THE CAMBRIDGE EXPERIMENTAL VIDEODISC PROJECT

Alan Macfarlane

(*Anthropology Today*, vol. 6, no.1, February 1990).

Largely inspired by the films and photographs of Professor von Furer-Haimendorf, we decided in April 1985 to make an experimental videodisc about the Naga peoples of the Assam Burma border. (1)

The Nagas seemed a good choice for such an experiment. The precipitous terrain and forest, as well as the warlike head-hunting reputation of the peoples, deterred outsiders from entering the area until very late. The period of contact, starting effectively in the 1840s, was unusually gradual, lasting over a century until Indian Independence in 1947. The relative lateness of the contact meant that the second fifty years of documentation were within the era of easily portable cameras ;and the last fifty years within that of moving film. But how well was this process documented, and what remained'?

Good fortune brought to the Naga Hills a series of very gifted observers. These men and women became so involved with the Nagas that they assembled large collections of material in very difficult circumstances. The chief collections we were given access were those of R.G. Woodthorpe, J.H. Hutton. J.P. Mills, C.von Furer-Haimendorf. Ursula Graham Bower ;and W.G. Archer. Between them. They collected over 5,000 artifacts, took over 7,000 black and white photographs, made a number of sound recordings and made over six hours of moving film. They also kept extensive diaries and collected many pages of fieldnotes as well as publishing eight books and many articles on the Nagas. There was clearly no shortage of material. But how was this to be formed into a usable and distributable archive?

Making a videodisc

A videodisc or optical disc is a silver object which looks like a gramophone record. Information is engraved on its surface which is then coated with plastic. The information is read off each separate track by a laser beam, using a standard videodisc player. This produces a virtually indestructible storage format which is not damaged by dust, normal changes of temperature, electric current, damp, insects, etc.

A videodisc can hold a very large quantity of information. A standard disc can play moving film for 36 minutes per side in interactive mode or hold 54,000 separate pictures per side, or a combination of these. It can store at least 300 megabytes of information per side (the entire **Encyclopaedia Britannica** with pictures takes about 200 megabytes of store).

A videodisc can hold copies of almost all kinds of recordable information: photographs, slides, moving films, x-rays, sound recordings, graphics. The discs are double-sided and once a master has been created, copies can be made relatively cheaply.

This sounded an ideal medium for our pictorial materials. But how does one make a master? Here there was almost no guidance. These were early days and no videodisc of the kind we were attempting had been made in Europe. With the cooperation of the Audio Visual Aids Unit at

Cambridge and the Open University Production Unit at Milton Keynes, we therefore invented the methods as we went along. How this was done is documented elsewhere but can be summarised as follows.

We photographed about 1,200 Naga artifacts as colour slides and transferred these by tele-cine to the one inch master tape from which a videodisc is made. We re-photographed some 7,000 black-and-white negatives with a half-frame camera and 'Illumitrans' and transferred the strips of films onto one inch tape. We constructed and photographed some 200 maps. We extracted 150 moving sequences of film and 1,000 still frames from moving films. We re-recorded a number of sound recordings, from early wax cylinders to recent missionary songs. After three years' work we had a master disc and 100 copies.

Dealing with texts

For a variety of reasons, including the need to update and change textual data and the far greater cost of a machine that could (as with the BBC Domesday disc) read digital data, we decided to keep the textual materials separately. We decided therefore to keep these materials on a computer.

The main categories of material were as follows. There were the equivalent of about 500 pages of manuscript fieldnotes. There were manuscript field diaries, from the earliest in 1872, through the field diaries of Christoph von Furer-Haimendorf, up to the diary of Mildred Archer in 1847; approximately 1,000 printed pages or equivalent in all. Over 100 official tour diaries by J.H. Hutton, manuscript letters, and other typed and manuscript materials were also available. All these were typed into the computer.

The other main way of getting materials in was through optical character recognition, where the published book can be directly scanned into the computer. We did this with the eight monographs, which saves a great deal of labour, though it still leaves a good deal of cleaning up of the material to be done by hand.

