.

9. Terminological requirements

But successful retrieval depends not only on the systematic arrangement, which helps to suggest ways of broadening or narrowing a search, it is also very heavily dependent upon matching terms and the use of recognized vocabulary. The examples cited earlier of DDC’s mixture of lay and scientific terminology demonstrate the difficulties. In most disciplines there is a problem of terminology because there is no universally understood language. The layman may well search for German Measles, rather than Rubella, and to a German they are Röteln anyway! Should one search under cancer, oncology, neoplasms? Even parts of the body are spelt differently in English/English and American/English - fetus/foetus for example, so that even if the decision to standardize on English is taken, there is no guarantee that everyone will search on the same term. At the very least, there is the possibility of searching, both in medicine and in biology, in Latin as well as in some sort of standardized English, and work is going on all the time on producing acceptable lists of terms in English for the international exchange of information, with due allowance for the problems of variant spellings, and for “lay” English and commonly used terms - Asian flu, mad-cow disease, AIDS, dabchick, peewit. In addition, there are examples of eponymous nomenclature - e.g. Parkinson’s disease, Sabine’s gull, Cetti’s warbler, Pere David’s deer, Creutzfeldt-Jakob Disease. Again, certain words are used indiscriminately to mean the same species, e.g. Parrot, Parakeet, Lorolet, Lorikeet and Rosella. There are lists that aim to correlate some of these terms, but the problem is vast.

A clear distinction can be drawn between the use of a standard English nomenclature and the use of familiar words in everyday conversation. In addition, not everyone wishes to access information in English, and English is not the only language used on the Internet. Variant terms will undoubtedly continue, but retrieval systems need to adopt some standards that are formalized, even if they are not always identical to common parlance. Problems have arisen, however, when those responsible for drawing up such lists have got carried away and propose changes to names that are perfectly adequate. The addition of the word “Common”, for example, to describe such birds as wigeon, cuckoos, hobbies or nuthatches does not cause difficulties, especially in a system that permits searching on words, since if the more familiar term is used, the information will still be retrieved. But more extreme examples, including dispensing with personal names, can obscure as much information as is revealed. A further problem in a computerized world is the unstandardized use of the hyphen which can result in the loss of a great deal of information.
10. Conclusion

Most of the sciences have bodies actively concerned with nomenclature at both national and international level. Many of these are the result of collaborative work generated through international conferences. More and more of these listings are now becoming readily available on the Internet - I have already referred to one for animals, and the Index Kewensis, the principal source for the names of Angiosperms is similarly available. There are Approved lists of bacterial names prepared under the auspices of the International Committee on Systematic Bacteriology. BIOSIS also maintains a machine-readable Register of Bacterial nomenclature. There is an International Code of Botanical Nomenclature adopted by the 14th Botanical Congress in 1987 and regularly updated, and similarly an International Code of Zoological Nomenclature in French and English, the most recent version of which came into force in 1983, and this is maintained by the British Museum (Natural History). The World Health Organization, in addition to being responsible for generating an international classification of diseases produces an International nomenclature of diseases whose principal objective is «to provide, for every morbid entity, a single recommended name» and the International Union of Biochemistry has a Nomenclature Committee. These are just a few examples of organizations and publications whose aim it is to clarify and maintain scientific nomenclatures. Many others exist for even more specialized areas such as ornithology, mammals or types of plant.

The quantity of standards and organizations all endeavouring to provide solutions in their specializations is overwhelming. What is important for those of us who are concerned with the creation and maintenance of tools that can retrieve information across the whole of knowledge, and at a wide variety of levels, both in the disorganized environment of the Internet, and for all the traditional applications is to capitalize on this endeavour. We must state clearly the sources for the arrangements that we adopt in our systematic displays and we must take advantage of the work done by specialist organizations concerned with nomenclature. We should consider incorporating their listings into our retrieval tools, so as to provide both a map of knowledge and a dictionary of terminology to assist the range of levels and linguistic approaches of those who seek information. We must make far greater efforts to co-ordinate what we are doing with the work of the specialists and reach some common agreement on the ways in which we handle both the systematic arrangement of concepts and control the vocabulary used in academic discourse. We also need to realize that any solution is going to be a pragmatic one that will not be universally accepted as correct, for no such thing as a universally ideal classification exists in reality, but that everyone, in the common interest of improving communication, must accept certain standards and make certain compromises.

References


