# Designing E-Catalogues for Archival Institutions: Overcoming the Skepticism of Scholars

***Graham Jackson and Andy White***

## Abstract

*This paper will describe the way in which the Public Record Office of Northern Ireland, through its eCATNI project, will convert its hard copy calendars into a digital catalogue freely available on the Internet. The universal applicability of some of eCATNI's methodologies, particularly in relation to transcription and the measures taken to ensure that the e-catalogue has scholarly integrity, will be emphasized as a means of promoting best practice.*

## Introduction

Since 2002 the Public Record Office of Northern Ireland (PRONI) has been systematically converting its collection of printed calendars (catalogues) into an electronic format for eventual delivery on the Internet and on fully networked PCs in the organization's public search room. This paper will start with a review of the existing system of retrieving documents in the public search room, with a particular focus on the institutional, cultural, legislative and technological factors that have created the desire for this ambitious project to automate the system of searching PRONI's archives.

In the main body of the paper, two main challenges will be highlighted, namely: the difficulty of achieving high standards of accuracy both in relation to transcription and in the correction of textual errors that occur in the original hardbound calendars; and the development of a sophisticated and user-friendly retrieval engine for such a large amount of information. The way in which the eCATNI (Electronic Catalogue for Northern Ireland) project has approached these challenges hopefully will be useful to other archival institutions looking to create their own electronic catalogues.

## A history of cataloguing at PRONI

The Public Record Office of Northern Ireland is a multi-disciplinary organization with over 100 curatorial, conservation, technical and support staff, whose main responsibility, under the Public Records Act 1923, involves identifying and preserving Northern Ireland's unique archival heritage and ensuring access to that heritage (especially in the current context of recent legislation such as the Freedom of Information Act 2000).

Records held fall into three broad categories: those originating from the Northern Ireland Government Departments (1921 to the present day), those relating to the various functional responsibilities of the courts of law, local authorities, etc., and finally, privately deposited archives (which constitute around 40% of PRONI's total record holdings and comprise the various records of business, landed estates, church and genealogical papers).

Under the existing system, the descriptive lists for these records are currently accessible by the public as paper catalogues, numbering around 450 volumes (estimated to total about 170,000 A4 pages). While these catalogues are of a very high quality (in terms of serving their purpose to accurately summarize the scope and content of a unit of description), they do reflect over eighty years' of fluctuating cataloguing practice and stylistic approach and as such display a pronounced lack of consistency.

A major consequence of this is that the information is often stored in a range of outdated formats and computer systems. Pre-computer, manually typed lists sit on the shelves of the Public Search Room alongside those lists originally created on 'AmiPro' and 'Word Perfect' programs (this problem is often exacerbated by the necessity to support multiple versions of any single catalogue entry, especially in the case of lists where, in order to properly comply with legislation such as the Data Protection Act 1998, sensitive information may need to be omitted from a publicly accessible version).

This largely paper based system is in turn supported by a relatively outdated and deficient network of document search, order and location management tools (originally intended to act as a key operational system supporting a range of archival management processes) and a number of searchable indexes which were previously held on obsolete (System 4 MSDOS) systems.

It is clear, therefore, that this existing approach is fragmentary, inconsistent and inaccurate, especially in regard to PRONI's legislative obligations under the terms of FOI and the consequent necessity to find information on any given record quickly and efficiently. In effect, PRONI is, at present, unable to communicate exactly what records it holds and, in the current climate of openness, accountability and accessibility, this is insupportable.

The eCATNI project is a three year undertaking by PRONI, charged with the task of successfully migrating this existing catalogue information into a searchable database and, ultimately, making available to the public Northern Ireland's rich and unique archival heritage, using Internet technology to overcome the limitations of geography and promote accessibility.

Beyond the need to properly comply with legislative obligations, there are a number of other obvious benefits to be realized through the design and launch of such an archival project.

Addressing the issue of consistency and style in archival description, one central driver of this project will be to impose a degree of uniformity on catalogue description, in line with those guidelines provided by the General International Standard of Archival Description or ISAD (G).

Furthermore, as the 'archival memory' for a country plagued by the legacy of many decades of politically motivated violence and civil unrest (the last three decades of which are known to most only as 'The Troubles'), PRONI staff are in a unique position to act as mediators of the archival resources it holds, and as such have a responsibility to support cultural diversity and social inclusion.

It is also expected that there will be the potential for digital images of archives to be made available to customers and,

eventually, to allow for the provision of a link to some form of common portal, such as 'archives UK' (aUK), a collaborative initiative across UK archives that would create a common on-line entry point to the sources in various institutions, thus enabling the user to gain access to a myriad of national and international archive, library and museum resources.

