Indexing the Domesday Project

David Lee

A general account of the indexing of a large-scale computerized videodisc project, describing the background of the enterprise and the various processes involved.

The project and its technology

The BBC (British Broadcasting Corporation) planned its Domesday Project as an act of commemoration for the 900th anniversary of the Domesday Book (1086), in addition to the more usual sort of programme, written and hosted in a television series by Michael Wood. The Domesday Project, the idea of Peter Armstrong, an Executive Producer in TV Network Features, would not be a programme or set of programmes but an artefact akin to Domesday Book, albeit more self-conscious, a monument to the 1980s. The Domesday Project would be a survey of Britain not on paper or in print, but would be in videodisc form, accessed by microcomputer.

The Laservision technology of the videodisc certainly existed, from Philips, but full interactivity with the computer had to be developed for Domesday. The project was at times very much a development project, resulting in the named product Advanced Interactive Videodisc (AIV). Previous videodiscs can certainly be read on the Domesday system, but as with so much developing technology, the older versions would not take the new Domesday discs. Bearing in mind the BBC’s involvement with Acorn Computers over the BBC micro, this was the logical micro to use with the videodisc, although a connection was kept with RML, makers of the Nimbus micro, and their version of the system was launched in November 1986, at the same time as the BBC Master series version. It is expected that compatibility with other micros such as IBM PC will be developed in 1987. Almost all the technical ideas which were hoped for from AIV have been realized, despite many problems encountered during the rather less than two years when the project was being prepared. Funding for a project of this developmental type, and of such potentially great size, could not come from the usual BBC funds (licence fees), and was obtained from Philips, the Department of Trade and Industry, and BBC Enterprises. The cost is generally quoted as around £2.4 million, though all three major partners (Philips, BBC Enterprises and Acorn Computers) needed to apply resources not necessarily obvious from the original budgets. It was always intended that the Domesday Project would be a commercial undertaking, placing interactive video more firmly on the technological map, with sales to libraries, all kinds of educational organizations, and every kind of body which requires information of many kinds.

The nature of the project

The BBC Domesday Project is not at all concerned with the Domesday Book, inspiration though the latter was. It is historical only in the sense that it will provide a remarkable source of interest to historians in the future. Of quirky interest to the Domesday team, however, have been the parallel problems which both faced, even to the element of incompleteness in some areas. There was so much more which was considered than in the end was incorporated into our two discs. Whilst encyclopaedic in range, the Project cannot be called an encyclopaedia. There are many subjects not covered at all, left out because other things seemed more important or were less covered elsewhere. The discs may be seen as a contribution to the technology of displaying information of every type, and as an historical monument in their own right, as well as an electronic reference source making material available publicly which was much less so before.

The two discs contain text, pictures (including the surrogate walks), maps, data, moving film and sound, though in varying proportions on the two discs. The text and data is digitized; the pictures and maps are in analog form, put in as photographs. The sources for the material were broadly as follows:

Community disc

- **text** from schools and other local groups: two large sections via the Automobile Association and the Geographical Association.
- **pictures** mainly from schools and other local groups, but a large contribution of aerial photos, and land satellite photography.
- **maps** mainly from the Ordnance Survey (all 1:50,000 series for instance). 70 towns covered by street plans.

National disc

- **text** chosen by a small team of researchers if already published, but some specially commissioned (e.g., 45 essays on wide range of topics).
- Text written also to accompany data sets.

---

Based on a talk given to the Society of Indexers, London, 9 October 1986.
pictures obtained from picture libraries and agencies, but also several thousand from a national picture competition. ‘Surrogate journey’ pictures taken specially on location—house types, town walk, etc.
moving film mainly from BBC news.
sound apart from that accompanying films, is used for ‘Help’ sequences only. On the whole the sound track was used for software.
maps on this disc specially drawn outline maps upon which to plot digitized data. Boundary files in digital form were obtained.
data apart from the schools survey of amenities and land cover, the data is national in scope, obtained from a variety of sources including the Economic and Social Research Council archive at Essex University, the Centre for Urban and Regional Development Studies (CURDS) at Newcastle upon Tyne, the Institute of Terrestrial Ecology at Bangor, and Birkbeck College, London. The list of organizations initiating the data is too long to give here, but credits are given at length on the disc.

