**Newspaper digitisation pilot project**

Final Report of Waterford City Library newspaper digitisation project, a case for further newspaper digitisation in Irish Public Libraries and recommendations and costings for a national newspaper digitisation scheme.

Organisation: Waterford City Library
Date:03/03/03
Author: Emer Fitzgerald

# Contents

# Introduction

This report will evaluate Waterford City Library's newspaper digitisation and automated indexing pilot project and extract a number of conclusions and recommendation based on the evaluation.

A case for newspaper digitisation for the Irish public library sector is established based on the statistics extracted from the library newspaper holdings survey in Stage 4.

A set of recommendations for a National Newspaper Digitisation Scheme is submitted.

Finally, a set of costings for a National Newspaper Digitisation Scheme is outlined based on the finds of the pilot project.

# Chapter 1 – Waterford City Library newspaper digitisation and automated indexing pilot project –Evaluation, Conclusions and Recommendations

## Introduction

In September 2002, Waterford City Library started a 6 month project to investigate how to make historical newspapers full-text searchable via the Cultural & Heritage portal site, http://www.askaboutireland.com/pilots/three/index_main.html

Waterford City Library's newspaper digitisation project is now complete. The project team is proud to conclude that the pilot project has successfully reached its goals of:

- Researching and learning how to create a full-text searchable online newspaper archive
- Making a small selection of Waterford City Library' s newspaper collection full-text searchable via the CHP portal
- Producing guidelines and reports on the digitisation process and related issues. Click on the link above to search Waterford City Library's Newspaper digitisation project

The project has been a fantastic learning experience for everyone involved. The project team is excited to have worked on a project, which is progressive and groundbreaking. Though the project presented many challenges to the project team, it is encouraging for future project teams to know that all problems were understood and resolved where a resolution was possible. Sometimes answers were found through research and other times solutions developed through trial and error.

This report is the final report of the Newspaper Digitisation Project. The aim of this report is to include an outline of the previous stage reports, evaluate the project by highlighting the main challenges of the project and formulate a list of recommendations for future newspaper digitisation projects. A list of reference material consulted by the project team during the project will also be included.

## Summary of Stage reports

### *Stage 1*

The objective of Stage 1 was to carry out Internet research into the scanners and OCR software available. This report outlines how to conduct successful internet research and provides details of the scanners and software found through the research.

***Stage 2***

Stage 2 of the project required the project team to select the newspapers and/or microfilm to test the software and the scanners researched in Stage 1.The report outlines guidelines for the selection of material to be digitised and gives the results of the software and hardware tested.

***Stage 3***

In Stage 3 of the historical newspaper digitisation project the test results in stage 2 were considered and the best performing equipment was used, by the project team, for the actual scanning, OCRing and automated indexing of the selected material. By the end of this stage the selected material is 'keyword' searchable via the CHP portal site

***Stage 4&5***

This report provides an overview of newspaper digitisation on a national scale. It provides an audit of the number, format and variety of newspaper collections held by Public Libraries in Ireland.

This report also attempts to assess the level of experience, knowledge and interest Public Library staff have in relation to digitisation.

Finally, the report outlines the copyright issues involved in newspaper digitisation, the costs of in-house and outsourced digitisation and a breakdown of the timescales involved in the digitisation process.

***Stage 6***

Stage 6 of the Waterford City Library's Newspaper digitisation project investigates current electronically –produced newspaper files and researches access to these files.

***Stage 7***

This report deals firstly, with the short-term or day to day preservation of the digital image. Secondly, the report deals with the long-term preservation of the digital image, including the development of a preservation strategy from a technical and organisational perspective. Thirdly, the report outlines the various hints and tips for long-term digital image preservation. Finally, there is a list of relevant references.

***Stage 8***

The aim of the Stage 8 report was to:

- Establish a list of the training requirements for the pilot project

- Establish a list of the skills the project team had going into the pilot project
- Establish an amended list of the training requirements on completion of the pilot project necessary to successfully complete a national or local authority newspaper digitisation project.

## Evaluation

Waterford City Library's newspaper digitisation and automated indexing pilot project has been a success and all the goals of the project have been met. The project team has successfully created a full-text searchable online newspaper archive, on the CHP portal site, from a selection of Waterford City Library's newspaper collection.

There are a number of observations the project team made during the project, which had not been foreseen prior to commencement of the project and are important to highlight at this stage.

### *Microfilm could not be used*

It had been presumed that the project team would be digitising from microfilm rather then the hardcopy originals because:

- Microfilming is seen as the best practice preservation method for newspapers[1]
- If properly microfilmed and stored the microfilm will be of a good quality
- Microfilming was normally done a long time ago when the original newspapers would have been in better condition then they are in now. Therefore the microfilm image will be clearer then the original in its present state.

