Providing Web access to a catalogue of British newspaper cartoons

J D Bovey

October 11, 2002

The Author

Dr Bovey is a Lecturer in the Computing Laboratory, University of Kent, Canterbury UK. E-mail: jdb@ukc.ac.uk

Keywords

Cartoons, Catalogue, Internet, Images,Information Retrieval, Access statistics

Word Count: 4799

Abstract

This paper describes how the University of Kent Cartoon Centre catalogue was made accessible on the World Wide Web and analyses the effectiveness of the search site during the four years it has been online.

The Cartoon Centre catalogue covers a wide range of British newspaper cartoon drawings and is unusual in being a substantial (over 90,000 records) online catalogue that also includes digital images of all the catalogued material. The paper describes some of the decisions that were made in putting the catalogue on the Web and then uses the evidence of the Web server logs to draw some conclusions about how successful the Web interface has been and how it might be improved in the future.

1 Introduction

The University of Kent Centre for the Study of Cartoons and Caricature (more usually just called the Cartoon Centre) was established in 1973 when the university became the custodian of a collection of original artwork of 20th Century British newspaper cartoons. The collection,

which included large bodies of work by, among others, David Low, Sidney George Strube and Victor Weisz (Vicky), was recognised as being a useful historical research resource and so the Cartoon Centre was set up and work started on cataloguing. The original catalogue was based on cards and photographs, but eventually it was replaced by an online catalogue with digitised images. Since the Cartoon Centre was set up, the collection of original artwork has changed as new donations have been acquired and other, loaned, collections returned, but the online catalogue has gradually expanded to become a research resource in its own right. Although the catalogue originated as a way to make the original drawings more accessible, it has gradually taken over as the main product of the Cartoon Centre. For example, the recently completed *CartoonHub* project, funded by the Research Support Libraries Program (RSLP) expanded the catalogue to include cartoon collections at John Rylands University Library of Manchester, the National Library of Wales and the library of the London School of Economics. This means that, in addition to covering original artwork held in the Cartoon Centre, the catalogue now covers external collections as well as a large number of cartoons catalogued and scanned from newspaper and magazine cuttings.

Until 1998 the catalogue was only accessible from within the local area network at the University of Kent. Researchers who wanted to use the catalogue and collection had to travel to Kent or phone in their query and get the search results sent through the post. This initially made sense when the catalogue was an access tool for the original drawings, but the rise of the World Wide Web and the growth of the catalogue as a useful research resource in its own right meant that providing Internet access became the natural thing to do.

The main Cartoon Centre (now CartoonHub) Web site can be found at

`URL: http://library.ukc.ac.uk/cartoons`

and has links to the catalogue described in the rest of this paper.

2 What is Different about Cartoons

The most important part of a cartoon is the drawing, and it does not really make sense to provide a catalogue of cartoons without including the digitised images in some form. This obviously implies an overhead in terms of storage, as well as down-load time when viewing records in a Web browser. On the other hand, as images go, cartoons are relatively easy to show online. Most of them are ink drawings designed to be reproduced in black and white and so they compress well and are relatively cheap to store and transmit.

2.1 The Catalogue Data

Before describing how Web access to the cartoon catalogue works, we need to say something about the catalogue records, images and other associated data.

Each cartoon has a catalogue record and a digital image. The records are subdivided into fields, including mechanically derivable ones like the date and place of publication, the name

of the artist, the caption and any text that can be transcribed from the drawing (from speech-bubbles, for example). Unfortunately, useful though these fields are, they do not contain enough information to allow cartoons to be retrieved by subject. The only real way to provide this is for the cartoons to be manually indexed with additional subject terms. This cataloguing has always been a major activity within the Cartoon Centre and was started long before there was any idea of having an online retrieval system. As well as subject terms, the cataloguers also add:

- names of any real or fictional people caricatured in the drawing
- any piece of text that is implied by the drawing but is not explicitly included - this is usually a quotation or a line from a well known song or poem.