This data input proceeded alongside the making of the disc. It will result finally in the production of a 25 megabyte database of materials which provide the context for the visual and sound materials. It is obvious that pictures and text reinforce and enrich each other.

Principles of selection

Elsewhere we have described at greater length the principles we used in selecting visual and textual materials. Very briefly, they were as follows. We reduced the six hours of moving film to 30 minutes by trying to include the material that was most intellectually and academically interesting, and, all else being equal, rejecting those sequences that were out of focus, badly composed, unsteady, from too great a distance, damaged, the colour fading, and so on.

There are likely to be over 15.000 Naga objects in European museums and private collections, of which we photographed a little over 1,200. We decided to confine ourselves to British collections. We then sought a representative selection, in terms of the types and functions of objects and their origins in different groups. We tried to use Naga criteria of significance rather than our own. We chose well documented pieces and those that fitted in with other materials on the disc.

Only a few hundred of the roughly 7,000 black-and-white photographs we discovered have been omitted. These were left out on the following grounds: they were duplicates of, or very similar to. other images; their quality was poor; they were outside the delimited geographical area; they were outside our time span; or they fell on the side of 'private experience' as opposed to 'public experience'. We did not censor photographs because their content was embarrassing or shocking in any way, or might do damage to the reputation of individuals, the British, anthropology as a

discipline, or for any other reasons. As far as textual materials are concerned, we limited ourselves mainly to materials written before 1947 and in English. If the material was likely to cause personal offence or political embarrassment to living persons. or was repetitive or trivial and of only personal interest, we omitted it. In all, this meant omitting at the most half a dozen paragraphs as compared to the twenty thousand we included.

Retrieval Systems

From very early on we were aware that the possibilities of the new media, a combination of computer and optical disc storage, would mean that we would have very large sets of data which it would be difficult to manage. The materials on the videodisc comprised about 10,000 'items' (maps, photographs, artifacts, films, etc.). The 25 megabytes equivalent of text represented about 20,000 paragraphs of writing. Supposing one wanted to search this, finding all the information in visual and textual materials relating to a specific person, place, date or subject, how could this be done? To search through 1,000 photographs or 1,000 pages of manuscript can be a long business. Recording materials of this diversity and scale, the disc would be unusable without an appropriate information retrieval system. None of the database management systems which had been developed for commercial or academic applications seemed appropriate for this project, so we developed our own. We worked in partnership with Dr Martin Porter to adapt his MUSCAT (Museum Cataloguing System) for these purposes. This system had been developed for use on mainframe and 'midi' computers and seemed ideal for our use. Among its advantages were the following.

It combines 'free text' with structured (boolean) searching techniques. The majority of current databases are based on 'boolean' retrieval (and/or/not). Though suitable for some purposes, boolean retrieval suffers from major limitations: the number of answers is usually too large or too small; users often require or expect to compose boolean expressions; the retrieved set of answers is not ranked in any way and so it is necessary to inspect the entire list in the search for relevance. The MUSCAT system incorporates boolean retrieval, but overcomes these weaknesses by adding 'probabilistic' retrieval, where answers are ranked in order of their probable usefulness. Terms are weighted according to formulae derived from probability theory.

In effect, this means that it is very easy to put in a natural language query such as 'show me all the photographs of women weaving on backstrap looms' or whatever. The 'best' answer will be given, then the next best, and so on, in decreasing order of probability of matching the query.

The added features of 'relevance feedback' and 'query expansion' turns this into a semi-intelligent system with considerable heuristic power. If an answer is 'relevant' (the kind of thing one was looking for) then it is 'marked'. All the marked records can then be examined by the machine. The computer provides a list of items in order of their probable statistically significant correlation with the marked answer. Any of these items can then be added to the query so that it becomes 'expanded', that is to say more powerful. This is a creative alternative to a synonym list. It is also a way of using computer and human together - combining the mathematical power of the machine with the intuitive knowledge of the human.

