(DigitalOrient; lecture given at Shanghai & Tsinghua Universities Sept. 2004)

Digital Orient: An Experiment

Alan Macfarlane and Xiaoxiao Yan

(About three thousand words at present; also show parts of web-sites etc.)

HISTORY AND AIMS

Alan Macfarlane and Sarah Harrison have been involved in the development of computer databases, information retrieval and multi-media projects for about thirty years. At the University of Cambridge they have pioneered the use of relational and probabilistic search engines, taken part in several multi-media projects including the BBC 'Domesday Disc' and developed a number of web-sites. The work described below, in collaboration with Xiaoxiao Yan and others, builds on this.

A combination of motives lay behind the setting up of Digital Orient (DO).

Returning culture

One of these was the idea that it would be good to make materials held in one part of the world (Cambridge, England), available in the places from where it was obtained.

This had been the idea behind the setting up of the **Naga Project**, the first anthropological videodisc project in the world, in 1986, where very large sets of materials in British archives, museums and private holdings were made available on a videodisc and database for return to the Naga peoples of the Burma-India border.

Approximately 30 minutes of moving film and 10,000 still images collected by five travellers and anthropologists in the first half of the 20th century were transferred to an analogue videodisc, in addition to thousands of pages of text from published works and unpublished manuscripts. The entire system could be searched with a powerful probabilistic information retrieval system. Although videodisc was the height of technology at the time, it is no longer in widespread use and this valuable resource is now inaccessible to most users. We are currently nearly ready to re-launch the system on the web.

As it became possible at the start of the twenty-first century to digitize and make available unique resources held in western collections, we started a web-site called **Digital Himalaya**, with Mark Turin and Sara Shneiderman.

The aim was to assemble ethnographic materials from the Himalayan region. Based in the Department of Social Anthropology at Cambridge University, the project began in December 2000. The Digital Himalaya project has three primary objectives:

1. to preserve in a digital medium archival anthropological materials from the Himalayan region that are quickly degenerating in their current forms, including films in various formats, still photographs, sound recordings, and field notes.

2. to make these resources available on DVD and/or over broadband internet connections, coupled with an accurate search and retrieval system useful to contemporary researchers and
3. to make these resources available on DVD to the descendants of the people from whom the materials were collected by making them both easily transportable and viewable in a digital medium

This web-site is now available at www.digitalhimaya.com

Digital Himalaya dealt with the Himalayan region, but much of the work of Professor Macfarlane was in eastern Asia, particularly Japan and increasingly China. So we thought it would be interesting to set up a parallel site to cover those regions with some of the same goals as digital himalaya.

Representing culture

Part of the aim of anthropology has always been to increase understanding of other cultures; to destroy stereotypes, ethnocentric and orientalizing images. The representation of Japan and China in the media is often negative and out of date. We felt that it would be helpful if we made available a set of recently filmed and textual materials which show what is happening in eastern Asia and, where possible, provide some analysis of what is occurring in those regions today. This would make it possible for those interested in, say, what is happening in Shanghai or rural China to get some glimpses of what an anthropologist with a camera can see in a series of travels across China.

Creating culture

A third motive arose out of the research of Xiaoxiao Yan. She is studying Broadband in Britain and China and as part of this research needs to study the use of this new medium in creating new opportunities for community. Anthropologists usually have their communities given to them, but in this day and age communities are increasingly virtual and distributed, and the web is increasing this tendency.

What better way to study virtual communities than to set one up oneself and study how it is constructed, how it grows, and how people use it? This is the nearest to participant-observation fieldwork one can get to in this kind of research. Participating in building and moderating an on-line community can then be combined with observing the effects and methods employed by those who join in it.

Testing technologies

Another aim is to develop and test out new technologies and new opportunities which are emerging very rapidly. Alongside broadband roll-out, there are new experiments in digital storage such as D-space, and new compression possibilities such as MPEG4. There are advances in search engines of a new probabilistic generation which make searching over large sets of data possible

for the first time. All these need to be developed and tested in particular contexts. D.O. will give us a test vehicle for this.

E-learning and e-archiving

Finally, we are interested in experimenting, in collaboration with our colleagues in China (at Shanghai Jiao Tong University and Tsinghua University, Beijing), with the possibilities opened up by broadband for e-learning and providing e-resources. This is another aspect of Xiaoxiao's research, and again the development of DO makes it possible to test out approaches and effectiveness in an environment which we can study from the inside.

The New Potentials which have become available since 2000.

