

Dealing with Missing Pages

Author: Rose Holley (Manager - ANDP)

Version: 2.0

Date: 25 January 2008

Introduction

At the start of the program it had been agreed that flagging known missing pages to users and adding missing pages into the resource if they were found later was of critical importance. Most other newspaper projects have come to the 'missing page issue' at a much later stage of the project when it is impossible to change workflows and therefore add in missing pages when they are found in an efficient way. However since the missing page issue had been identified at the start of the program as something that needed addressing workflows have been set up and solutions implemented so that missing pages can be dealt with in the best way. The missing page scenario is described below.

Why are pages missing?

There are two reasons why pages are missing:

1. The original microfilm has missing pages or entire issues since an incomplete copy of the newspaper was sent for microfilming. This is particularly evident in the Sydney Morning Herald. Hard copies may or may not be still available for these pages.
2. Some types of microfilm scanning software 'jump frames' when the microfilm is being scanned resulting in missing digital pages (where pages that are on the microfilm are not converted into digital images).

What is the extent of missing pages?

At the start of the ANDP this was unknown. No Library had ever tried to calculate the extent of missing pages on newspaper microfilm before. Staff performing quality assurance on the ANDP newspaper pages perceived the missing page issue to be greater than 5% of pages. On the strength of this it was decided to formulate some solutions. In addition it was also agreed to set up a simple software algorithm which would automatically calculate the number of missing pages and issues in the digital collection, so that the exact extent of the issue could be identified and solutions reviewed. After 9 months the following figures for extent of missing pages are now available. The extent is far less than originally perceived. These pages do not include the Sydney Morning Herald. It is perceived that this paper has a higher rate of missing pages, though this is yet to be verified or analysed further.

From the Pilot 50,000 pages (a representative sample of the 3 million corpus):

534 missing pages + 43 missing issues (at 8 pages each~ 344) = 878 missing pages

1.7% missing page rate

From the first 370,000 pages scanned:

1057 missing pages + 225 missing issues (at 8 pages each~ 1510) = 2567 missing pages

0.7% missing page rate

How are missing pages identified?

Missing pages can only be identified by manually checking the sequence of all digital images. Sometimes microfilm targets are included indicating where pages are missing. If frames have been jumped on the microfilm this is very hard to detect unless the microfilm is compared to the digital resource. The Library is unable to do this level of checking since it is too time consuming and in addition in most cases the Library does not have the microfilm on the premises.

Solutions

Several solutions were identified and are in place:

1. Only Nextstar microfilming software is now being used to scan microfilm. This does not 'jump frames' and the Library is confident that all pages present on the microfilm are being converted into digital images.
2. The first 200,000 pages filmed using other software were re-scanned when the amount of missing pages was much higher than would be expected. In these cases it was much more efficient to simply scan the entire reel again rather than try to identify and replace individual images.
3. The Quality Assurance module of the Content Management System was developed so that missing pages and issues could be flagged during the Quality Assurance process, so that subsequent workflows for missing pages can take place.
4. The Library created 'digital targets' which act as placeholders for missing pages and issues. These replace the microfilm targets (if any exist) and have a virtual sequence number. This enables them to be easily replaced with found real pages. The targets also have a consistent look and contain a clear citation so that they may be used in the search and delivery system as a way of showing users that pages are missing.
5. Lists of missing pages for titles can be generated from the Content Management System so that stakeholders can assist in searching for hard copies of missing pages for re-filming. This will be necessary in particular with the Sydney Morning Herald.
6. Both scanning and OCR contractors were asked to be able to support the workflow and processing for found missing pages and issues out of sequence. These workflows have been tested and are workable. However for efficiency it is intended to group missing pages together at one time, rather than process odd pages each week.

Workflow Example

The Library is undertaking quality assurance of digitised newspaper pages from microfilm to a high level. This involves entering data for each page which includes giving every page a virtual page number. The pages are also placed in the correct sequence. Often the order pages have been microfilmed in was not the correct sequence. Often the pages do not have a number appearing so the page number must be guessed. Through a simple algorithm the quality assurance system predicts which pages and issues are thought to be missing, the operator will verify these and the system generates digital targets. An operator can also manually create targets for known missing pages/issues.

These targets will be clearly visible to end users in the search and delivery system as they browse through an issue page by page.

When a missing page or issue is found the workflow enables it to be scanned out of sequence and then sent for OCR processing out of sequence. This is largely because page number

metadata is separate and distinct from the file names and persistent identifiers. Because the digital targets act like placeholders and have a virtual page number the real page can easily be switched with the digital target without affecting the page sequencing or persistent identifiers.

Fig 1. Identification of missing pages.

Notes: Entire issue filmed with background page visible

P9 [238413] [Edit](#)



P10 [238414] [Edit](#) P11 [238415] [Edit](#)



Missing page target
 Select as well if the page blank
 Page:
 Edtn seq:
 Edtn name:
 Supl seq:
 Supl name:
 Sect seq:
 Sect name:

Missing page target
 Select as well if the page blank
 Page:
 Edtn seq:
 Edtn name:
 Supl seq:
 Supl name:
 Sect seq:
 Sect name:

P14 [2384



Notes: Page missing

Issue: 1920-09-07 15 pages [add missing page](#)

P1 [238417] [Edit](#)



Notes: Filmed with background page visible

P1 [238419] [Edit](#)



Notes: Entire issue filmed with background page visible

[UNRESOLVED DUPLICATE](#)

P2 [238420] [Edit](#)

P3 [238421] [Edit](#)

P4 [238422] [Edit](#)

P5 [2

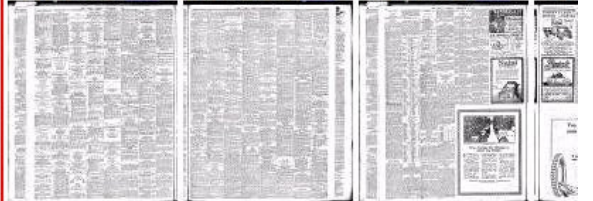



Fig 2. Verification of missing pages causes automatic target generation.