2.2 The thesaurus

For collections like ours, where subject indexing has to be done manually, there are strong arguments for using a controlled index language. In the pre-computer catalogue we used a list of approved keywords but now we use an inhouse thesaurus (see Bovey, 1995). This is a conventional thesaurus with a linked structure of preferred terms, entry terms, broader terms and narrower terms (Aitchison, 1987; International Organisation for Standardisation, 1986), and currently contains about 11,000 individual preferred subject terms. The existence of the thesaurus has an impact on the design of the Web interface because any search interface to the catalogue also needs to provide searchers with a way to select appropriate index terms from the thesaurus.

2.3 The problem of copyright

One thing that makes a catalogue of cartoon drawings different to, say, a catalogue of printed books is that there is an issue of copyright. Although the copyright of the catalogue records belongs to the Cartoon Centre, the copyright of each cartoon drawing belongs to either the artist or the original publisher. This means that each copyright holder's permission has to be obtained before their cartoons can be included. Copyright holders have to be convinced that the images in the catalogue cannot easily be downloaded or used without permission (in a book or magazine article, for example). Since an image on the Web is, by definition, downloadable, the only protection is to try and ensure that the images in the catalogue are not suitable for reproducing in print. For the lower resolution images this is not really a problem, but the full resolution images need to be marked in a way that spoils them for printing but does not hide any detail. We do this by adding a Cartoon Centre logo to the full-sized images before they are downloaded.

2.4 Viewing search results

Another decision that has to be made in designing a Web-based cartoon catalogue is how to present the initial results of a search. Ideally, the system should present a compact list of summaries of matching cartoons, with some information about each, so that the searcher can select those that look worth viewing in more detail. Unfortunately, there is no reliable way to mechanically summarise a cartoon. They do not usually have titles, and many cartoon captions do not, by themselves, give much idea of what the cartoon is about. A reduced *postage-stamp* sized image of the cartoon drawing is compact in screen space but the amount of information it gives about the full cartoon is very dependent on the style and nature of the drawing. A simple drawing with bold lines may be clear when reduced but a scaled down copy of a detailed drawing may not give much information at all, especially if the drawing contains much text. On the other hand, whatever the drawbacks, it is hard to see any practical alternative to using small images with a few text details.

2.5 User registration and logging

The Cartoon Centre is funded as an academic research resource and so, when the catalogue was made freely available on the Internet, we wanted to try and keep a record of who was using it and what they were using it for. The Web server will automatically keep logs of the requests it receives and these contain a record of the queries submitted and the records displayed but they don't normally contain any information about who submitted the query. The only way to get this information is to ask users to register and then flag each logged request with a user identification code. The next decision that needed to be made is whether to use heavy-weight or light-weight registration. A heavy-weight registration procedure would be one in which prospective users are asked to fill in a Web form with details about themselves and why they want to use the catalogue. In return they would be given a login name and password that they could use to get access to the catalogue itself. The advantage of this approach is that it would provide fairly reliable user information but the drawback is that we felt that it would put people off. Instead, we opted for a more light-weight process in which new users are required to register by entering a few personal details but, having done this, they can then go on and use the catalogue without having to login or use a password. This is implemented using persistent cookies so that, as long as a user keeps using the same PC (their own desktop or home PC say) they should only need to register once.

3 The Web Search Interface

The pre-Web search program that was used (and is still used internally) is called *prism* and is shown in Figure 1. A searcher can type a boolean query into any of the five boxes at the top. A summary of the search results is shown as a time series in the area underneath and the searcher can look at individual records by clicking on a chosen date. The selected cartoon is shown in the lower half of the window, with the image and the catalogue data side by side.

When planning the Web interface, we considered trying to implement something like the *prism* user interface in a Web page. While this would certainly have been possible, in the end, we abandoned this approach as being too expensive in download bandwidth, not sufficiently intuitive and too dependent on features that may not be available in all Web browsers. Instead, we decided to use a more conventional, form-based, approach. A searcher who wishes to search the catalogue sees a page like that in Figure 2: a form, followed by an explanation of how the catalogue works. The searcher can choose to see the results of the search with either one or 16 records per page. When summarised at 16 cartoons per page, the images are shown reduced to postage stamp size and with just the artist's name, date of publication and caption. The searcher can then click on any of the tiny images to see that cartoon's full catalogue record (Figure 3). Finally, a user who wants to see an image in more detail can click on the displayed image to view it in the highest resolution available. These full-resolution images are usually much too big to fit on a PC screen but can be scrolled to show fine detail in the drawing. They are also the only images that are of printable quality and so they are over-printed with the Cartoon Centre logo made up of randomly distributed dots (see Figure 5).

