Designing Effective Retrieval Systems for Digital Archives of Historical Documents

Andy White
Centre for Media Research, University of Ulster
Coleraine, Northern Ireland

Abstract

This paper uses two high profile digitisation projects to demonstrate the way in which effective retrieval strategies can be designed for digital resources. The main theme is the relationship between the accuracy of the natural language of the database and the effectiveness of the various search functions. It will be argued that successful retrieval strategies can only be based on ASCII text of an exceedingly high standard. To reach this standard requires rigorous proofreading and, as such, would appear to call into question the creation of databases comprising millions of words; the projects cited by the author each contain less than one million. If ICT is not a panacea for converting enormous amounts of original historical documents into easily retrievable digital archives, much smaller digital collections can yield results. The circular, as opposed to more traditional linear, media enables the design of content in a multi-layered fashion. Thus, material can be catalogued and tagged for metadata in an academic way but, with the aid of additional multimedia features, can provide different entries of access for people of varying abilities and interests. In this context, highly accurate retrieval systems using both controlled vocabularies and natural language can greatly aid researchers.

Introduction

With the aid of two case studies, this paper explores some of the major issues regarding the design of the architecture of digital archives and the way in which that design influences the interrogative strategies of users'. Crucial to the design of digital archives is the appreciation of the fundamental difference between the linear narrative of original historical documents and the non-linear form that database driven digital resources invariably take. As such, this paper will begin by considering what effect this form of non-linear media has on the presentation of historical documents. It must be stressed, however, that non-linearity does not equate to anarchy and therefore there will be an exploration of what kind of structures are most appropriate for this media. Only after those structures are in place, can attention be focused on the design of effective retrieval systems. There will be a particular emphasis on two issues: the accuracy of the ASCII text in representing the original sources; and the exploitation of multimedia both as a means of improving navigation of the digital archives and of improving knowledge and conceptualisation of the subject matter.

Non-linearity and the Importance of Hypertext

As Greenfield points out, the non-linear narrative has been around for thousands of years; this he illustrates by highlighting the layout of the *Talmud* with its central text bounded by annotations directing the reader to different sections.[1] But the invention of the printing press and then the novel essentially embedded the linear narrative in our literary culture.[2] Other forms of media, notably film and theatre, have experimented with non-linear narratives, but the printed form has remained basically linear in construction. The advent of CD-ROMs and the Internet has changed all that. CD-ROMs allow users' to navigate their way through the vast amounts of information held by relational databases. The Internet has accelerated this trend towards non-linearity, as users' are able to enter and exit websites at any point and time, and surf the seemingly limitless resources of the World Wide Web without leaving their seat.

Cultural theorists have been quick to assert that this trend has altered the relationship between reader and writer:

> They [digital environments] also seem to offer a privileged place to explore theorist Roland Barthes's valorization of "writerly" textuality, wherein the reader does not encounter a work whose meaning is fixed, but rather (re)writes the text through the process of reading. The "writerly" is opposed to the "readerly" qualities of classical fiction, wherein the art object is static and the hierarchy of creator and consumer is rigidly maintained.[3]

This form of interactive writing is commonly referred to as 'hypertext'.

While this new form of writing is to be welcomed for its capacity to enable users' to navigate their way through large amounts of 'linked' information, there are concerns over its impact on traditional pedagogy. Creators of digital resources are aware of the dangers inherent in users' accessing non-

linear information on subject areas in which they have no prior knowledge. They have attempted to overcome this potential problem by repetition of certain themes through the linking of small textual units, or, as Barthes referred to them, 'lexia'.[4] This leads to a paradox, in that the non-linearity of the digital text increases our knowledge but does not necessarily improve our conceptual understanding of historical events.[5]

The conundrum is to invent some kind of structure that will enable users' to exploit the flexibility of the non-linear narrative to gain knowledge within a discursive framework that also promotes conceptual understanding. Sharp, McKinney and Ross sees this conundrum being solved by the judicious use of multimedia:

> The emphasis would be placed on the use of graphics, sound, animation, spatiality and interaction to best reflect the ability of hypertext to connect ideas and create analogies, so corresponding with the patterns at play in the human mind. Hypertext could then be seen more clearly for what it is: a departure rather than a denial of language.[6]

Lunenfeld argues that this is best resolved through the use of the concept 'mise-en-abyme' – a series of mini-narratives that reflect the wider structure of the digital resource.[7] In other words, the small textual units representing knowledge should reflect the wider structures representing conceptualisation.