For our data, the MUSCAT system had a number of other advantages. It is very flexible, both in terms of size and structure. The system will deal with data sets of any size. It is possible to hold as many datasets as are needed. The number of records per dataset is unlimited. The number of fields per record and of characters per field is unlimited, within the one restriction that no single record must exceed 64,000 characters (about 20 printed pages).

There is no need for precoding and the data structure in a record can mirror the material one is dealing with. Hence we were able to adapt it easily to deal with the varied structures of records describing artifacts, photographs, maps, sketches, films, sound. manuscripts and printed texts.

Our aim was as follows. We wanted to adapt the very general 'Muscat' system so that it would be useful for historians, anthropologists, museum curators, etc. To do this it was necessary to bring it down from a mainframe to bring it down from a mainframe to a micro. This has been done and it works on all IBM XT/AT compatibles, running in less than 300k of free RAM. It needed a friendly icon-driven screen system, which it now has. It needed special programs to deal with the kinds of data which anthropologists produce. And it needed simple documentation on how to use and set up such a system. All this has been done and the system, re-named the 'Cambridge Database System', is now completed in a prototype version.

It has exceeded our expectations. To take just the question of speed. Using a fairly slow IBM-PC compatible microcomputer, a search of the 20,000 records which we currently have will take from one to five seconds. A structured (boolean) query for a place name, for instance, retrieved the first 1000 answers out of 1775 in two seconds. A 'free text' search on three terms, occurring respectively 148, 263 and 48 times in the database, assembled the several hundred answers, in decreasing order of probability of interest to the user, in just over a second.

It is possible, using the Naga materials as a test bed, to find the information about any person, any place, and date (day/month/year), any archive, any medium (e.g. photograph) or any ethnic group, or a combination of these (combined with any subject) more or less instantaneously. For instance, if one asks to see all the photographs taken by a certain photographer in a certain month in a certain village, which concerns carved village gates, the photographs will be presented almost immediately.

Of course, the retrieval will depend very heavily on the care and accuracy of the descriptions of the visual and textual items. These are based on the ethnographic and other texts and our accumulating knowledge of the materials. If the ethnographer appears to have made a mistake, this is indicated. Since it is possible to modify the descriptions within the database, this can be an 'open' system which reflects the growing knowledge of the compilers.

Work still to be done

The main task remaining is how to make the methods and materials available to a more general public. Here we can only outline what we hope to do.

The videodisc will be used in a special exhibition on the Nagas in the Andrews Gallery of the Museum of Archaeology and Anthropology when the whole anthropology galleries re-open in Spring 1990. A reconstructed Naga long-house (**morung**) will hold a videodisc and computer, so that visitors can look at films and photographs and listen to sound as a background to the exhibition. It will also be used in teaching at all levels. For instance, it is being used to give undergraduates a simulation of how to asks questions of anthropological data in a 'practical' exercise in their first year. In the second year, it is linked to courses in 'Visual Anthropology'. At postgraduate level, it is being used to show how information retrieval works on anthropological materials.

The videodisc is being used to help prepare a large book with about 500 photographs and analytic text which will accompany the exhibition.

The videodisc itself, as well as the texts and computer discs, are to be marketed. Any profits from this or other parts of the project will go back into a university fund for future research on anthropology. Likewise the Cambridge Database System software will be marketed both for use with optical discs (videodiscs and compact disc) and also as an advanced database system on micros.

NOTES

(1) We would like to acknowledge the support of the following: the Economic and Social Research

Council, the Nuffield Foundation, the Leverhulme Trust, Trinity and King's Colleges, Cambridge, the University of Cambridge and the Department of Social Anthropology, Cambridge. The major members of the team are Sarah Harrison, Julian Jacobs; the computer advisers are Michael Bryant and Dr Martin Porter; the co-director is Martin Gienke.

(2) Macfarlane, Alan and Martin Gienke, 1989. 'The Principles Use in Selecting, Editing and Transferring Materials for an Archival Videodisc', **Journal of Educational Television**, vol.15, no.3, 131-141.