Mention has already been made of a number of converging technologies which enable various new things to be dreamt of. It is worth elaborating these a little bit:

- firewire, USB 2, and proper digital communications between hardware (connecting cameras, computers etc.)
- video editing suites and particularly iMovie
- large temporary hard-disc storage on firewire and USB hard disks
- permanent archives for editing film – Dspace, THDL (University of Virginia) etc
- web-site hosting costs and facilities improved
- scripting languages and web-designing tools, in particular the suite of programs from Macromedia (Dreamweaver, Flash etc).
- web-delivery, especially Broadband speed and coverage
- compression and cleaning software, esp. 'Cleaner'

To this may be added broader political and social changes, like the rise of digital sophistication in China and India, the spread of the internet in every sphere of life, increasing governmental interest etc. Added to this is the huge potential of the English language market, the brand potential of top academic institutions like Cambridge University, the intrinsic interest of the discipline of anthropology as a subject for the web with its comparative and global perspectives and highly visual component suitable for multi-media.

Digital Orient: an outline of its history and development to the present.

Early history and nature: The setting up of DO in December 2003, ideas in the design of the pages, architecture, content, features etc. as at May 2004. Early materials on. (Xx to write….)

Supplementary features: JADE (Joint Academic Digital Education)

As stated, Digital Orient is a test vehicle and part of a wider project to investigate the uses of broadband and of various designer tools. We are currently developing a suite of programmes to help here. These include:

BAMBOO (SEE longer account in Appendix if needed)

A powerful information retrieval system written as an up-date of the concepts behind the Museum Cataloguing System (MUSCAT). It combines boolean and probabilistic searching, and is based on thirty years of experience in using information retrieval in relation to social and historical materials. It holds the material in XML and outputs the results in HTML, suitable for the web. It is designed to work with multi-media applications, including film, sound files, photographs and texts. It is being developed in association with Lemur Computing of Cambridge.

SILK (Simple Interface for Linking Knowledge) (see longer account in Appendix if needed)

A Website Content Management and Dynamic Page Generation System

The SILK software is designed to help build a collaborative web-site. It provides a simple interface to enable people to input materials from their own work into a web-site without having to learn the deeper structure of the site, HTML scripting languages or Dreamweaver etc. It provides templates which automatically generate consistency of style and links. Major control of the site is maintained by the administrator. Developed by the Digital Orient team.

RICE (Relational Interactive Collaborative Experience)

A specific set of modified 'Blog' software for use in relation to social science and historical materials. Being developed by the Digital Orient team.

PAPER (Private and Public Exchange Resource)

A specific application of the 'Forum' concept, an area to exchange views in relation to social science and historical materials, with various levels of privacy. Being developed by the Digital Orient team.

Broadening out the project: TEA (Treasury of Educational Assets)

The DO project is part of a larger project to make resources available in accordance with the aims stated at the start of this paper. This will be worth describing briefly, and the outlines can be seen on my web-site (www.alanmacfarlane.com)

A wide range of educational resources to be used for research and teaching. These include a database of 80,000 quotations; seven sets of video archives; books and articles; video lectures; travel films and other materials assembled in Cambridge. This links materials on a number of web-sites, including alanmacfarlane.com, digitalhimalaya, digital orient.

There are e-learning resources including several series of lectures and writings and books and articles on a number of themes. Linked to this is a web-site

connected to a new book of thirty letters synthesising all this material on 'How the World Works'.

There is a 60 hour overview of global history, made in association with Windfall television company and C4 television. This includes film from all over the world and seminars and films concerning the development of all the world over the last ten thousand years.

There are archival interviews of over fifty internationally recognized social scientists, travellers and others talking about their experiences and methods of research all around the world. The interviews last from 15 to 180 minutes each.

There will be a database of 80,000 quotations and facts, covering many topics and assembled over thirty years

There are the complete indexed records of an English village covering the period 1380-1850

There will be the Naga materials alluded to above, as well as the Digital Himalaya web-site

There will be a large quantity of material relating to the social history of a Himalayan community in Nepal, including over fifty hours of film and covering the detailed history of a community during the second half of the twentieth century. Film on research methods and how to film in the field is included.

Other materials are planned (e.g. family films, further work on history of the world, working methods etc.)

The longer-term future: some speculations

What could the situation be like by the time of the Beijing Olympics in 2008?

According to Moore's law (computing power doubles every 18 months) and Metcalfe's law (each new link in a network increases the value of all previous links and hence value increases exponentially), things will be very different in even four years.