According to the results of the Stage 2 report, in order for OCR software to work to a workable level of accuracy the quality of the imaging or original being scanned has to be good. Digiscan and Olive Software (digitisation bureaus) both concluded that the quality of the microfilm holdings of Waterford City Library was insufficient to

[1] Most other libraries hold the view expressed at the 1995 RLG digital selection symposium, that "digitisation appears to have a preservation role in reducing use of originals ... but appears not to be suitable for preservation of information that is preserved in no other form." The consensus among most American preservation officers is that digitisation alone does not constitute preservation:

- digitisation is access-driven,
- it is not a replacement for microfilming or other preservation activities,
- selection for digitisation does not arise from the same motivation as selection for preservation,
- the original object or analog copy serves as the long-lived preservation medium, while the digital version serves extremely well as the access medium,
- special collections and visual materials take priority for digitisation because the technology suits them and they suit the technology.

To cite several examples, Harvard uses digital versions for access, and analog media for long-term storage; Columbia and Chapel Hill apply scanning for access to primary sources which are retained afterwards and for access to high-use course-related materials; Indiana uses digitisation to print out acid-free copies instead of photocopying; and Yale's Project Open Book created digital versions from already-existing preservation microfilm. Joint RLG and NPOPreservation Conference Guidelines for Digital Imaging . [WWW] http://www.rlg.org/preserv/joint/gertz.html

produce effective OCR accuracy results. The advice given was that the originals would have to be used in order to get a clearer, sharper image.

Working from the originals is not ideal because it is slower and may include unbinding and rebinding of bound volumes, therefore, overall it is less cost-effective. However, due to a lack of a workable alternative, the project team proceeded to digitise from the original newspapers.

### *Quality of the original newspapers*

Generally speaking, apart from Olive Software[2], OCR software is designed to recognise laser-printed text. This text is sharp, clear and each character is easy for the software to recognise.

Newspapers present very different print text characters including:

- Blotching
- Dirt
- Fading
- Through Bleed
- Holes
- Tears
- Irregular font sizes
- Columns
- Text density
- Missing text
- Old language
- Irregular text types
- Variety in layout from one ear to another

Because of all these issues, OCR software will not meet the level of advertised accuracy it claims to be able to achieve with modern day office generated text. In order to make OCR worthwhile and avoid a situation where so much editing is involved in the OCR stage that the project team is effectively re-keying the text, the quality of the scanned image must be very good. In turn, in order to create a very good quality scanned image, the original (or microfilm) must be very good. The project team found that they had to divide Waterford City library's newspaper collection into Quality Batches of:

- Good
- Medium
- Bad.

Bad quality newspapers were immediately dismissed as being of insufficient quality. Medium quality newspapers were tested to see what level of OCR accuracy was being achieved. This often varied greatly from page to page. Generally, good quality

[2] Olive Software is specifically designed to OCR historical newspapers

newspapers were the ideal as the medium quality still needed a lot of post-OCR editing. Luckily a lot of the collection was of sufficiently good quality.

### *Image creation*

Creation of the scanned/digitised image from the original newspaper was relatively straightforward. Advice from TASI, Digiscan and other sources gave the same advice to scan at 300-400dpi to create a TIFF master copy. The only variance in advice came with regard to using greyscale or black and white. The team used greyscale having tested the OCR success of both greyscale and black and white.

In order to know what image quality to achieve the project team needs to be clear of the potential use of the image. In this case, the image quality had to be good enough to help the OCR software achieve as high a level of character recognition accuracy as possible.

### *OCR software*

The project team was surprised at the amount of OCR software available and the variety of software available. OCR software ranged from approx. 60-500 Euro and functionalities and complexities also varied greatly. Having tested numerous types of OCR software the project team realised that some were not suitable at all, mainly because they could not cope with the column layout of newspapers.

Firstly, the project team concluded that any OCR software used to digitise newspapers and especially old newspapers would need good 'zoning' functions to cope with the column lay-out by dividing the page into it's various units i.e. articles, advertisements, etc. Secondly, the software would need good editing and proofing functions, because even with the best quality images, the quality of historical newspapers is inherently poorer then what the software is designed to deal with.

### *Post –OCR*

The OCR software had successfully produced editable text files of the images, but the project team was uncertain of how to make this text searchable via the CHP portal site. The project team decided to rely on the automated indexing facility available on the CHP portal search engine. Once the project team had decided this, the project team had make the output files web compatible by converting the text to HTML so that the CHP portal sites search engine index could read the text.

For a while, the project team wondered if the OCR stage was necessary at all but then realised that to create the HTML text to enable the CHP portal search engine to find an image there were 3 possibilities:

- Manually keyword index the image in the appropriate metadata field .
- Rekey the whole text into HTML.