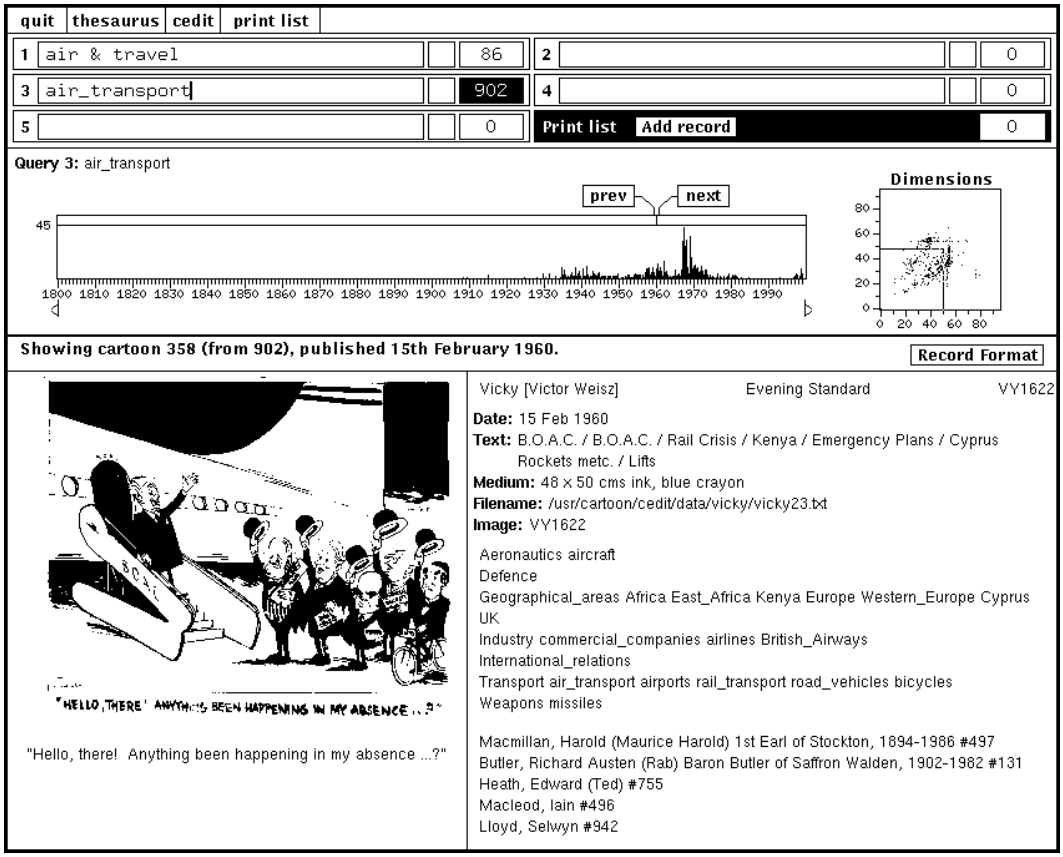

Figure 1: The *prism* search program

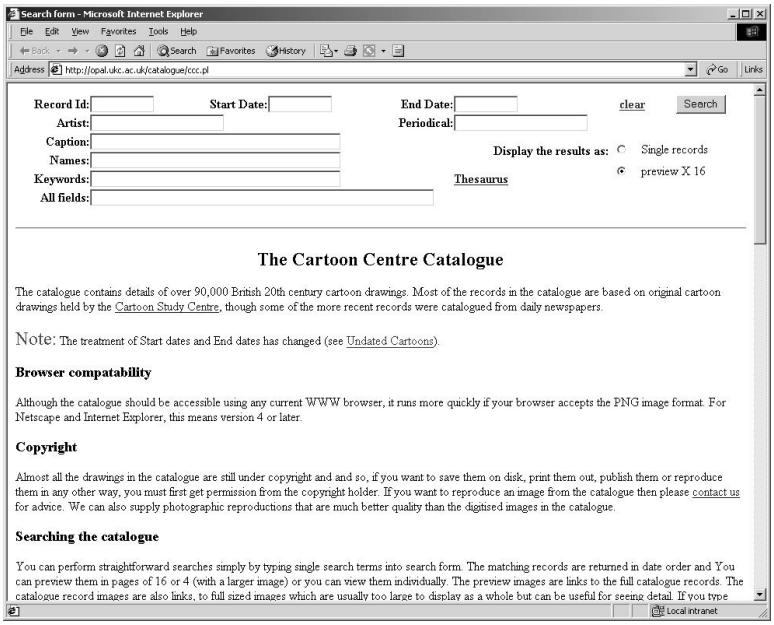

Figure 2: The Web search form

3.1 The thesaurus

Searchers who want to use subject keywords need to be able to look them up in the thesaurus. One possibility would be to provide the thesaurus as a pdf file that can be downloaded and printed, but a printed version would soon become out of date and, anyway, it is really too big for many people to want to print it. The natural alternative is to provide thesaurus searching and browsing in a Web page that is linked from the main searching Web page. This is what we did and the Web page is shown in Figure 4 – a user can search for all the the subject keywords containing a given string and, having found a valid subject term, can follow links to broader, narrower and related terms.

3.2 Other image sites

When we originally set up Web access in the summer of 1998 we did not know of any other substantial, Web accessible catalogues of images so we had to design our own. Since then, of course, other image