The need to create some sort of hierarchy within the non-linear format was exemplified by an exercise carried out by the convenors of an undergraduate module, 'Investigating Cyberspace', at HATII at the University of Glasgow. Students were encouraged to create a work of essentially non-hierarchical hyperfiction:

> When discussing the structure of their collaborative piece, the students continually found themselves returning to some form of hierarchy, which they were initially trying to avoid. With further consideration, they realised that a small element of hierarchy is forgivable and often inevitable, so long as that structure does not solely determine the course of the narrative.[8]

The Case Studies

To explore the influence that this theoretical debate has on the architecture of digital archives, the author has chosen two digitisation projects with which he had some practical involvement. The 'The Act of Union Virtual Library' (AUVLP) - of which the author was the Project Leader - was, under the aegis of the Library and Information Services Council (Northern Ireland), a collaborative project based at Queen's University Belfast between the latter, Linen Hall Library, Belfast Central Library, Public Record Office of Northern Ireland, Ulster Museum and the Union Theological College. The AUVLP (www.actofunion.ie, www. actofunion .ac.uk) was funded by the UK Government's £50 million New Opportunities Fund Digitise Programme, started in January 2002 and formally was launched in September 2003. It contains some 20,000 pages of digital versions of pamphlets, parliamentary papers, manuscripts and newspapers contemporary with the 1800 Act of Union between Britain and Ireland.

In March 1999, as Research Officer at Belfast's Linen Hall Library, the author began cataloguing the collection of political posters held in the library's Northern Ireland Political Collection. Later that year, the library received funding from the European Union, through its PROTEUS Programme, to create a CD-ROM of images relating to Northern Irish politics from 1966-2001. Launched in October 2001, the CD-ROM was placed second in the electronic category of the annual Chartered Institute of Library and Information Professionals/Nielson Book Data Awards; and in early 2003 the CD-ROM was the recipient of the 17th annual Christopher Ewart-Biggs Memorial Prize.

Troubled Images: A Visual Representation of Northern Irish Politics, 1966-2001

Selection

The Northern Ireland Political Collection of Belfast's Linen Hall Library contains a unique holding of primary source material from 1966 onwards associated with Northern Irish politics. Among this material are a substantial collection of political posters – in excess of three thousand – and a panoply of artefacts. The posters constituted a valuable academic resource, especially as a significant proportion of them were from illegal organisations that, by and large, did not have unmediated access to the mainstream media. Though the posters were arranged in a loosely thematic way, they were stored in very tightly packed drawers and, as such, the library staff restricted access to them. At the time, the posters were not only used by academic researchers, but also by broadcasters, such as BBC Northern Ireland, who displayed them as background to political interviews and documentaries. When the posters were accessed, that they were not catalogued meant that staff had potentially to sift through hundreds of posters before they located one appropriate to the research request. Over the long term, this type of practice was likely to lead to the physical deterioration of the posters.

The decision to first catalogue and then digitise the posters was therefore motivated by concerns about their ongoing preservation. A catalogue of the posters drastically would curtail the amount of sifting of the original collection, and digitisation would achieve an even more significant reduction. Apart from duplicates and posters that were clearly not related to Northern Irish politics, all the posters were catalogued on a 'talis' database, a basic library OPAC. Obviously, at this stage the selection were relatively straightforward. However, once funding was secured for the

digitisation of the posters, thoughts turned to construction of an appropriate content architecture.