- Or ‘cut’ and ‘paste’ the whole text from the editable OCR output file to either the metadata of the text or a separate webpage.

The first option is limited and does not create full-text searchability. The second option is far too time consuming. Therefore, the third option was chosen for speed and functionality. The images on the web would be jpeg file types so that they are small enough to download quickly but the legibility at jpeg resolution can be low. It was decided that the text and the image should both be available to the user, and separate webpages were built for each image. The user can clearly read the HTML text but can also see the image of the article or the page in order to get the context and historical ambience of the source.

### *Metadata*

The purpose of Metadata only became clear towards the end of the project. However, further reading helped clarify that there are two kinds of metadata. The first type of metadata is a searchable record of keywords that can help retrieve the record and image, i.e. subjects, location, names, placenames etc. This was not as essential to the newspaper digitisation project, as we had made the images themselves searchable but it could be used to add context and thesauri to the image. The second use of metadata is to record technical data so that the original image can be traced or located, i.e. disc no. file name, file size etc.

### *Skill –level of staff*

By the end of the project, the project team was surprised at the level and range of skill required for a digitisation project. The project team highlighted good I.T. skills as the most important skill. Every stage of the process involves using I.T. from scanning to webpage design. Therefore, for troubleshooting and problem-solving reasons, a digitisation team would need a very good understanding of the I.T. implications of the project e.g. File size, naming, file types and their uses etc

### *Proofing and Editing*

Despite the fact that Waterford City Library’s newspaper project was successful, the project team was surprised at the amount of proofing and editing (manual intervention) needed. The main reason for this is that despite the advances in OCR technology, some manual editing is needed to compensate for the inherent poor quality of historical newspapers.

# Recommendations

## 1. Conduct a pilot project or feasibility study

A full-scale digitisation project is a costly and time consuming undertaking. Experience of the Waterford City Library pilot project has shown that each project can be unique in the obstacles a particular collection may present. Therefore, Waterford City Library pilot project would concur with the National Library of New Zealand in recommending a pilot project because, "digitisation is all about choices, picking the most suitable approach and goal for the material, budget and institution in question."

TASI states that at the pilot project stage if the answer to the following questions are found to be in any way wanting, serious thought must be put into the viability of the project. It might be concluded that the project should be postponed until it can be assured of better support and results:

- Is there funding and time available to complete the project?
- Are all the works available and are there any conservation issues that might affect their capture?
- Will the proposed infrastructure be able to support the digitisation and deliver the end product?
- Can the project have the full support of the institute or owners of the collection as well as from the staff and user-base?
- Does the risk assessment show any other likely problems and difficulties for the project that have not or cannot be addressed?
- Can the institute provide ongoing support for the image resources into the future, guaranteeing its long-term viability?

## 2. Define the purpose of scanned materials

As well as testing the hardware and software using the relevant source material, the project team need to decide from the beginning the **purpose or intended use of the scans**. In this instance, the images had to be of sufficient quality to allow OCR software to successfully recognise the text. If this cannot be achieved with reasonable manual editing then the project team may have to rethink the projects goals.

## 3. Plan the project holistically

One can also look at the questions outlined above when considering the project plan. It is important to **plan the project holistically**. By this, a project manager should not just consider the technical problems and the desired result but must weave sideline considerations such as copyright, metadata, tracking, and preservation, into the plan from the beginning. All these things take time and cost money if staff time has to be used, therefore, it is important when preparing the

projects budget and time plan to include every aspect of the project from the beginning, building as many aspects into the workflow lifecycle of the project.

**4. Selection of Material**

The **selection of material** to be digitised will probably become evident from the pilot project, however, it is worth looking at in isolation. Once a project is considered feasible, selection is the next step because digitisation of whole collections are rare due to lack of digitisation potential of the collection and/or lack of sufficient time or money. One of the most important and decisive selection criteria for newspapers digitisation is to establish whether to digitise from the original material or surrogate. This selection question will outline the course for the rest of the project e.g. Necessary equipment, potential results etc. The decision the manager has to make is based on the fact that scanning from microfilm is cheaper, quicker and meets preservation criteria, however, all too often scanning from originals is the only way you produce OCR friendly images because microfilm quality is often too poor but it is more expensive and time-consuming and if the original doesn't have an existing microfilm copy, long-term preservation requirements have not been met.

**5. Costs**

**Watch out for unforeseen costs** – in the case of Waterford City Library's newspaper digitisation project the microfilm is of insufficient quality to use for OCR purposes. If this is the case and the originals have to be used, there may be an additional cost of unbinding bound newspaper volumes.

**6. Copyright**

**Clarify any copyright issues** from the outset. It is too late at the end of a costly digitisation project to find out that the newspaper publisher objects to the use of the papers for digitisation.