## The perils of transcription: doing justice to the original

Large-scale digitization projects generate such an enormous amount of text that quality assurance is extremely problematic. There are essentially two options for migrating paper-based text to a digital platform: manual re-keying or through the use of optical character recognition software (OCR). The first option can be extremely time-consuming, financially costly and prone to human error. The second option is not without its drawbacks too, as OCR software is less than one hundred per cent accurate on anything less than pristine, laser-printed, unblemished text. The approach of many projects that use OCR software is to embed a "fuzzy" logic searching function into their databases, essentially allowing users to set the parameters of their search by, in a typical example, omitting one or two characters in a word so that misspelled words are easily located. However, establishing equilibrium between coverage and precision is not easily achieved. Sophisticated "fuzzy" logic searches that enable users to locate badly misspelled words (perhaps where two or three characters are missing or incorrect) will retrieve large numbers of irrelevant terms. As Lanham and Harrison have argued, the attention span of the average Internet user is so short that he/she will not tolerate sifting through large amounts of redundant information:

> whenever we "persuade" someone, we do so by getting that person to "look at things from our point of view," share our attention structure. It is in the nature of human life that attention should be in short supply, but in an information economy it becomes the crucial scarce commodity. Just as economics has been the study of how we allocate scarce resources in a goods economy, we now will use a variety of rhetoric as the "economics" of human attention-structure ... a vital activity in our information society[1].

Further, "fuzzy" logic searching cannot guarantee to identify every single word. Leaving aside the fact that calibrating the search to retrieve words with one or two characters omitted or incorrect means that the user has precisely to locate the part(s) of the word that are most likely to contain inaccuracies, much more serious errors, like the omission of whole words, cannot be rectified by even the most sophisticated retrieval system.

If OCR'ing does not achieve adequate textual accuracy, what about the alternative of re-keying the text? Though it is undoubtedly more accurate, it is also more expensive – Chapman arguing that it can cost up to ten times as much as OCRing[2]. He has proposed a hybrid solution:

> If near 100 per cent accuracy of searching is required, it might be less expensive to key than to undertake the three-step process of scan, OCR, and correct OCR errors. Several reliable studies report that a trained technician can correct 6-10 pages per hour. Depending upon salary, this task alone could easily exceed the cost of keying[3].

Whatever approach is taken it is difficult to achieve one hundred per cent accuracy. The Library of Congress's National Digital Library Program stipulates that its textual error rate must not exceed 0.5%[4]. This essentially is a character error rate of 1 in 200 and, if we assume that the average length of a word is around 6 characters, a word error rate of less than one in thirty. Expressed in these terms that is a relatively high error rate. Later, this paper will look at the approach of the eCATNI project to textual reproduction, including an estimation of its accuracy.

## Designing functional retrieval systems for digital archives

Retrieval system design is largely based on factors enumerated in the last section, namely the accuracy of the digitized text. But there are other factors of a more intellectual or schematic nature that must be considered. Designers must decide what types of metadata can be searched. While most of these categories are relatively straightforward, provision sometimes need to be made for absent and/or incomplete information, varying consistency of information and the incorporation of a controlled vocabulary. The identification of types of metadata for searching is relatively straightforward; for archives this can simply be the main bibliographic elements: authorship/provenance; title; date; description; reference number; access decision. Within this, it may be feasible to introduce a search that takes into account the fact that some dates are expressed in the form of a range rather than a precise time. In order to maximize the functionality of the search engine inconsistencies in the original paper-based catalogues need to be ironed out and omitted elements should, wherever possible, be kept to a minimum.

Decisions over whether to incorporate controlled vocabularies in retrieval systems are essentially intellectual in nature. In the archival environment, this option is considered for two reasons: where contemporary words have mutated from their antecedents; and where historical terms are contested. In another digitization project of Irish historical documents, the Act of Union Virtual Library (www.actofunion.ac.uk), the designers created a controlled vocabulary for terms – like 'Huguenot' (spelt 'Hugonot' two centuries ago) – whose contemporary version is different from that of its eighteenth and nineteenth century antecedent[5]. Also, Northern Ireland's troubled history is reflected in Irish nationalists and Ulster unionists having their own nomenclature for certain geopolitical terms, the most notable example being contestation over the name of Northern Ireland's second city – 'Derry' for nationalists, 'Londonderry' for unionists. In all these examples it is much better to create controlled vocabularies rather than second-guess users or, worse still, replace historical terms with their modern-day equivalents.

## Who makes these decisions?

As has been illustrated the creation of e-catalogues involves numerous intellectual, if not political, decisions. While there are myriad technical guides on metadata, digital preservation and security, the literature on the intellectual issues that designers face

is rather sparse. Rhyne has expressed concern that scholars have little or no involvement in what are essentially academic resources[6]. Lynch too has raised similar concerns:

> As a case in point, with the availability of substantial number of digital images from museums, we are seeing universities employing these collections both for teaching and research. But we don't seem to be seeing the dialogue I would have expected between the scholars in the university who study this material and teach it, and the people in the museums who curate and exhibit it. (I do recognize there are some long-standing cultural divides here.) I don't think we have seen much change in the shape and practices of the scholarly literature that uses and interprets these images, or the teaching materials that build on them[7].