Design Architecture

Initially, a Filemaker Pro database was established and populated with the bibliographic metadata. The posters were photographed and slides were produced from the negatives. Given the worries over the sustainability of digital resources over the long term, this was a particularly effective way of ensuring that there is a viable analogue alternative to the CD-ROM. The slides were scanned to a high resolution and the images were used to populate the Filemaker Pro database to complement the bibliographic metadata.

It was becoming clear by this stage that the technology of the CD-ROM should fully be exploited to deliver a resource that could be used in an educative way. This meant expanding the range of sources in order to encompass as much as the symbolism of Northern Irish politics as possible. For this, the creators mined the Northern Ireland Political Collection's display of artefacts, hitherto mainly used to impress visiting researchers'. Once these were photographed it was therefore possible to present in digital form many of the myriad flags and emblems associated with Northern Irish politics, as well as historically valuable sources, like the 'comms' that were used to smuggle messages in and out of the Maze Prison during the republican struggle against the prison regime in the 1970s and 1980s.

Methods of Retrieval

The non-linearity of the medium was utilised by the designers' to make a series of small textual units, in the form of annotations of around fifty words in length, to accompany the metadata of each image. It can even be argued that Lunenfeld's concept of the 'mise-en-abyme' was evident in the way in which these mini-narratives echoed the wider narrative of the overall project. This was particularly important because the annotations were an integral part of the full text search facility that was developed. In order to optimise search efficiency, it was important for the writers' of the annotations to include words that were reflective of the wider narrative. In other words, if a user searched for images on 'civil rights', a full text search with this term would only identify those posters that contained that precise term within its accompanying text. An awareness of this allowed the writers' to create a narrative for the whole resource.

As there is little scope for using hyperlinks, the CD-ROM format limits the extent to which hypertext can be utilised. But, as emphasised earlier by McKinney, Sharp and Ross, exploiting the gamut of multimedia features now in existence can largely make up for this limitation. 'Troubled Images' certainly utilised various modes of media to enhance interactivity. As well as standard retrieval by bibliographic elements and full text, it is possible to search for images with a graphical timeline and a map of the world. The notepad function, where images can be saved and annotations added, enables users' to write their own hypertext. There are also a number of short explanatory essays on key themes of Northern Irish politics and nine thematic interviews with academics and political activists.

Quality Assurance

It is imperative that the textual content of academic resources in digital format is almost wholly free of grammatical and spelling errors. The main advantage that a database-driven resource has over an analogue resource is that it can enable the user to search enormous amounts of texts for precise references. This search facility is undermined if there are large numbers of inaccuracies in the body text. With this in mind, 'Troubled Images' employed a professional proof-reader to complement the factual checking carried out by the two main researchers. The latter process was particularly important as there is a tendency for printed material to be rigorously proofread for grammatical errors, while the same level of diligence is not applied to checking basic historical facts. The material in 'Troubled Images', as the name suggests, addressed a number of extremely sensitive issues, including the deaths of over three thousand people in the recent Northern Irish conflict. Factual inaccuracies about the number of people killed in particular incidents, or the names of people involved in those incidents could have caused offence; anything that was potentially libellous was also expunged from the digital resource. The Northern Ireland Political Collection is renowned for the 'neutral' space that it offers researchers' in their study of a conflict that is both contemporary and in close geographical proximity to the library. The CD-ROM had to reflect that 'neutrality' and the researchers' scrupulously adhered to this; a small number of people representing the full spectrum of political opinion in Northern Ireland were shown the contents of the CD-ROM before project completion to ensure that this objectivity was achieved. It was noted that there was a disproportionate number of republican posters. Because of this we removed a few posters that duplicated particular campaigns or political initiatives from the final version. The reason we removed so few was that we realised that posters' were a form of media much patronised by republicans, as they found it difficult to access much of the mainstream media due to censorship laws and because some of it, like the main broadcasters', was either state (British) owned or regulated stringently by the state.