P9 [238413] [Edit](#) P10 [238414] [Edit](#) P11 [238415] [Edit](#) P12 [238790] [Edit](#) P13 [238791] [Edit](#)



NLA generated **Missing page target** NLA generated **Missing page target**

Issue: 1920-09-07 15 pages [add missing page](#)

P1 [238417] [Edit](#) P1 [238419] [Edit](#) P2 [238420] [Edit](#) P3 [238421] [Edit](#) P4 [238422] [Edit](#)



Notes: Filmed with background page visible
SELECTED DUPLICATE

Notes: Entire issue filmed with background page visible
DISCARDED DUPLICATE
WILL BE SUPPRESSED

Fig 3. NLA Digital Newspaper Targets – Issues

NATIONAL LIBRARY OF AUSTRALIA

Colonial Times
Issue Missing: 1856-05-14

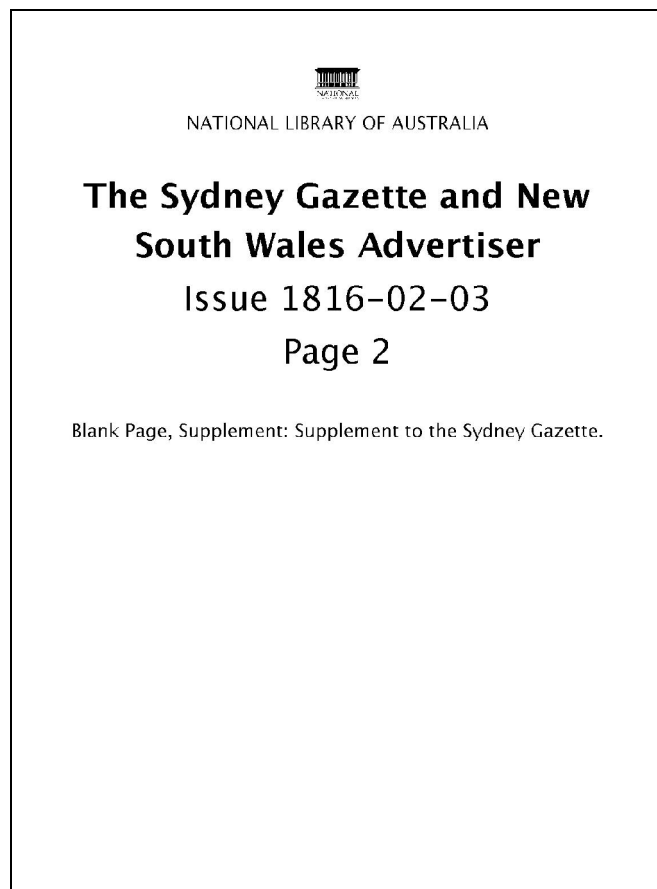
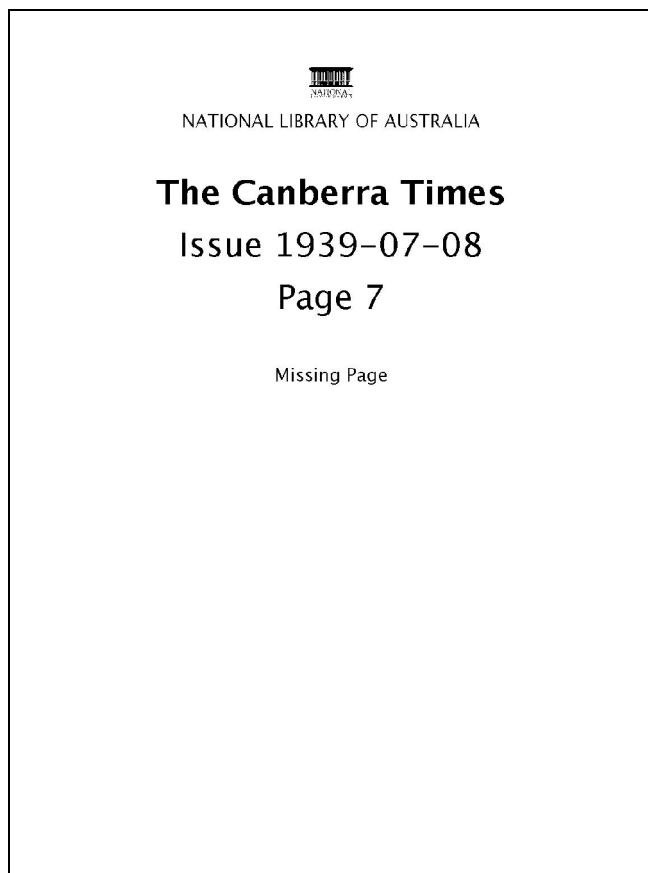
Issue is missing

NATIONAL LIBRARY OF AUSTRALIA

The Canberra Times
Issue Missing: 1941-04-11

Issue was not published

Fig 4. NLA Digital Newspaper Targets – Pages



Summary

The Library has now identified that the extent of missing pages is a much smaller % than was originally thought. Workflows and systems have been established to allow integration of 'found' missing pages. Despite the % being lower than the Library anticipated the Library sees value for the end user in having addressed this issue at an early stage of the program. It is anticipated that by working closely with ANPlan and members of the public missing pages may be sourced in the future. Some individual missing pages hold high value to researchers.