In terms of data storage. Currently the largest external Hard Disks available off the shelf are about 1 terabyte (1000 gigabytes) in size and cost about £500. By 2008 they will be at least 5 terabytes in size, and a 1 terabyte external hard disk will cost £80 or less. Five terabytes would hold about 5000 hours of film compressed to MP4.

Dspace currently offers our project half a terabyte of storage on the University server. By then it would be about five times that amount.

Editors and computers. The new generation of G5 Macintoshes will compress 1 hour of film in real time (I guess – the G4 takes about double real time). So by then it should be possible to compress an hour of film in 12 minutes or so.

Broadband. The standard bandwidth being rolled out over Britain gives about 1 mb. Download. By 2008 it could be at least 5 mb per second.

So what is still needed?

APPENDICES

1. Nature and development of Bamboo

Bamboo is a fast, flexible, user-friendly information retrieval system, specially designed for use with large databases on a PC or the Web. It is among the most powerful academic database systems available because it combines two sets of skills and experiences stretching over nearly thirty years at the University of Cambridge.

One of these is a set of high level developments in retrieval software pioneered at Cambridge. Four generations of programmers have developed a system combining the best of earlier relational (Boolean) retrieval with the powerful system of probabilistic retrieval developed in Cambridge in the late 1970's. First pioneered by Professor Keith van Rijsbergen and Dr Martin Porter and implemented in the MUSCAT (Museum Cataloguing System) software, the approach has been tested on a variety of historical and social science databases. Recently it has been completely re-written for web work in the 'Bamboo' application.

The programmers have worked alongside historians and anthropologists who have needed powerful, intuitive, software to make it possible to store and retrieve large sets of multi-media data in an effective way.

Many commercial databases are written by expert programmers, but not geared to the experiences and needs of working researchers. They are often expensive and not entirely satisfying. Thirty years of thinking about and using the developing system has ensured that the Bamboo software can deal with all the standard demands of the working researcher as well as those of a range of organizations.

Some features of Bamboo

- powerful searches using 'and', 'or', 'not' (Boolean) logic
- flexible searches using probabilistic ('best', 'next best' etc. answers)
- a combination of the above to improve searches
- terms are weighted by their retrieval potential
- there is powerful 'suffix stripping' or normalizing
- the system caters for the development of multi-media (text, sound, film, photos) databases
- the data is easily entered through an input page or from text files
- different databases with varying indexing strategies can be developed
- searching is very fast and intuitive

- the data is highly compressed (it is held in XML and pages are generated in HTML as needed)
- Bamboo has been proved effective for a number of academic applications
- it can be integrated with accompanying programs to build a collaborative web-site, forum and blog software system
- advice and support is available from Lemur Consulting.

2. Simple Interface for Linking Knowledge (SILK)

A Website Content Management and Dynamic Page Generation System

Introduction

The SILK software is designed to help build a collaborative web-site. Because of the tendency for web-sites to start simply and then become richer and more complex, it is worth considering how to automate parts of the process of adding further links and content. It is very easy to lose track of what needs to be changed each time a new page or category is added to a site. It is very easy for the consistency of style and structure at different levels to be lost, even with the use of style sheets and good web authoring tools such as Dreamweaver.

These problem are magnified hugely if it is hoped to add diverse knowledge from a number of users, some of them with little or no experience of web-site design. This leads to considerable problems. Either the administrator has to do all the work, or guest contributors have to learn the basics of web design, coding in HTML, and the hidden architecture of the particular web-site they are adding to. This latter alternative will debar the majority of potential contributors.

So we have written a simple tool which takes the following hierarchical form:

Front page
A selection of Themes (e.g. countries)
A selection Topics within each theme
A sub-video field (a choice of videos within a topic)
A sub-text field (a choice of texts within a topic)

The last of these takes one to a page which displays the film or the text. To take the example of our web-site, digitalorient.org, the structure might look like this:

(Diagram showing a tree of the structure)

In effect what the system does is as follows.

The administrator retains control of the top three levels (Front, Themes, Topics), as these need to be individually designed and should not be changed too often.

Then an authorized contributor could add complexity and information at a lower level. Through an interface which is accessible on the www, they can type in certain information into a standardized input window for each of the following levels:

Sub-video, sub-text, video record, text record

(Xx – examples of these input pages should be given in the manual, with a brief description of what goes in each box to help people)
When a contributor has filled in the boxes on the input page, the management system will input the content to the database (web-site).

Contributing users do not need to learn how to design web pages. They do not need to worry about corrupting the web-site. They can add their own local knowledge to a collaborative effort, including texts, photographs and films.

So the system will allow a new kind of simplified, collaborative, web-site to develop. The aim is to test it on D.O.