Act of Union Virtual Library: Primary Source Material Contemporary with the 1800 Act of Union between Britain and Ireland

Selection

In the late 1990s the Library and Information Services Council (Northern Ireland) committed itself to digitising parts of Ireland's archival heritage. Given that the bicentennial of the 1800 Act of Union (operative from 1 January 1801) was nearing, it was felt that a dedicated website of primary sources contemporary with the Act would be useful not only to scholars, but also to those whose interest was awakened by exposure to the many

commemorative events taking place at the time. Shortly after funding was secured, the selection and digitisation of material began at Queen's University Belfast.

After consultation with Irish historian Jonathan Bardon and the collaborating institutions, LISC(NI) decided to digitise four different sources: pamphlets, parliamentary papers, newspapers and letters written to and from some of the main protagonists in the debate leading up to the Act. In an era when political activists were not able to exploit the technologies that today's politicians take for granted and utilise so effectively, the pamphlet was as effective a medium for conveying one's political message as anything else. 'The Pamphlet War' of 1797-1800 generated over four hundred pamphlets, some of which were pseudonymous, arguing the cases for and against the union. Most of these were polemical, though some replicated parliamentary speeches. W.J. McCormack's 'The pamphlet debate on the Union between Great Britain and Ireland, 1797-1800' – the most useful list of the pro- and anti-union pamphlets held in institutions in Britain and Ireland, even if there are significant omissions – was used to identify the number of different versions of each pamphlet and their location in Belfast. The bibliographic information of pamphlets held in Belfast-based institutions was inputted into a spreadsheet until eventually there was an alphabetical list of titles of pamphlets. By cross-referencing with McCormack's book, it was possible to highlight the particular pamphlets that the project intended to digitise. Pamphlets held in Special Collections in Queen's University's library, the Irish Collection of the Linen Hall Library, the Belfast Ulster and Irish Studies section of Belfast Central Library and the Union Theological College were collected by project staff and taken to the Centre for Data Digitisation and Analysis at Queen's University and digitised with a high quality book page scanner.

There are a three types of parliamentary papers' located within the AUVLP website: British House of Commons and House of Lords debates; Irish House of Commons and House of Lords debates; and Irish statutes, including the Act of Union itself. Parliamentary debates during the period 1797 to 1804 that were deemed to have relevance to the digital archive were digitised; it should be noted, however, that it was not possible to locate Irish House of Commons and House of Lords papers' from July 1797 onwards (of course, from 1801 onwards there was no separate Irish parliament). This significant omission was partly compensated by digitising the annual volumes of Irish statutes from 1797-1800, thus giving users' an indication of the kind of legislation that was being passed in the dying days of Ireland's last parliament before partition in 1921.

It was felt that digitising a range of Irish newspapers contemporary with the Act of Union would augment the other sources. Three newspapers that reflected both pro-and anti-union sentiments were chosen: the Belfast Newsletter, Londonderry Journal and the Dublin Evening Post. We discovered at an early stage of the project that it was not possible to produce ASCII text versions of the newspapers without considerable cost in terms of time and money. Optical character recognition software could not properly identify the text because it was so small; such was the volume of work that keying in the text was not a viable alternative. Therefore, full runs of the newspapers are presented as images that can be enlarged and divided into columns for ease of viewing.

Finally, the images of manuscripts held at the Public Record Office of Northern Ireland of correspondence to and from very prominent politicians, like Castlereagh, were captured. Due to a number of staffing and resource difficulties this part of the project was not as successful as the others: the pages of the manuscripts are displayed haphazardly and without accompanying bibliographic details.