The archival staff at PRONI possess an array of skills, historical knowledge and accrued expertise in various fields. Gaining the support for the eCATNI project of PRONI's researchers, many of whom are established scholars, is probably largely dependent on the input of these archivists.

## How has the eCATNI project dealt with these challenges?

During the data capture and quality assurance stages of the project, a number of measures were introduced in order to address the legacy of inherited stylistic difference in cataloguing practice. As previously noted, the sheer variability and inconsistency of the different types of catalogue generated a diverse array of challenges.

During the lifespan of the project (which is still some way from completion), a small team of experienced archivists has worked in close cooperation with a 'project dedicated' unit from PRONI's Information Systems section. It was (and still is) a critical factor in the ongoing and successful progression of the project that core archival decisions be left in the hands of the archivists and that the IT unit cater (where at all feasible) to these requirements. A number of core specifications to the present incarnation have arisen directly from this hand-in-glove relationship between the two disciplines. Of course, certain demands that, from an archival perspective may seem logical and reasonable, are often unrealistic from the perspective of the pragmatic IT specialist. In such instances, there is a necessary compromise. Nevertheless, the fact that this symbiotic relationship has so far proved successful is evidenced by the amount of progress made up to this point and in the way that certain seemingly insurmountable obstacles have been overcome.

A pertinent example of this is the repeated occurrence in PRONI catalogues of 'bundles' of records (where an unspecified number of documents had previously been treated as a single unit or bundle). This practice was no longer deemed acceptable under the new system, as it was imperative that, as far as possible, all documents be recorded at 'Item' level. Therefore, in order to accommodate such instances, 'Dummy' and 'Clone' records were generated. The former denotes the use of a blank Item level entry, identifying it for future cataloguing priority, whilst the latter denotes that part of general description which has been duplicated to indicate the item's place in a greater whole or bundle.

One major portion of 'private' archives comprises those records transcribed from original documents (whose catalogues were previously electronically captured in a separate project in 1998). Unfortunately, these data remained unused and 'dirty' until the current project's inception and consequently demanded a considerable degree of clean-up. Primarily, certain specific fields were prioritized (such as *Title*, *Date* and *Reference Number*) and the data rectified accordingly, through a process of quality assurance and manual amendment of any errors found (this also included the addition of titles to many archives, which had previously been omitted).

A substantial percentage of the core descriptive data remained flawed. One crucial reason for this lay with the type of operating system employed by overseas keying staff in 1998. Standard ASCII code values terminate at 127, but the ASCII designation for the £ sign happened to be 163 (values over 127 being determined by the individual operating system). In the case of the system used at the time, there was no £ sign available to keying staff, resulting in all instances of £ being supplanted with the term *XX*. The consequence of this unforeseen oversight is that, although not fatal in terms of any search capability, it was necessary to allocate important resources in order to properly clean this particular set of data.

Another common problem has been the concern over how to deal with errors, colloquialism or localized phraseology inherent in many documents, especially prevalent in transcribed letters or diaries (that is, where the responsibility for the error lies with the original author).

To alter the misspelled word or place name is indeed, from an archival standpoint, incorrect, as this can be construed as compromising the integrity of an original document. Yet the resulting consequences for any search engine are obvious.

This is further complicated by the variation in spelling of personal and place names, especially the 'Townland' (a uniquely Irish term denoting a small division of land), where a single place name may have a number of versions. This inconsistency is largely due to the difficulties of representing the pronunciation of Irish language names in English spelling. Hence, the area of *Tanderagee* can be found in some records to be spelt as *Tandragee*, and, if either name is searched for, two quite different sets of results can be expected.

The same can be said of personal names, where the surname '*Wyman*' may arguably be derived from the surname '*Weyman*,' but a search will unfortunately not seek out both versions of the name.

With this in mind, it is anticipated that the initial incarnation of the project search facility will encourage the researcher to input several variations of a place name or surname to expedite any comprehensive search. Thus, in reference to the aforementioned and politically contentious case of the City of Londonderry / Derry, any keyword search for the word '*Derry*' will not locate the words '*Londonderry*,' '*L/Derry*' or '*L'Derry*' and will require that the researcher qualifies the search by inputting two or more versions.

Similarly, project staff encountered difficulties when attempting to accommodate the large number of date field variations associated with differing types of records. For example,

in the case of official court records, it was not uncommon for *Equity Civil Bill* papers to adopt terms of 'sitting,' such as *Trinity*, *Hilary* or *Michaelmas,* to represent the date on which a court heard cases. Equally, in private archives, it was quite usual for an individual, in the course of a large volume of correspondence, to use only the time or day of writing, such as *Friday, midnight,* in place of a specific calendar date. At the time of writing, the project has verified upwards of 233 separate, valid date formats.