An Editorial Board oversaw the content of the discs, the members being eminent researchers in a variety of fields. A number of other committees was formed, of the commercial partners, of data suppliers, of educationists, for marketing, to name a few. The internal BBC team was small, including subject specialists, picture researchers, software writers, text editors, technical people, and indexers. With such a large amount of material—we have estimated that a person reading and consulting all the text, data, maps, etc., would spend seven years proceeding through sides one to four—naturally we needed to co-operate with the outside world, and more than a million people had some connection with the Project. Our bulging files of copyright forms prove that. As far as indexing was concerned the work of many people was involved indirectly, though it remained the direct responsibility of a team of two: one fulltime indexer and myself as Data Services Manager.

The indexing

There were four phases in the indexing of the Domesday Project:
1. schools project keywords
2. national disc thesaurus or hierarchy
3. national disc keywords
4. correction, cross-referencing and editing
Each will be discussed in turn, though it will be realized that to some extent all ran parallel.

Schools project keywords

The Domesday Project was very much dependent on the ‘D-block’ of 4 x 3 km, carved out of the National Grid framework. Note the resemblance in format to the television screen! There are some 23,000 of these blocks in the UK. Twenty pages maximum were allocated to each block, and the software allowed four keywords per page. Three pictures could be provided, again with a possible four keywords each. The sum is simple: 23,000 x 80 plus 23,000 x 12 = 2,116,000 possible keywords for the community disc alone. It struck me that this indexing could be handled by two indexers in a year—I didn’t know too much about the slowness of computers under pressure then—but we were not likely to get funding for the eight indexers that calculation would show were needed for the three months’ indexing time envisaged. Clearly, in the event, self-indexing at source would be required. Advice was prepared for the schools (and other groups) on making keywords, and this, with a list of the main keywords (see figure 1), was printed in the Survey guide,1 as there was an idea of using a general keyword and specific related keywords. This did not come to pass, as totally varied topics were often dealt with on one page.

The self-indexing varied greatly from perfect to appalling, and included some amusing oddities such as those where the space was used for advertisements. All the keywords were checked by a team of Loughborough University graduates/undergraduates, whose task also included the search for libel, blasphemy and one or two other problems. Whilst the text was rarely altered, keywords frequently had to be, if the material was to be retrieved. The indexing of the community disc material is probably rather too good, given quite a generous allowance for keywords. We didn’t incidentally fill all the 23,000 D-blocks, but some 9,000; an excellent score in a

<table>
<thead>
<tr>
<th>Armed Forces</th>
<th>Industry</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arts and Crafts</td>
<td>Information Systems</td>
</tr>
<tr>
<td>Attitudes</td>
<td>Landscape</td>
</tr>
<tr>
<td>Biography</td>
<td>Language</td>
</tr>
<tr>
<td>Clubs</td>
<td>Law and Order</td>
</tr>
<tr>
<td>Communications</td>
<td>Life Styles</td>
</tr>
<tr>
<td>Customs</td>
<td>Local History</td>
</tr>
<tr>
<td>Ecology</td>
<td>Media</td>
</tr>
<tr>
<td>Economics</td>
<td>Politics</td>
</tr>
<tr>
<td>Education</td>
<td>Population</td>
</tr>
<tr>
<td>Employment</td>
<td>Recreation</td>
</tr>
<tr>
<td>Energy</td>
<td>Religion</td>
</tr>
<tr>
<td>Entertainment</td>
<td>Science and Technology</td>
</tr>
<tr>
<td>Events</td>
<td>Social Concerns</td>
</tr>
<tr>
<td>Farming</td>
<td>Supernatural</td>
</tr>
<tr>
<td>Government</td>
<td>Tourism</td>
</tr>
<tr>
<td>Health</td>
<td>Trade</td>
</tr>
<tr>
<td>Home Life</td>
<td>Transport</td>
</tr>
</tbody>
</table>

Figure 1. Primary keywords

These are the primary keywords which were intended to be used with each item on the Community Disc, in conjunction with more specific keywords. They give a good idea of the scope of the topics on the disc.
difficult year that included teachers’ strikes and other unforeseen problems.