**7. File tracking system**

**Set up a good tracking system** from the beginning of the project to track file creation at each stage of the process. A good tracking system will show what stage of the digitisation process any image is at, at any given time. This tracking system will include information such as: file type, file name, date created, creator's name, storage information.

**8. Preservation Policy**

Similar to tracking, the project manager should sort out a **digital image preservation policy** from the outset, so that from the beginning files are being created to a consistent specification or file naming procedure.

**9. Upskilling Staff**

Allow enough time in the project plan for the project team to become completely **familiar with the necessary equipment and software**. This will help the process proceed quicker and with less problems when the real digitisation begins.

**10. Outsourcing**

A project manager embarking on a newspaper digitisation project should be familiar with outsourcing options available. A scanning or digitisation bureau may be able to deal with some or all of the digitisation process more efficiently than doing it in-house.

**11. Consider alternatives**

If the feasibility study of a library shows that a library's collection is of insufficient quality to produce the desired results, then alternatives such as re-microfilming or sourcing better quality film should be investigated.

**12. Watch this Space**

OCR technology is improving all the time and is increasingly able to deal with the characteristics of newspapers. Also, like all new technology, the hardware necessary for the digitisation process will become cheaper. Therefore, even if digitisation is not feasible now, a project manager should keep afoot of technological advances in the future.

**13. Don't reinvent the wheel**

There are fantastic free resources available which give reliable advise on all aspects of digitisation e.g. TASI, PADI, ADHS. There are also similar projects in other countries, like those, referred to throughout this paper and soon the guidelines paper produced as a results of Waterford City Library's newspaper digitisation pilot project, will be published by the Library Council.

Therefore, whilst one is carrying out the feasibility study, refer to these resources to highlight issues of important and highlight possible problems and solutions and most of all to reassure the project manager that every aspect of the project has been considered. One should not rely completely on these resources as each project is different and the materials and collections have individual and particular characteristics.

# Chapter 2 – A case for newspaper digitisation in Irish Public Libraries

## Why should Irish Public Libraries digitise their newspaper collections?

Two findings of the newspaper holdings survey detailed in the Stage 4 report justify research into why Irish Public Libraries should digitise their historical newspaper collection. The factors are:

- 87.5% of libraries in Ireland are interested in newspaper digitisation
- 34.3% of libraries already run a manual newspaper indexing scheme, in order to unlock the information contained in the newspapers.

Therefore, Irish Public libraries clearly want to digitise their newspapers and it is becoming more imperative that libraries digitise their historical newspaper collection. Research shows that there are 4 reasons why libraries should digitise their historical newspaper collections:

- User demand for historical newspapers
- Cultural and Heritage value of newspapers
- Preservation of newspapers
- Promotion and marketing of analogue newspaper collection

### *User demand for historical newspapers*

The National Preservation office (UK) says that, "it would be a poor investment to digitise collections that will only have a small audience and limited time span". User demand is a reason for digitisation forwarded by most leading sources on digitisation including TASI, AHDS and Lee. If user demand for newspapers is low then the cost of digitisation is not justified. But if the user demand is thought to be high for historical newspaper, then a library would be meeting the communities needs by digitising,[3]especially if the digitisation prioritises and 'accommodates peaks in demand'[4].

In relation to the Irish Public Libraries, 50% of all libraries surveyed[5] say their newspaper collection is used very frequently and 40.5% say that their newspaper collection is used frequently. Most libraries have microfilm readers and booking systems are common place. Also peaks in demand were noted around the Irish Famine era, 1916 civil war, World War 1, World War 11 and local events of significance. It is therefore quite clear that the Irish Public Library statistics meet the criteria of user demand warranting digitisation.

[3] http://www.ahds.ac.uk/checklist.htm
[4] Edinburgh University Library
[5] See Stage 4 Report

User demand would most likely be increased if the digitisation were to be used on the web. It became apparent from the telephone survey that the libraries with the most developed local history sections on their websites showed greater remote enquiry level, especially from remote users. These Libraries stated that remote user demand increased as they made material available through their website.[6]

### *Value of newspapers*

Of all the local history material that could and probably should be digitised, why newspapers? The National Preservation Office states that digitisation priority should be given to high value, at risk material of national interest. One can argue about whether newspapers are at risk and this will be discussed under preservation. We need to ask the question are historical newspaper of high value? The British Library sum it up best when they say that, " there is no other medium in our history that records every aspect of human life over the last 30 years – on a daily basis, like newspaper." The British Library go on to say that as newspapers became more complex in the 19th century with the inclusion of advertising, recipes, pictures, reviews, letters, etc. Newspapers are a detailed source of how people lived, dressed, thought and ate.

Irish history over that last 200 years has seen so much change, war, immigration, famine that the newspapers must be seen as a valued historical source for Irish people in Ireland and for the families of the millions who have immigrated. Digitisation for web access could open up this source to Irish people everywhere.