collections have appeared, and they invite comparison with the cartoon catalogue. One example is the Visual Arts Data Service (VADS) (Purdy, 2001), part of the UK Arts and Humanities Data Service. VADS provides a hosting service for image collections to the British Academic Community, and also publishes guidelines on *Best Practice* for academic groups which are planning to set up image collections. The VADS retrieval service is powered by a commercial retrieval system and provides some kinds of searching that we do not, for example searching on adjacent terms. On the other hand, the way in which VADS presents the results of a search is very similar indeed to the way the cartoon catalogue does it, with rectangular arrays of thumbprint images leading to catalogue records, that lead, in turn, to

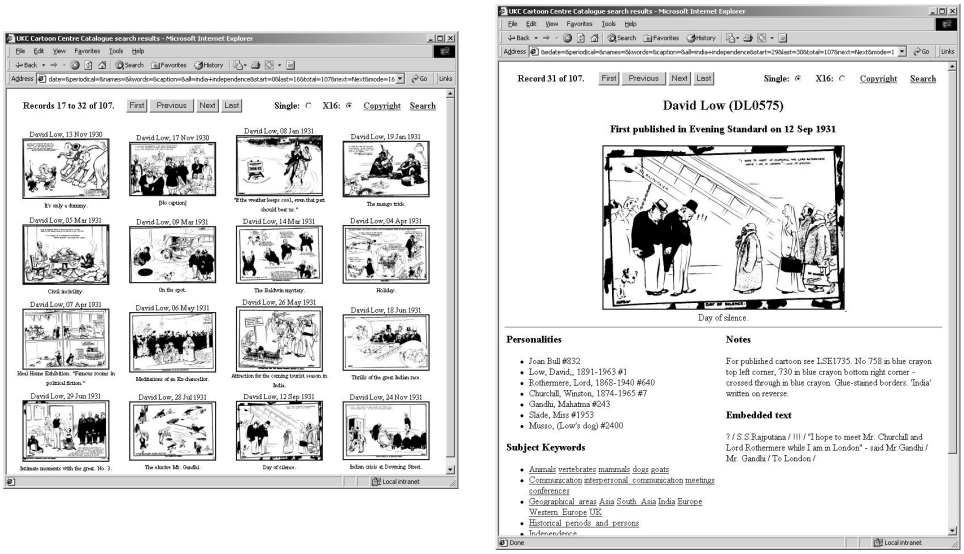

Figure 3: Search summary and a single record

large images with more detail. The VADS system differs from ours though in that even their large images are limited to 800 by 800 pixels and do not have logos to discourage copying.