National disc hierarchy

Whereas the Community Disc’s main access is through the maps (which are served also by an alphabetical gazetteer), the keywords being secondary, the subject arrangement is paramount on the National Disc. There are in fact two main subject routes into the contents of the National Disc. The first is the hierarchical structure, accessed through an option called CONTENTS. I prepared a hierarchy of subjects, after much consideration of existing classifications, such as that of UNESCO, which was basically used at Essex Archive, a major data supplier; SHIC (the Social history and industrial classification) from the Museums Documentation Association; or the BBC's own Telclass/Relclass system. In the end, a new scheme unique to Domesday was created, as each of the other schemes had a bent of its own, partialities which did not coincide with those of Domesday itself. Domesday's coverage of knowledge is wide, but by no means universal, and its hierarchy may well not be applicable to any other project. In practice it does work well for the needs of the browser through the contents of the disc, particularly in the broader stages. There are four sections into which knowledge is placed, and which are shown to the user who selects the CONTENTS option: Society, The Economy, Culture and Environment. Each of these Level Ones is divided into as many Level Twos as seems necessary (see figure 2). There are 45 terms at this level, each of which is further divided, and so on through further levels. The degree of specificity, the number of levels reached, is dependent on the amount of material to be covered (data, pictures, text). A limit of division to twenty was laid down for each level, which worked quite well until the bottom level was reached, that upon which the actual named items hung. At times the terms numbered over twenty, and rather artificial divisions had to be created, or items would simply not be found through the hierarchy. In no way could one declare that the structure we have ended with was totally logical, for there had to be continual interplay between the ideal and the practical, and Katherine Dawson, Domesday's indexer, often had to make decisions which resulted in slightly awkward terminology, or placing. No term could be repeated in two parts of the hierarchy; i.e., each term must be totally unique, or the computer would retrieve (wrongly) both sets of material.

A further drive away from central control as the indexers might have seen it come from the influence of the Editorial Board, which would demand that a subject should be represented on the Domesday disc, and that it should be placed in such and such a place. Members would fight quite hard to place a subject within their own sphere of interest. Planning as a subject could be seen as part of Society, or as part of Environment; Drugs might be within Society or Culture. The library classifier's daily

SOCIETY
Armed Forces & Defence
Education
Health
Housing
The Law
People & Events
Politics & Government
Population
Science & Technology
Social Welfare
Transport & Communication

CULTURE
Arts & Entertainment
Beliefs & Attitudes
Consumption
Customs & Traditions
Home & Community Life
Language
Leisure & Recreation
Life Style
Mass Communications
Religion & Philosophy
Sport & Games
Tourism

THE ECONOMY
Economy as a Whole
Finance
Industry
Overseas Trade
People at Work
Personal Finance
Public Sector

ENVIRONMENT
Agriculture
Climate
Conservation
Environmental Pollution
Forestry
Geology
Industry in Environment
Landscape
People in the Environment
Planning Issues
Soils
The Seas
Urban Environments
Water Resources
Wildlife

Figure 2. Level 2 topics

These are the second-level terms in the hierarchy on the National Disc, within the four groups, SOCIETY, THE ECONOMY, CULTURE and ENVIRONMENT.

problem of placement of a physical object, and coverage elsewhere by cross-references or added entries, was hard to convey to the board members. It is true that in the final presentation of material on the screen, the ‘added entries’ are subordinated to the main lists of material, and can be overlooked.

A major crisis arose with The Economy, where one board member pressurized the team gently but firmly towards accepting his own view of the structure of the subject. As he is a Professor of Economics his scheme is dynamically better than my own, but did exclude what one might call ‘natural’ places for subjects which were less part of the academic study of economics. The adoption of the term ‘The Economy’ itself, rather than ‘Economics’, was a direct decision by the Board, indexers tending not to like definite and indefinite articles in such circumstances. ‘The Seas’ and ‘The Law’ are the two other such Board choices.