### *Access*

Most leading digitisation sources, Lee, TASI, British Library, RLG, would see increased accessibility to historical newspapers as the main reason why libraries should digitise. As well as this, libraries from abroad that have digitised their newspapers see: " newspaper digitisation as an important part of the library's long-term commitment to digitisation as a primary way to increase access to the library's collection.[7]

There are 2 aspects to accessibility in relation to digitisation:

- Digitisation means that people can access newspapers through more flexible media then bound volumes or microfilm i.e.PCs. This also means that collections can be shared more easily between branches, between library authorities and can even be accessed from the homes, schools and internet cafes i.e. users don't have to go to the library.
- Secondly, digitisation may allow users to find relevant material with the newspaper collection, using retrieval software.

[6] Waterford County Library, http://www.waterfordcoco.ie/index.cgi?cat_id=4

[7] National Library NewZealand.

75% of the libraries surveyed stated that the main reason they would like to digitise there newspaper collection is because it would mean being able to share the historical newspaper collection between branches. All 32 Library Authorities in Ireland are made up of 2 or more branches. Microfilm or hard copy is held usually in one branch, so users have to travel to that particular branch to access the papers. It also means that the collection is lost to various local history, groups and schools not located near the collection. In 1999, British Libraries recorded figures of 407,000 readers visiting the library to look at its holdings, but a staggering 10 million looked at their website and the online collections during the same period.

The logistical benefits to access are clear, but accessibility for newspapers with digital technology can have another aspect. The main problem with newspapers, in original or microfilm format, is that the image is static and there is so much information on various topics with different value available that, "it takes dedicated researchers to handle broadsheet –sized bound volumes of crumbling paper or miles of microfilm, especially when most newspapers are minimally index".[8]
It would be a desire of every local history librarian to have the contents of their newspaper collection unlocked and easily searchable. 75% of librarians surveyed cited searchability as one of the main advantages of digitisation. This backed up by the fact that 34.4% of librarians painstakingly, manually index their newspaper to make them more accessible and useful to day to day users, shows that newspapers, more then any other historical documents, would benefit significantly from digitisation.

***Preservation of newspapers***

Newspapers are particularly difficult both to preserve and access. They are large in format, prolific in output. Their creators intend them as essentially ephemeral – important today discarded tomorrow and so they print them on paper, which is, produced with economy in mind rather the survival.[9]Most libraries in Ireland either hold their newspapers in hardcopy or microfilm format. Findings of the survey carried out by the project team show that 23.5% of newspaper titles in Irish Libraries are hardcopy, 64.4% are microfilm, 7.8% hold a mixture of both. Microfilming is the main preservation method for newspapers and digitisation technology at this stage cannot overtake microfilm, as the long-term preservation method because the longevity of digital image storage mediums are unknown. However, there are problems related to microfilm preservation of historical newspapers.

The UK and the US (Library of Congress and British Library) have become concerned about microfilm preservation as the only method of newspaper preservation because some libraries (due to storage) throw out originals once they have been microfilmed. This is the case in some libraries in Ireland e.g. Mayo County Library, also 64.4% of newspaper titles are held as microfilm only with no original available. Another fear the British Library has is that if microfilm isn't done to a high standard, the deterioration of microfilm over time and use of old, rough readers is severely damaging the condition of microfilm holdings. Issues such as this have lead digitisation commentators to suggest that though microfilm is the main preservation

[8] British Library report,PG.4
[9] British Library report,PG.3

method, digitisation has a role to play in deflecting handling of both originals and microfilm.

***Promote and Market Services***

TASI suggests that digitisation can enhance public knowledge, recognition and understanding of the collection. Even if digitisation of whole newspaper collections is not possible, libraries may choose to digitise 'peak demand' papers, titles or eras and make them accessible through their website. We have already seen that as Irish Public Libraries increase the material available on the website, usage levels increase. The more people that see the online services, the more people become aware of the newspaper, prompting people to go beyond the collection on the website if they are interested. At the very least it promotes the library, as a new, forward-looking service not confined to the walls of the library and the lending of books.

## Digitisation Potential for Irish Public Libraries' newspaper collection

There is no doubt that technically, there is nothing stopping Irish Public Library creating searchable online archives of the newspaper files using OCR software.[10] This can be done in-house using varying degrees of sophistication for retrieval depending on the budget. But realistically to achieve a cost effective, efficient system the key is to keep the amount of editing and proofing at the minimum, so that the process is as automated as possible. Editing and proofing can only be kept low if the OCR accuracy is reasonably high and in turn OCR accuracy directly depends on the quality of the TIFF, which in turn is directly dependent on the accuracy of the source material be it hardcopy originals or microfilm.