Initial quality assurance of returned data has been conducted by comparing the original marked-up text with the new electronic version. In the initial specification of requirements with the overseas data capture contractor, PRONI insisted upon a level of no less than 99.95% accuracy. As the project has progressed, the relationship with the contractor has generated confidence at both ends, in terms of the capabilities of each. The actual exercise of validation has in itself generated a plethora of unexpected problems, each of which has spawned its own (considerable) workload. Records that possessed no logical 'parent' or 'child' (that is, no ISAD (G) level directly above or below it) needed parents or children created retrospectively. Records with unacceptable date formats, acronyms (a curse common in the domain of official records), abbreviations or the use of such informalities as 'ditto,' were manually amended accordingly.

Unit description length presented a unique dilemma (one whose solution was of particular significance and, in terms of supporting future search capability, potentially damaging) to the success of the data capture exercise. Early project software sustained only 64,000 characters, which precluded many of the larger document transcriptions (a single transcribed diary description can easily reach over 100,000 characters) and jeopardized any intended search functionality. However, to counteract this, specific software was purchased as a 'bolt on' to the existing catalogue package. This now permits unit description length of up to 480,000 characters (more than enough to cover even the lengthiest of diary entries!). As previously mentioned, the combined issues of 'accountability' and 'access' feature prominently on the list of mandatory criteria for project functionality. This entails that, in the case of compliance with Freedom of Information legislation, records are to be easily and rapidly accessed and their existence communicated to the inquirer within a set time period. Conversely, under the terms of compliance with Data Protection Act legislation, records bearing any data that could be construed as information of a personal or sensitive nature and relating to a living person may not be displayed on a public facing database. These issues have further complicated the work of the electronic cataloguer, inasmuch as every individual catalogue entry may have the potential to breach a particular set of legislation. This has resulted in a dual faced catalogue (and search facility), with one side supporting a restricted version, accessible only to PRONI staff, whilst the other one constitutes a public facing, fully searchable version.

## Conclusion

The eCATNI project visibly demonstrates the intellectual and technical challenges that all digitization projects face. Despite the rapid technological developments in this field over the last decade the stage has yet to be reached where the conversion of hard copy catalogues into machine readable text is completely automated. Even if perfect automation could be achieved in technical matters, the intellectual challenges cannot similarly be overcome. It is therefore imperative that archivists and/or academics are integrally involved in these projects as a means of ensuring that a digital resource has scholarly integrity, both for its intrinsic worth and to ensure that it commands the respect of all potential researchers. The eCATNI project generated a huge number of intellectual challenges, ranging from differences in nomenclature to the omission of useful information, but by employing archivists as well as technical staff it has given itself every chance of overcoming them.

## References


[1] R.A. Lanham, The electronic word: democracy, technology, and the arts. (The University of Chicago Press, Chicago and London, 1993), cited in C. Harrison, "Hypertext Links: whither thou goest and why". First Monday, 7, 10 (October 2002), available at http://www.firstmonday.org/issues/issue7_10/harrison/

[2] S. Chapman, "Working with printed text and manuscripts", in M.K. Sitts (ed.), Handbook for Digital Projects: A Management Tool for Preservation and Access (North East Document Conservation Center, Andover, MA, 2000), pg. 114, available at http://www.nedcc.org/digital/dman.pdf.

[3] *Ibid.*

[4] National Initiative for a Networked Cultural Heritage, "The NINCH guide to good practice in the digital representation and management of cultural heritage materials" (HATII and NINCH, Glasgow, New York, 2003), available at http://www.nyu.edu/its/humanities/ninchguide/.

[5] A. White, Designing Effective Retrieval Systems for Digital Archives of Historical Documents, Proc. IS&T's Washington Conference, pg. 41 (April 2005).

[6] S. Chapman, "Scholar community: an end-user speaks up", in M.K. Sitts (ed.), Handbook for Digital Projects: A Management Tool for Preservation and Access (North East Document Conservation Center, Andover, MA, 2000), pg. 178, available at http://www.nedcc.org/digital/dman.pdf.

[7] C. Lynch, "Digital collections, digital libraries and the digitization of cultural heritage information". First Monday, 7, 5 (May 2002), available at http://www.firstmonday.org/issues/issue7_5/lynch/index.html.


## Author Biography


*Graham Jackson is employed as an archivist at the Public Record Office of Northern Ireland and is a project team leader on eCATNI. Andy White is a Research Associate in the Centre for Media Research at Northern Ireland's University of Ulster and formerly an archivist at the Public Record Office of Northern Ireland.*