Two other points are worth mentioning about the hierarchy:

(a) it became a shopping list. This was indexing as prescription, and it therefore often became a battleground between and amongst the Board members and Domesday’s own specialists, the indexers perhaps as punchbags in between. Discussions of contents were by

The Indexer Vol. 15 No. 3 April 1987 147
thesaurus term, huge lists of items being produced for every Board meeting, showing the latest stage data/picture/text collection had reached. Internal specialists and outside suppliers, represented on the Board, fought for items to go or not go into their sections. The indexers' knowledge and terminological skills were often called upon.

(b) as with the keywords, described in the next section, the providers of material had to allocate terms from the thesaurus, one principal term, and up to four more added entries. There was central control to ensure uniformity, but the very speed of operation, and the amount of material to be covered (9,700 data sets, 2,000 pieces of text, 500 picture sets—some 25,000 pictures in all) meant that keeping to the same vocabulary and recording changes was a problem to outside suppliers and to us. One supplier behaved badly by not providing thesaurus terms for a large number of items, no doubt under pressure, and none found it easy.

National disc keywords

Each picture set, text item and data set on the National Disc could have its five hierarchical places, as mentioned in the last section, and also a maximum of twenty keywords per item. Whereas the thesaurus terms were fed into the CONTENTS system, as a browsing device, the keywords were intended as a freer entry to material whatever its origin or structured place in knowledge. In some instances, say a short piece on SDP (Social Democratic Party) policy on unemployment, the allocation of twenty was generous, but in the case of a picture set on buildings or the multi-role royal family, it was very limited. Keywords were applied by the data/text/picture providers. Some were our colleagues, with varying degrees of skill; others we never met. All textual matter came through the hands of the indexing team at quite an early stage, but picture sets and data sets were less likely to do so, and indeed some data indexing was never checked in house. Moving film and surrogate walks were indexed at a late stage only, but were very largely in the indexers' hands, and the most thought given to them. Without doubt the indexing would have been better (more consistently, more imaginatively) done if it had all been done in house. Without much doubt also, we would never have finished it all, particularly if every item had to be seen by one of the two of us. The publishing date of 25 November 1986 had been set quite early and it was not a date we could afford to miss, in view of other Domesday celebrations.

Keyword cross-referencing, editing, etc.

So far this account of Domesday indexing has dealt with the co-operative processes in which schoolchildren, researchers and suppliers have supplied either keywords or terms in a hierarchy, all saving the time of the in house staff. We turn now to the index editing processes. The keywords on the Community Disc and on the National Disc were built into two quite separate alphabetical indexes on our VAX computer, and the indexing team had access to these at more or less all times. We were able to see which terms were building up in what numbers. We could also usually make alterations, and to this end were supplied with error reports in computer printout. We were thus able to approve potential keywords or alternatively to kill them. A stop list had been created of words like 'a', 'and', 'but', which could not be used in indexing. The number of mis-spellings was enormous, on both discs. It seemed almost impossible for people to spell 'abattoir' or 'restaurant' correctly. Correction was long and tedious work, as we had to go into the different systems used for text, data and pictures. Although our computer colleagues had made these systems fairly friendly, continuous hours of work made us less so. Correcting mistakes easily led to fresh mistakes too, in a situation where the work had to be finished by particular dates to enable technical operations to take place. Although the size of the editing process was large, the need for knowledge of what had gone on before made the addition of new staff pointless. In addition, the more systems in use on the same operation, the slower computer access became. Even with two workers at times, the speed of computer processes was intolerably slow.

The actual retrieval system, evolved by Dr Martin Porter of Cambridge, involves the use of stemming. He had devised an algorithm by which words were stripped of their endings 'ation', 'ity', etc., according to quite subtle rules. Even so, there are oddities to those working within the system, if not, one hopes, to the user. In the print out, RAI indicated RAY/RAYS, GA=GAS, LINSE=LINSEED, and HOGWE does not immediately express HOGWEED.