4 Implementation

The Web searching is implemented using straightforward perl cgi scripts that construct html pages of search results – either summary pages or individual records. These perl scripts need to be able to search the database and retrieve lists of records to present (*prism* uses an inverted file database that is rebuilt every night from the raw text catalogue records). The accepted standard for interfacing software to bibliographic retrieval systems is Z39.50(Z39.50, ???) and we did look into the feasibility of implementing a Z39.50 gateway for the cartoon catalogue database. In the end, we decided against doing this, mainly because Z39.50 has a stateful, connection based, protocol that is not very compatible with the stateless http protocol used by Web servers. Interfacing http and Z39.50 means either:

- opening a new Z39.50 connection to handle each Web page and then closing it again – this is rather inefficient;
- using an intermediate process that sits between the cgi scripts and the Z39.50 gateway and keeps the connection open – this seems to add a lot of unnecessary complication and would inevitably make the system more unreliable.

In the end, we decided to go our own way and use a simple search program built from *prism*'s back-end modules. This runs continuously and provides a network socket that the perl cgi scripts can use to submit search requests and retrieve records. The search requests are in our

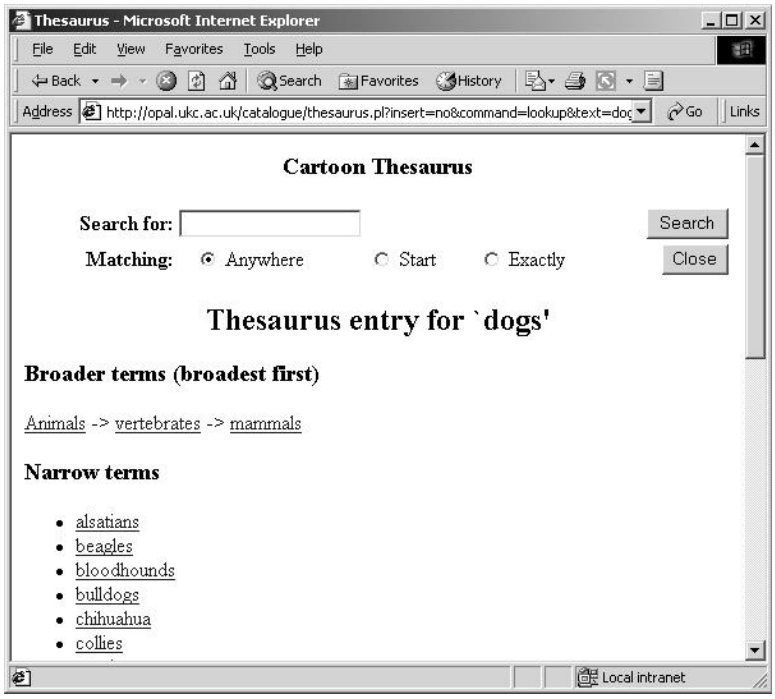

Figure 4: A thesaurus browsing Web page

own boolean query language (the same as is used in *prism*) and the retrieved data are raw records straight from the database. This means that the interface is very non-standard but at least it was simple to get working and has been reliable in practice.

4.1 The Images

The digitised images are stored in separate files in a Unix directory tree and are simply retrieved by name when they are needed. For reasons of efficiency, each image is actually stored in three sizes.

full – the highest resolution image that we have. These vary in size and format depending on when and where they were scanned, but the bulk of them are monochrome PNG (Portable Network Graphics) or, less often, JPEG (Joint Photographics Experts Group). These full-sized images are generally too large to fit in a Web browser window.

standard – medium sized images for viewing individually. These are generated from the full images and are either gray-scale PNG or JPEG, depending on the format of the full image.

tiny – postage stamp sized images for displaying 16 at a time in search summaries. These are either 4-bit grayscale PNG or JPEG depending on what kind of image turns out to be smaller. The script that generates the tiny images from the standard images actually creates a PNG and a JPEG version, compares the sizes of the resulting files and throws away the larger one.

Figure 5: A full resolution image with superimposed logo

4.2 Adding the logo

For the reasons discussed above, the full-sized images are displayed on the Web with an added logo. This is similar to the addition of a digital watermark in that it identifies the source of the image and discourages illicit copying but, unlike a digital watermark, it is visible rather than hidden. The logo needs to be added without obscuring detail in the image and, for obvious reasons, should not be too easy to remove using software. For coloured images this could be done by superimposing a tinted logo of some kind, but this is not possible with most of our images since they are monochrome. The approach we use instead is to impose a logo made up of a random pattern of inverted pixels. If the density of inverted pixels is not too high (we use 1 in 4) it creates a clearly visible mark without hiding too much detail. The randomised pattern of dots is different in each image, making it hard to remove mechanically.