Design Architecture

A number of descriptive metadata elements, based on Dublin Core, were drawn up. The bibliographic elements and details of the work processes were held in an Access database. In addition, there were two main directories: one holding 400 dpi TIFF images of the material; the other holding full text versions of the pamphlets and parliamentary papers, separated by page (tagged, for example, <NEW PAGE> 2). The text was generated either by using optical character recognition software or, where this was not feasible, by rekeying. All this information was inputted into a MySQL database for delivery on the web via PHP. The database driven website was receptive to a range of search options, the results of which were displayed in a format broadly similar to that of an Internet search engine, with the search term highlighted in the event of a full text query. The target image is a JPEG image of the actual page. The size of the image files, both in terms of the amount of memory it holds and its physical presence on the screen, was determined by carrying out experiments with a 14" laptop with a Pentium II processor and 56K modem in order to balance download times with legibility. Before going live, a password-protected website was made available to a select number of academics and information scientists for feedback on any aspect of the design.

The AUVLP did not use the full range of multimedia employed by the 'Troubled Images' project, mainly because its purpose was to present an archive of fully retrievable primary source documents in an unmediated fashion. Its narrative structure was based instead on the four pillars of parliamentary papers, pamphlets, newspapers and manuscripts, all linked by the discursive thread of Dublin Core metadata. That said, some historical context was necessary, and this was achieved through introductory text on the sources used on the website, an essay on the Act of Union by historian Jonathan Bardon and a timeline of events. There are also hyperlinks to the partner institutions and similar projects.

Methods of Retrieval

Search retrieval is by basic bibliographic elements, namely author, title, date, publisher/printer, place of publication, and like 'Troubled Images', by full text; though

the newspapers and manuscripts do not have the latter facility. There are additional features, including browsing, a search by the institutions in which the original sources are held, and by the different types of material: parliamentary papers, pamphlets, newspapers and manuscripts.

Like 'Troubled Images', the AUVLP contains a number of controlled terms. While in the former project these were in the form of keywords, the latter was concerned with the way in which two-centuries-old words and place names are often spelt differently in the present. For instance, the word 'Huguenot' was spelt 'Hugonot' at the turn of the nineteenth century. It is, of course, possible to amend the ASCII text to reflect the modern variant of the word, but this is a questionable practice, not least because the creators then become interpreters rather than disseminators of the primary sources and because professional historians are likely to use the terms contemporary with the period of study. The controlled vocabulary for a number of important terms enables users' of differing abilities to identify the same sources using differing variants.

Quality Assurance

Unlike in the 'Troubled Images' project, the text in the AUVLP was replicated not generated and was of a much greater volume. Replicated text is likely to be more prone to error than generated text and so quality assurance is a much more time and resource intensive process, though the concern about objectivity does not apply with text that is merely a reproduction of the original. This is particularly important with text that has been replicated with the aid of optical character recognition software.

In its discussion of the use of optical character recognition software for the creation of digital archives, NINCH comments, in relation to the Victorian Women Writers Project, that:

> OCR [optical character recognition] is a good option for capturing the text, but since the project's audience is a scholarly one, careful proofreading will also be necessary to ensure that requirements for accuracy have been met.[9]

Similarly, the creators of the AUVLP took the view that the full text search should be taken literally and, accordingly, the ASCII text should be virtually error-free. OCR is extremely accurate when used on modern laser printed font.[10] However, the AUVLP comprises documentation that is two-centuries old with all its consequent imperfections, namely discoloration, unevenness of surface, fading and the 'bleed through' of text that was a common trait of printed material back then. This meant that around 40% of the documents prepared for full text searching were reproduced using rekeying rather than OCR. Even those documents on which OCR feasibly could be used could not be considered an accurate replication until rigorously proofread. Indeed, as Tanner has illustrated, even an OCR process that is 99% accurate is not necessarily acceptable for academic use:

