THE ENGINEERING OF SCIENTIFIC KNOWLEDGE

Derek deSolla Price
Yale University
New Haven, Connecticut 06520

ABSTRACT

With the rapid maturation of knowledge engineering it will be possible at last to do what the encyclopaedists attempted and failed - to make a perfect constantly updated and packed down file of universal scientific knowledge. There are a couple of snags. In the first place only less than half of scientific information is universal. The other half is local, domestic, relevant to only a small subsection of the world and must be handled at the parochial level. In the second place science is not a single linear updatable file. It is a very complex two dimensional array of such files, in constant rebranching and recombination. Will it be possible for very smart artificial intelligence to monitor and display this system in a fashion which takes librarianship from a passive managerial role to the highly creative.

L'INGENIERIE DE LA CONNAISSANCE SCIENTIFIQUE

RESUME

Avec la maturation rapide de l'ingénierie de la connaissance, il deviendra enfin possible de réaliser ce que les encyclopédistes avaient déjà tenté et échoué: un parfait fichier de la connaissance scientifique universelle, compact et constamment à jour. Il y a toutefois deux obstacles à surmonter. En premier lieu, moins de la moitié de l'information scientifique est universelle. L'autre moitié est purement locale, domestique, relative à une petite sous-section du monde et doit donc être traitée à un niveau "paroissial". En second lieu, la science n'est pas un unique fichier linéaire que l'on peut mettre à jour facilement. Elle est plutôt un alignement complexe et bi-dimensionnel de tels fichiers en constants raccords et recombinaisons. Sera-t-il possible, un jour, à une intelligence artificielle très futée, de contrôler et d'offrir un tel système de façon à donner à la bibliothéconomie une vocation hautement créative plutôt qu'un rôle de gestion passive?

ENGINEERING SCIENTIFIC KNOWLEDGE

A large part of the Scientific Revolution of the 17th century was wrought by the miraculous invention of the scientific paper. It happened as a direct result of the technological advance that made possible the first publication of newspapers and magazines, and in an almost explosive chemical combination with something in the structure of science the world's first two scientific journals were born almost simultaneously.

As is now well known, the exponential growth of the scientific literature so successfully outstripped all else in human development that within a century science had moved from books to journals and the burgeoning was so rapid that it created the 18th century mass movements to journals of abstracts and also the idea of the encyclopaedia. We can now see these movements as palliative. For a time it looked as if the knowledge generated in journals could be abstracted in briefer form after the fact, and that fairly rapidly it could be packed down into handy book form, even if the entire corpus of learning might well take a multi-volume set to encompass.

The problem is patently still with us, and in an aggravated form because we now realize that packing down is a complex problem with no ultimate solution in sight. Each decade the volume of knowledge doubles so that as much has been produced in the last ten years as in all previous time. There will never be quite enough expert reviewers to write the scholarly reviews of the state of the art on a sub-field basis, and even at a level more distant in time from the research front it is the best we can do to work at the level of textbooks for the graduate students emerging in the fields. The constant updating is perhaps not much more rapid than it always was because science has grown in its internal ecology the trick of keeping sub-fields almost constant in size by generating more of them. Specialization into more and more areas has left each area constant size. So the problem has shifted from simple updating to that of organizing the rapidly growing set of sub-field in their wild state of flux. They are always splitting by bifurcation and then recombining the strands as they weave together in new patterns.

Two new advances in computerized information enable us to solve a pair of problems that dominate the handling of scientific information. The updating problem is nowadays simply that of using the machine to manage a file that is constantly growing. It used to be taken as a matter of course that the literature of papers generated as extended letters between colleagues at the research front -- an on-line real-time operation of communication produced a product in journal form which when transferred from the periodical room to the stacks (via the bindery) would pack down into an archive which could be available for all time. Today the computer, though we have not yet quite taught it how to do better, can at least operate like a super wordprocessor and keep the file up to date once the new paragraphs of scientific contribution have been written.

The big intellectual contribution comes from our new ability to monitor the changing map of science as the threads separate and weave together again, so that at any time, using this feature of an extremely

orderly and well-mannered road map structure, we can spread out the files of knowledge and lay them all over the road map. All the knowledge that there is can therefore be in principle displayed as such a road map of super wordprocessed mini-encyclopaedias. We can leave it to the machine to aggregate and disaggregate such encyclopaedias as the state of the art changes.

Now even if this problem of knowledge engineering were solved there would remain another most interesting problem. We used to think that all scientific knowledge was of this special universal variety, valid and identical in all cultures and places, and simply structured because of the universality and objectivity of the science which gives it such an edge over non-science. Now we know that this is not entirely so. By analyzing the flow of citational transactions from journal to journal, nation to nation, and language to language, it turns out that only about half of all the "scientific" material published, even in high reputation journals, is of this sort. The other half of all publication is particular to a certain journal, or nation, or language. It is, so to speak, relevant to that particular group, parochial in its behavior and function, rather than universal.

This means now that even ideally we shall need one universal system, as previously described, and in addition to this we shall need numerous local systems particular to the needs and interests that fall within a single field, nation or language -- or worse still in all three categories simultaneously.

It looks to me as if the needs of universal science can be rather well met by a combination of super wordprocessing to create mini-encyclopedias and by the annotated road maps of science which can be kept on line and in real time. Local systems in which we do not have the economic and organizational advantages of scale will necessarily falter by comparison. The international science game will have the new feature of being liberated from the necessity of writing and publishing papers in such a style that they must serve both as research front communication and as packed down archive. The machine can do the packing down ultimately by processes more efficient than binding the journal, indexing it in every which way, and packing it into book stacks in all the world's great libraries. Presumably we can have instant satellite access to the updated world archive at any point on the globe. Another snag then will be the economic one of how we are to let the consumers somehow pay the cost of all this artificial intelligence and information management.