5 Has Web Access been a Success?

We originally introduced Web access as a convenience for researchers who already knew about and used the catalogue. It has clearly been a success in this respect since the number of researchers needing to visit the Centre in person fell from more than 50 in 1997 to single figures

in 2001, and the number of telephone queries has reduced from more than 500 to almost none in the same period. This reduction is clearly not because cartoons have gone out of fashion since the number of requests for photographs and digital images of cartoons has increased.

5.1 Locations of Registered Users

Another test of the effectiveness of the Web interface is the extent to which it has introduced the catalogue to new users. One source of information on this is the file of details entered by people registering to use the catalogue. After filtering out multiple registrations and also registered users who never managed to retrieve any records (discussed below), there were 4141 registered users by June 2002. It would be interesting to know where in the world our users are based but, unfortunately, the registration form does not ask for a physical address. On the other hand, the form does ask for an e-mail address and a proportion of these have domains with country codes giving some information about where people live. These codes show a surprising worldwide spread of interest in what is, after all, a national collection. In addition to clusters of users in English-speaking countries (UK(1159), Australia(63), Canada(65), New Zealand(39)) and Europe (Germany(78), France(37) and the Netherlands(65)), there are significant numbers of registered users in South America, Japan, Korea, Israel and many other countries. There are also, of course, large numbers of users from non-national domains (for example, .com(1890), .edu(166), .org(69) and .net(275)). It is clear from these figures alone that the catalogue must have acquired many new users since being on the Web.

5.2 Evidence from the Server Logs

Another source of information about use is the logs kept by the Web server. There are a number of problems with analysing Web-server logs – for example, a proportion of requests are satisfied by Web cache machines and never reach the Web server at all. Another problem is that, although a user typically undertakes a *search session* with searches alternating with looking at the retrieved records, the server logs a sequence of unlinked requests for pages or images. This means that an essential first step in analysing the logs is to try and group the logged requests into sessions, with each session representing an attempt by a single user to search the catalogue. It is difficult to identify sessions reliably but we used sequences of requests by a single registered user that have no breaks of longer than 20 minutes. After removing search sessions that did not retrieve any records at all, this left 12,124 search sessions between October 1998 and May 2002. While that number inevitably includes a high proportion of frivolous searches, there were also many substantial search sessions. For example, 40% of searchers looked at six or more full catalogue records and about 6% (724) looked at more than 50. About 30% of searchers retrieved at least one full sized image. There were also a few very substantial searches that lasted several hours and must have involved the searcher looking at thousands of cartoon records.

5.3 Searches that Retrieved Nothing

Another use of the logs is as a source of ideas for improvements to the catalogue and its Web interface. One natural place to look for things that do not work is the search sessions that got no hits at all – there have been several thousand of these. While many are the result of well-formed searches for subjects that are not in the catalogue (cartoon films or non-UK cartoonists for example), others failed because the searcher misused the search form in some way. Of these, some kinds of misuse occur repeatedly and could probably be avoided by a better designed search interface. For example:

- Many searches failed because people typed free text into the keywords field.
- The date fields seem to be a particularly common source of problems – particularly the non-intuitive way that start dates are inclusive whereas end dates are exclusive.
- Other searches failed because the searcher did not understand how the search fields combine – for example, some searchers respond to a failed search by adding more detail to the form.

These failures suggest some potential improvements that are discussed in the next section.

6 Future Improvements

The current Web interface was aimed at academic researchers and they are still the most important users, but it is clear that since going on the Web the catalogue has acquired many different kinds of searcher. For example, the list of registered users shows evidence of there being quite a lot of use in schools by school children. This suggests that it is time to produce a better, more user-friendly, Web interface to the catalogue. Among the likely improvements, suggested in part by the failed searches, are:

- Better integration of the thesaurus browsing pages and the search form so that keywords can be selected from the thesaurus rather than typed in.
- Follow popular search engines and provide a choice between simple and advanced search interfaces. The simple alternative would have a single query box and would present a ranked list of matching records whereas the advanced interface would be an improved version of the current one.
- Give better feedback when a search fails to retrieve anything. The most obvious thing to do here would be to break the search into separate terms, search on them individually and point out any that do not have any matching records.