It had always been the intention to make cross-references in the keyword structures, as items found by the hierarchy could also be called up at other points in it. But whereas the cross-references in the hierarchy have positively to be called up by the user, with keywords, if the cross-references are made, material shown under one heading, e.g., PHONE, can also be found by asking for TELEPHONE, and vice versa. There were many such instances of true synonyms, but also many where there was some overlap, such as WATERWAYS/Canals/Rivers. In such cases, one-way references were made, in effect 'WATERWAYS see also CANALS' etc.

Much care had to be taken not to be too clever in making cross-references. One thought of, between MASS and COMMUNION, was ditched when second thought made one realize that MASS was more likely to occur as part of MASS MEDIA, or MASS COMMUNICATIONS, than actually meaning the religious service. Quick checking of context was not possible, and caution had to be exercised often. What one might call abstract cross-references, i.e., those where one of the words was used and the other not, were not made. On one disc GIPSYES was used only;
on the other GYPSIES only. In retrospect one can see that this could have been got round by falsely indexing one usage, which indeed leads one to a conclusion that many common mis-spellings could be catered for by allowing cross-referencing under words like RESTAURANT and ABBATOIR. Perhaps correct spelling is a wholly outdated concept again.

A few problems

Indexing for and with the computer was a new world for me, and although all indexing depends on assessment of content and the control of words, there were a few new tricks to be learned.

(a) Alphabetization Here the computer undoubtedly has an influence. Compound words are preferable to pairs with hyphens. Fancy typography is nearly always a mistake. Standards created must be adhered to.

(b) Keywords, thesaurus terms and titles We had to consider the relationship between these on the National Disc. In the event it was decided that the hierarchical thesaurus terms would also be in the keyword index; but the sooner a decision like this is made, the better, so that assignment of terms during indexing can take account of the result. Obviously, sometimes the same words would be involved. Item names (titles), on the other hand, were not included as component words in the keyword index. Items can be retrieved by their names, if known, by placing them within double quotation marks. I am pleased we took this decision, bearing in mind the allusiveness of titles and the false items which would have been retrieved.

(c) Space allowed per item All computerized work demands a well thought-out pattern, down to questions such as spacing, or capitalization. We allowed 24 spaces for keywords, which was quite adequate; 30 spaces for thesaurus terms, which was not too bad; and 30 for item names, which was definitely too little. Deviation from the rules always leads to an error, if only a wrong column, and software answers to correct such can be very expensive.

(d) Treatment of different materials Did the various materials demand different treatment by the indexer? I think the answer to this is yes, provided that the anonymous final user must be the person to whom the indexer addresses him/herself. The following terms were provided by different areas of work: M.WKG.MACH.: EMP % CH 78-81 (27 characters for a data set dealing with employment in metalworking machine tool industry, not a potential index entry one would recommend); ART AND OPINION ON THE STREET (29 characters for the title of a 65 item picture set); HIPPY CONVOY IN LOCAL CONFLICT (30 characters for a text item). None of these is self-indexing; all are quite different in approach but require massaging for the final user.

(e) Personal names Originally these were not to appear on the Domesday discs, bearing in mind the Data Protection Act, but in the end the names both of authors of text, photographers and artists, and of people in the public eye, do appear and are indexed.

(f) Place names The Community Disc includes over a quarter of a million names, and the amalgamation of the Irish with the Ordnance Survey gazetteer presented some tricky problems, not all solved. SAINT, ST., ST., and any other versions I haven’t discovered, are still in separate groups.

(g) Abbreviations These are a problem, particularly in the hierarchical terms, which were incorporated in the keyword index, as mentioned above. Some were removed in editing; others remain and will only be found by those who play tricks with their FIND facilities. The real point is less what silly things are there, like Hh for HOUSEHOLD, as well as HHOLD, but that these should be covered by the full word as well, in the index.

(h) Plurals We preferred these to singular versions, but enough are irregular to need special attention during editing.

(i) Gallery In addition to the two approaches on the National Disc, FIND (keywords) and CONTENTS (hierarchy), sophisticated software was used to create GALLERY, which enables the user to walk around an artificial gallery consisting of several rooms based on some of the terms in the hierarchy, call up picture sets and essays, and walk through into other environments. I would describe it as a game rather than as a browsing mechanism, but it does give a third mode of entry into the disc’s materials.