The pilot project's discussions with professional bodies, such as DIS, Digiscan and Olive suggest that microfilm in Ireland produced before 1990 will probably be too poor to achieve any degree of automation in the digitisation and OCR process. Even Olive (software designed specifically to deal with historical newspaper digitisation) concluded that Waterford City Library's microfilm is too poor for them to use. This would suggest that the recognition level of OCR is so low that even the addition of fuzzy searching could not compensate for the mistakes. The implication of this is that 54.8% of libraries in Ireland that have their titles on microfilm only have a very low digitisation potential, unless they can source better copies of their microfilm.

The remaining 38.7% of libraries, whose titles are held on original format have a very good chance at achieving digitisation and OCR success, especially if the papers are medium to good quality. For those libraries concerned about preservation as well as

[10] See Stage 2 & 3 reports

access, the ideal situation would be to microfilm the hardcopy first to a very good quality standard, which is OCR friendly and continue the process from there.

# Chapter 3 – Recommendations for a National Newspaper Digitisation Scheme

This reports consists of 4 recommendations for a national newspaper digitisation scheme:

- Digitisation Bureaux
- National Newspaper Digitisation Standards and Standards' Body
- Hybrid Approach to Microfilm and Digital Technology
- Promotion and Marketing of Newspaper Digitisation

***Digitisation bureaux***

Microfilm or large flatbed scanners are very expensive. Therefore, it would be the ideal solution if this equipment could be shared. There are two options here:

1) Regional digitisation centres

Or

2) One National digitisation centre

Either way, the set-up of the centre is key to the success of equipment sharing. Microfilm and flatbed scanners are very big and ideally should not be moved too much to prevent damage. Therefore, digitisation centre/s would physically hold the equipment.

The next issue is staff. If the equipment is to be held at a central or regional centre, then the microfilm or hardcopy newspapers will have to be moved to that centre. The question is, which library's staff will be scanning the material. Are libraries going to send staff to the digitisation centre for the duration of the scanning or is the digitisation centre to have dedicated scanning staff. If each library had to provide their own staff, each set of library staff would have to be trained. As a result quality and standards would vary.

It is recommended that a national centre or regional centres be developed, equipped with the appropriate hardware and software. These centres would deal with digitisation jobs from participating libraries. In other words the digitisation centres would be professional digitisation bureau for the library, archives and museum sector. The digitisation centre/s would provide the staff. This would be the most efficient solution because the digitisation centre staff would only need initial training and the library community can be assured of a consistently high level of quality being produced.

### *National Newspaper Digitisation Standards and Standards' Body*

Lessons should be learned from newspaper microfilming in Ireland. Up until Newsplan was established in 1982, there were no microfilming standards in Ireland and as a result, much of the microfilm collections in Ireland are of very poor quality.

National Newspaper Digitisation Standards should be developed and enforced from the beginning of digitisation in Ireland. A different set of standards would need to be developed for each type of digitisation e.g. photographs, maps, audio and newspapers as each type of material has different digitisation requirements. For newspapers, standards should includes:

- guidelines on collection selection
- digital image creation
- digital image compression and web use
- OCR requirements
- file-naming
- tracking
- meta-data
- digital image preservation
- storage

The development of a national newspaper digitisation standard and a standards body would help to make sure that all newspaper digitisation project around the country are consistent in quality, preservation and delivery.

### *Hybrid Approach to Microfilm and Digital Technology*

According to the survey conducted by Waterford City Library as part of the newspaper digitisation and automated indexing pilot project, 54.8% of newspaper titles held by Public Libraries in Ireland are held on microfilm. Advice and feedback from various digitisation bureaus in Ireland suggested that microfilm more then 10 years old is probably of very poor quality and certainly not of sufficient quality to support digitisation or OCR technology.

Scanning from microfilm is far more cost effective than scanning from broadsheet hardcopies. Hardcopies are large and each page needs to be manually turned, whereas microfilm scanners, scan from the roll automatically.

Scanning from microfilm is also the preferred preservation solution. Microfilm remains the main long-term preservation method for newspapers, however, digitisation poses excellent potential for increasing access and searchability of newspapers. The ideal is to microfilm and scan.

The digitisation potential for the 54.8% of newspapers titles held on microfilm would have been greatly increased if proper microfilm standards had been introduced earlier. Digitisation technology is the future for newspaper access. Therefore, microfilming standards should be developed in –line with digitisation standards. In other words,

when a newspaper is microfilmed, the microfilm should be of a quality and standard that will achieve the optimum image quality for OCR technology when scanned.

### *Promotion and Marketing of Newspaper Digitisation*

Due to the every day demands on libraries and library staff, very few staff have the time to consider initiatives such as newspaper digitisation. The digitisation centres or centre should promote and market the digitisation of newspapers. This would included showcasing the potential uses of newspaper digitisation and outlining how the digitisation centres can overcome barriers such as staff time, lack of space, cost of equipment etc.