> For example: a page of 500 words with 2,500 characters. If the OCR engine gives a result of 98% accuracy this equals 50 characters incorrect. However, looked at in word terms this could convert to 50 words incorrect (one character per word) and thus in word accuracy terms would equal 90% accuracy. If 25 words are inaccurate (2 characters on average per word) then this gives 95% in word accuracy terms. If 10 words were inaccurate (average of 5 characters per word) then the word accuracy matters more than the character accuracy – we can see the possibility of 5 times the effort to correct to 100% across the word accuracy range shown in this simple example.[11]

In light of this, proofreading the entire content was a must. Because of time and resource constraints, the Project Leader alone carried out this proofreading. This was not entirely satisfactory, as it was really only possible to skim the twenty thousand pages of content. For this reason, two types of 'fuzzy' searching, soundex and metaphone, were created: both search for words similar with a similar string or structure to the search query. Despite these caveats, tests carried out during the construction of the 'fuzzy' search facility identified a word error rate in the ASCII text of around 0.4 per cent or 1 in 250 words.

Conclusion

'Troubled Images' and the AUVLP were designed in ways in which reflected their different media. The CD-ROM and the lack of core textual material of the former meant that it had to fully exploit its multimedia features. The latter had a huge textual core and, although not losing sight of the need to add historical context, was more concerned about developing effective interrogative strategies for this large corpus of material. It was possible, through the use of hyperlinks, to direct users' to the institutions that held the original documents, as well as similar digitisation projects.

Apart from the searches by bibliographic elements, the methods of retrieval in both projects were different, reflective again of the different types of primary source material used. There are, however, three elements of the design of effective retrieval methods that are common to both projects. The first was the need to order the primary source material in a way that fully exploits the flexibility afforded by non-linear media while also ensuring that an alternative narrative is constructed to convey to the user not merely knowledge but also conceptualisation. 'Troubled Images' attempted this through the development of annotations as mini-narratives, while the AUVLP was clearly structured according to the four types of printed material. Second was the need to exploit the flexibility of multimedia and hypertext within this structure to deliver a resource that is not merely a replication of the printed material, but uses sophisticated search strategies to promote interactivity. And finally, without the input of a thorough quality assurance

procedure, the benefits bestowed by the carefully calibrated design of the digital resources would have been lost.

References

1. D. Greenfield, Contextual links and non-linear narrative: a virtual Rashomon, Museums and the Web, Pittsburgh, 2000.
2. Ibid.
3. P. Lunenfeld, Snap to Grid: A User's Guide to Digital Arts, Media and Cultures, The MIT Press, Cambridge, Massachusetts, 2001, p. 46.
4. Ibid., p. 52.
5. Ibid., pp. 52/53.
6. R. Sharp, P. McKinney and S. Ross. Visual Text: concrete poetry, hyperfiction and the future of the narrative form, HATII, Glasgow, July 2003, p. 7.
7. Lunenfeld, op. cit., p. 54.
8. Sharp et al., p. 12.
9. National Initiative for a Networked Cultural Heritage (NINCH), The NINCH guide to good practice in the in the digital representation and management of cultural heritage materials, HATII and NINCH, Glasgow and New York, 2003, p. 86.
10. S. Tanner, Deciding whether Optical Character Recognition is feasible, King's Digital Consultancy Services, London, 2004, p. 5 and A. Morrison, M. Popham and K. Wikander, Oxford Text Archive: Creating and Documenting Electronic Texts, Oxbow Books and the Arts and Humanities Data Service, Oxford, 2000, pp. 18/19.
11. S. Tanner, Deciding whether Optical Character Recognition is feasible, King's Digital Consultancy Services, London, 2004, p. 5.

Biography

Andy White is a Research Associate in the Centre for Media Research at Northern Ireland's University of Ulster. Since completing a doctorate in Politics at Queen's University Belfast in 2000, he has been involved in a number of digitisation initiatives, most notably Belfast's Linen Hall Library's 'Troubled Images' project and Library & Information Council (Northern Ireland)'s 'Act of Union Virtual Library'. His main research interests are digital preservation, the development of the Internet as a pedagogical tool and British and Irish constitutional politics.