Conclusion

Finally, it is worth asking what the indexing systems of the Domesday discs will produce. Have we succeeded in providing a key to the almost boundless treasures therein, for there is much extremely useful and interesting material to be found? The answer is yes; an answer immediately qualified by saying that in some ways the indexing will produce sometimes more than you require. On the Community Disc, for instance, many thousands of entries will be revealed by keying in LANDSCAPE, for that was a most popular index keyword. In practice, the index on the Community Disc operates on the map displayed on screen, which may be the whole country (Northern Ireland, Isle of Man and Channel Islands are included as well as the mainland), so that for the level consulted, usually a 40 x 30 km block or a 4 x 3 km block, the hits would be able to be coped with.

The whole of the material—maps, text, film, data and pictures—is all traceable, but the user who will gain most from Domesday will be he/she who studies the system. There is a user handbook which comes with the system, and guide cards/posters are being considered to aid quick access to the discs and explain what in fact is a very simple grammar.
If we began again, the whole project, we would do a better indexing job. We would probably standardize more, lay down stricter rules to be followed. But then, we might lose some of the strong individualism of terminology which comes through from the grassroots of British society, on the Community Disc particularly. The voice of the people really does come over, and all that we indexers have done is bowl the hoop along a bit. It is hoped, incidentally, that there will be a revised edition of the Domesday discs, to enable late-coming material to be incorporated and add some fresh pieces to be included. As if seven years of potential consultation were not enough!

References

David Lee is Managing Editor, Domesday II, in the BBC Interactive Television Unit at Elstree. A former Chairman of the Society of Indexers and Deputy Editor of this journal, his substantive job is as Manager of the BBC Hulton Picture Library.

Food for thought in The golden bough

The golden bough by Sir James Frazer (Macmillan, 1925) is one of those books that can be opened at the index—or any other page—and provide stimulating reading. Often answers to unusual and awkward questions asked by precocious children can be found in The golden bough.

Do Australians eat kangaroos? We find in the index:

Kangaroo, eaten to make eater swift-footed, 496

This answers the question without even looking up the reference. The reference is worth looking up though, just to see what else is on the page. This gives good grounding on related food questions.

It must be admitted that there are omissions from this index. For example it should be considered fortunate that the food question wasn’t, why did a Chinaman eat a whole tiger? The only heading for tigers is:

Tigers, respected in Sumatra, 519

nor is there a heading under Chinese, and under China we find:

China, emperors of, 9; charms in, 35; geomancy in, 36; modes of compelling the rain-god to give rain in, 74; trees planted on graves in, 115; convulsions attributed to the action of demons in, 186; custom as to shadows at funerals in, 190; ceremony at the beginning of spring in, 468; popular superstitions in, 498; human scapegoats in, 566; expulsion of evils in, 567

The answer to the question is on page 496. This was found by chance whilst looking up kangaroos. It is still possible to save the day by reminding the questioner that nowadays a Chinaman is generally considered to be an off-break bowled to a right-handed batsman by a left-handed bowler.

The index suffers from most of the usual faults—strings of page references. Some headings should have been combined, as in several instances the singular and plural forms of nouns have been used as separate headings:


Identical names should have been differentiated;

Nails, used in magic, 44; knocked into trees, 127; used as charms against fairies, 226
Nails, parings of, used in magic, 13, 233; swallowed by attendants, 229; disposal of, 233–7

In spite of all its faults, the index makes fascinating reading. Just to give a few more examples:

Caterpillars, precautions against, 531
Footprints, contagious magic of, 44, 45
Lapland, tying up the wind in, 81
Washing, forbidden for magical reasons, 31

One last observation on the kangaroo meat: time has shown that eating it doesn’t make one swift-footed enough to catch rabbits, as any Australian will vouch. An indication, perhaps, that some of Frazer’s information is now out of date!

To end on a controversial heading:

Water-ousel, heart of, eaten to make eater wise and eloquent, 496

Could it be true as several writers have noted in the past—we are what we eat? If so, what should indexers be eating?

M. W. Gilchrist