Work is already underway on this improved Web interface and it will be installed and online by the time this paper appears in print.

6.1 Image quality

When we first started scanning cartoons for the database we were setting up a search tool for a collection of original drawings. This meant that the quality of the scanned images was not too important (as long as they were good enough to give a fair idea of what the cartoon was about) since the original drawings were always available if needed. This is no longer true – the online images are all that is available to most of our Web searchers and they need to be good enough to stand alone. As a result of this, we have improved the quality of our current scanning but we have also done quite a lot of rescanning of the more important drawings.

7 Concluding remarks

The development of the Cartoon Centre online catalogue seems to have been rather unusual in that we have used our own retrieval software, cataloguing tools and file formats. One reason for this is that we were in many ways ahead of our time. When we started converting catalogue cards into online records the available record format (for example MARC) did not have fields suitable for describing cartoons, and flexible data formats like XML were not available. When we started to include images, the only hardware that had appropriate, large, high resolution screens were expensive Unix-based workstations and there certainly was no suitable retrieval software available off-the-shelf. Also, when we first made the catalogue available on the Web, there were no other examples to follow or off-the-shelf software that we could use.

Another reason for our rather unconventional setup is the informal nature of the collaboration between the Cartoon Centre and the author. The Cartoon Centre was originally created as a separate Cost Centre within the University Social Science Faculty (with HEFC (Higher Education Funding Council) Special Factor Funding). It had enough money to pay for a curator and a cataloguer but could not afford expensive software packages or dedicated computer support. When I became involved, I was (and still am) a Computer Science lecturer with an interest in information retrieval. All the software development had to fit in along with my teaching and other research commitments. Essentially, this situation still continues – the Cartoon Centre now employs four people, is part of the University Library and has been funded recently by a series of one-off grants to do Cartoon related projects, but the software is still developed and maintained by myself in the time I can spare. This may sound like a rather risky way to proceed, with the software depending on a single person, but we have managed to get away with it and continue developing over a period of about 14 years; a time during which most bought retrieval packages (and the hardware they ran on) would have become obsolete several times over. That said, our current approach will clearly not be viable indefinitely.

The way in which the Cartoon Catalogue has developed obviously has advantages and disadvantages. Because the move to an online catalogue never depended on targeted funding (internal or external) we were able to just get on with it – solving technical problems when they arose and acquiring pieces of equipment when we could. In the process, we have created a substantial and useful resource that would not exist if we had had to proceed in a more formal way. That resource includes a great deal of cataloguing that would not have been done had it not been

in support of a live catalogue. The main drawback is that we have developed independently of standards that now exist but were not there when we needed them. If we were starting now from scratch then we would use a standard record format such as TEI (Text Encoding Enitiative) P4 (Sperberg-McQueen, 2002) and we would buy a comercially supported software package or use a data hosting service like the Visual Arts Data Service. As things stand, we will eventually have to convert our data into a standard form and deposit it with somewhere like VADS if it is to be available indefinitely.

References

Aitchison, J. and Gilchrist, A. (1987), *Thesaurus Construction*, 2nd Edition, Aslib, London.

Bovey, J.D. (1993) "A graphical retrieval system", *Journal of Information Science*, Vol. 19 No 3, pp 179-188.

Bovey, J.D. (1992), "Tools for preparing an online catalogue of cartoon drawings", *Program*, Vol. 26 No 1, pp 39-54.

Bovey, J.D. (1995), "Building a thesaurus for a collection of cartoon drawings", *Journal of Information Science*, Vol. 21 No 2, pp 115-122.

International Organisation for Standardisation ISO-2788 (1986), *Documentation – Guidelines for the establishment and development of monolingual thesauri*, ISO, Geneva

Z39.50 International Standard Maintenance Agency
URL: http://lcweb.loc.gov/z3950/agency/

Sperberg-McQueen, C.M. and Burnard, L. (eds.) (2002) Guidelines for Electronic Text Encoding and Interchange. Text Encoding Initiative Consortium. XML Version, University of Oxford, Oxford.

Purdy, P. (2001), "Digital Image Archiving and Advice: in Tandem with the Visual Arts Data Service (VADS)" *Cultivate Interactive*, issue 4,
URL: http://www.cultivate-int.org/issue4/vads/