### *Continuance of a National Cultural Portal Website*

The potential of digitisation would be lost if libraries only hosted their own digitised material on their own library websites. Libraries need to be encouraged to contribute digitised material to both their own website and the National Portal. This way, users can have the benefit of all the collections from around the country being accessible through one central portal with the option to go to specific library sites for more details.

# Chapter 4 – Costings for a National Digitisation Scheme

Based on the findings of the pilot project there are 6 stages to the digitisation and automated indexing process:

- Scanning — Approx. 3 mins per article
- Compression — Approx. 3 mins per article
- OCR — Approx. 12mins per article
- Webpage Creation — Approx. 9 mins per article
- Inserting image links — Approx. 3 mins per article
- Uploading of webpages — Approx. 6 mins per article

Some of these stages take longer then others but if the timescale of 3 mins per article is used for each stage then the following workstations in the digitisation production line would be required to complete 2 years worth of newspaper in 12 months.

The Production Line requirements are based on:

- Weekly, broadsheet newspapers consisting of approximately 30 articles per page, 10 pages per weekly paper.
- A full-time staff of 11, working an standard 35 hour week for 12 months at a rate of 3 mins per article per workstation simultaneously.

## Production Line

| Workstation | Functions | Staff | Hardware | Software |
|---|---|---|---|---|
| Workstation 1: Scanning & Clipping | • Scanning<br>• Selection of Articles<br>• Electronic clipping<br>• Saving output files to TIFF<br>• File naming*<br>• Tracking | 1 Grade 3 | • Scanner<br>• CD rewriter<br>• PC - RAM | • Scanner Software<br>• Paintshop Pro |
| Workstation 2: Image Compression | • Creation of compressed jpeg file.<br>• Creation of thumbnail jpeg file<br>• File Naming<br>• Tracking | 1 Grade 3 | • PC | • Paintshop pro |
| Workstation 3: OCR | • Putting Tiff through the OCR Software<br>• Editing and proofing OCR | 1 Grade 3 | • PC<br>• Printer | • OCR software<br><br>Abbyy or Acrobat |

| | | | | |
|---|---|---|---|---|
| | output<br>• Saving to PDF<br>• File Naming<br>• Tracking | | | |
| Workstation 4: OCR | Same as Workstation 3 | | | |
| Workstation 5: OCR | Same as Workstation 3 | | | |
| Workstation 6: OCR | Same as Workstation 3 | | | |
| Workstation 7: Webpage Creation - text | • Creation of webpage 2 templates<br>• Cut and Paste OCR output text to template 1<br>• Make sure dates of each page is changed for each new article<br>• Links<br>• File naming Tracking | 1 Grade 3 | • PC<br>• CD rewriter<br>• Printer | • Internet Explorer<br>• Frontpage<br>• Acrobat Reader<br>• Paintshop pro |
| Workstation 8: Webpage Creation - text | Same as workstation 7 | | | |
| Workstation 9: Webpage creation – images | • Insert thumbnail jpeg to template 1<br>• Insert expanded image jpeg to template 2<br>• Links<br>• File naming Tracking | 1 Grade 3 | • PC<br>• CD rewriter | • Internet Explorer<br>• Frontpage<br>• Acrobat Reader<br>• Paintshop pro |
| Workstation 10:Uploading & QA | • Proof- reading<br>• Final tracking<br>• Tracking checks<br>• Metadata<br>• Preservation of digital image – saving, printing, storage<br>• Uploading<br>• Quality Assurance | 1 Grade 5 | • PC<br>• CD rewriter<br>• Printer | • Frontpage<br>• Internet Explorer<br>• Excel<br>• Word |
| Workstation 11:Uploading & QA | Same as Workstation 10 | 1 Grade 3 | | |

## Equipment Costs:

(Price in Euro.)

| | € |
|---|---|
| 1 x Microfilm /Flatbed Scanner | Approx. 15,000. |
| 11 x PCs | Approx. 11,000 |
| 4 x CD Rewriters | Approx. 600 |
| 3 x Printers | Approx. 1,300 |
| 4 x OCR software (ABBYY) | Approx. 480 |
| 4 x Paintshop Pro | Approx. 480 |
| 4 x Frontpage | Approx. 1,116 |
| Total | 29,976 |

## Staff Costs:

| | |
|---|---|
| 10 x Grade 3 (average salary of 23,000) | Approx. Average 230,000 |
| 1 x Grade 5 (average salary of 33,000) | Approx. Average 33,000 |
| Total | 263,000 |

## Other Costs:

Room/ Space Rental
Web site maintenance and design
Transports

This would be the approximate cost of digitising and automatically indexing 2 years of 1 newspaper title using a staff of 11 working simultaneously on 11 workstations over a 12 month period.

If more newspapers need to be digitised in a 12-month period, simply increase the number of Production Lines.

# References


1) Edinburgh University Library (EUL)

2) Digitisation: a project planning checklist. Arts and humanities data service,2002. http://www.ahds.ac.uk/checklist.htm

3) The Nordic digital newspaper library. Nord info-nytt:Bremer,2001. http://www.nordinfo.helsinki.fi/publications/nordny#/nny#2_01/bremer.htm

4) Managing the digitisation of library, Archive and Museum material. National Preservation Office. http://www.bu.uk/npo/

5) Need to increase the access of cultural and heritage in libraries across Europe via the Internet. European Commission. [WWW] http://www.Digicult.info (accessed on 17/01/03)

6) Selection and preparation of material, 2002. [WWW] http://www.tasi.ac.uk/advice/creating/selection.html (accessed on 12/11/02)

7) PADI (Preservation access to digital information), a comprehensive and well-maintained clearinghouse to all types of information resources related to digital preservation.[WWW] http://www.nla.gov.au/padi/ (accessed on 16/12/02)

8) Beagrie, Neil and Greenstein, Daniel. A strategic policy framework for creating and preserving digital collections. Arts and Humanities Data Service UK, 2001. [WWW] http://ahds.ac.uk/managing.htm (accessed on 12/12/02)

9) Deegan M., King E., Steinvil E. Automated Indexing of newspapers and grey literature. The British Library Newspaper Library, The Refugees Studies Centre at Oxford University, The Malibu Hybrid Library Project, OCLC, and olive software Inc., Jan. 2001.

10) British Library Online Newspaper Archive. http://www.uk.olivesoftware.com/

11) OCLC historical newspaper service. [WWW] http://www.oclc.org/digitalpreservation ,(accessed on 20/01/03)

12) Deegan,M., King E., Steinvil E. Digitising Historcial Newspapers: Progress and Prospects. RLG Diginews, Aug.15th,2002, Vol.6, No.4.

13) Lee, Stuart D. Digital Imaging: a practical handbook. Library Association Publishing:London,2001

14) Swora,T. Selecting Library and archive collections for digital reformatting. Mountain View, 1996.

15) Papers Past. National Library of New Zealand. http://paperspast.natlib.givt.nz/about.html

16) Alexander, Michael.cBritish Library computing and telecommunications at the 2nd NPO conference, 1998. http://www.nla.gov.au/niac/meetings/npo98ma.html

17) Gertz, Janet. Selection for preservation in the digital age: an overview. Library resources and technical services, 44 No.2 97-104 April 2000.

18) Joining Forces – delivering libraries and information services in the information age. An Chomhairle Leabharlanna ( The Library Council), Sep 1999.

19) Branching out: a public library review. Department of the environment and local government. The Stationery Office: Dublin,1998.

20) Smith, A. Why digitise? and the future of the past: preservation in American research libraries. CLIR, 1999. Http://www.clir.org/pubs/reports/reports.html

21) Deciding to digitise. Technical advisory service for images. http://wwtasi.ac.uk/advice/managing/pdf/digitise.pdf

22) Ling, Ted. Ehy the archives introduced digitisation on demand. RLG diginews, August 15th, 2002, Vol.6 , No.4

23) Beagrie, Neil. Going digitial: issues in digitisation for public libraries. Joint Information systems committee (JISC), The Library Association and UKOLN. Http://www.ukoln.ac.uk/public/earl/issuepapers/digitisation.why

24) National Library of Australia digitisation policy 2000-2004. http://www.nla.gov.au/policy/digitisation.html

25) Ayris, Paul. RLG. http://www.rlg.org/preserv/joint/ayris.html

26) http://www.ahds.ac.uk

27) http://www.ota.ahds.ac.uk/documents/creating/chap3.html

28) Alkura, R., and Pieska, K.Optical Character recognition in microfilmed newspaper library collections – a feasibility study. Technial research centre of Finland, ESPOO, 1994.

29) http://www.digiscan-ace.com

30) http://www.olivesoftware.com

31) http://ahds.ac.uk/guides.htm

32) Kenney, Anne R., Rieger Oya Y. Moving theory into practice: digital imaging for libraries and archives. RLG,2000.

33) Caribbean Newspaper imaging Project. http://web.uflib.ufl.edu/digital/collections/cnip/eng/project.htm

34) Steven's 'Stute' newspaper digital archive project. http://stute.jacobus.stevens-tech.edu/

35) Chapman,Steven.http://www.oclc.org/oclc/presres/pubpres/corcpathfinder.htm

36) Hewlett –Packard. Choosing a scanner (online). http://www.scanjet.hp.com

37) Chapman, Steven. Guidelines for Image Capture. http://www.rlg.org/preserv/joint/chapman.html