

## Newspaper Article Categories

Author: Rose Holley (ANDP Manager)

Version: 1.3

Date: 23 May 2008

### 1. Overview

When pages are sent for OCR processing the Contractor categorises articles, as per the Library's specifications. The 4 categories being used are:

- News
- Family Notices
- Advertising
- Detailed Lists, Results and Guides

The purpose of applying a category to articles is to increase search success for users in the search and delivery system. Over the last year the Library has experimented with using a number of different article categories in order to determine:

- How accurately Contractors can assign categories to articles?
- How useful categories are for searching and why?

### 2. Category Specification

The Library's specification for article categories was finalised in May 2008 and is summarised below.

The full specification for content analysis and Optical Character Recognition (OCR) is available at:

[http://www.nla.gov.au/ndp/project\\_details/documents/ANDP\\_StatementofWorkSpecificationforContentAnalysisandOCR.pdf](http://www.nla.gov.au/ndp/project_details/documents/ANDP_StatementofWorkSpecificationforContentAnalysisandOCR.pdf)

#### News

News articles cover a wide range of subject matter, including current affairs, law courts and crime, official appointments and notices, commerce and business, sport and social news, obituaries, editorials, letters and correspondence (usually to the editor) and editorial or political cartoons.

Articles with subject matter related to shipping news or intelligence, including arrival and departure information, sailing schedules, and fares and services for ships are categorised as News articles.

Articles with subject matter related to art, literature, music, theatre, comics, shows, gardening, travel, crafts (such as crochet and knitting), stories, fiction and poetry are categorised as News articles.

News is the default category. If an article cannot be clearly identified as any of the specific article types listed below it will be categorised as News.

## **Family Notices**

Birth, death and marriage notices and related announcements including weddings, anniversaries, in memoriam, bereavement, birthdays, and congratulations are categorised as Family Notices.

## **Advertising**

This category contains both display advertising and classified advertising. Display advertising usually contains both text and graphic information such as logos, drawings, or other pictures or photographs. Display advertising is usually large in size spanning multiple columns or single entire columns, with large fonts and is placed outside of the classified advertisements section on whole pages or inserted amongst news items. All advertising in the newspaper masthead although small is display advertising.

Classified advertising generally appears in a specific section of the newspaper and contains text only. Classified advertising may include property notices, items for sale, employment notices, public and personal notices.

## **Detailed Lists, Results, Guides**

This category contains detailed sporting results, guides, radio and television guides, weather forecasts, election results, education results and courses and stock market lists, crossword puzzles, word games and quizzes.

### **3. Accuracy of Assigning Categories**

Article categories must be assigned accurately or they will not be useful in enabling the user to achieve high quality search results. The Library has quality acceptance criteria of 98% for article categorisation. Initially the Library commenced with 11 specific article categories, however through the pilot phase, there was inconsistent and incorrect application of these categories. This was because reading of the first lines of the article and comprehension of the article was really required in order to select the most appropriate article category, e.g. to define a sports article from a news article. As the Library's Contractor was based in India and production staff did not have English as their primary language, article categories were applied most often based on the visual layout of the newspaper page. As an outcome of the pilot process the Library decided to modify and reduce the number of article categories.

Another strategy that was considered to enable more accurate article categorisation is use of a vocabulary list. The vocabulary list would identify specific words associated with certain categories, e.g. the words 'weddings', 'marriages' would be associated with the category Family Notices. A vocabulary list works well on single newspaper title over a specific date range where the content, headings and layout of the newspaper is known or can be analysed easily. It is however, much harder to apply across a broad range of titles and dates such as the scope of the ANDP.

The four categories currently in the Library's specification can now be achieved to an accuracy level of 98% or higher. Categories are largely assigned based on visual layout cues and sometimes with use of a limited vocabulary list. Little if any intellectual 'comprehension' of the subject of the article is taking place. If at all, only the title and the first four lines of the article are read to assist with assigning the article category.

### **4. Usefulness of categories in the Search system**

Initially it was intended to use article categories to assist users to search for certain types of information, e.g. shipping notices, sports news. However it quickly became apparent that categories such as obituaries, death notices and shipping news could be easily retrieved by full text searching on these words in combination with keywords, e.g. "deaths and Peter Smith".. The strategy for use of categories then changed. Categories were seen as more useful for users to refine searches, in particular, to exclude content that was not useful for their research, e.g. classified advertisements, sports results, crossword puzzles. After much internal discussion it was determined that the 3 main areas of a newspaper were: articles; advertising and lists, e.g. sports results, and that these should be used as categories for

refinement. Due to the significant interest by genealogists and family historians it was also decided to retain the article category 'Family Notices' which includes birth, death, marriage and related notices.

As further development took place on the search and delivery system article categories were also used to relevancy rank articles in results sets. There are 6 elements in the relevancy ranking algorithm, of which article category is one. At present if all the keywords are found in an article and the user has not refined the search by category (already limited to a news article) then articles will be ranked lower if they are a detailed list (20%) and even lower if they are advertising (10%). For example, if three articles had each scored 500 in the other 4 parts of the relevancy ranking algorithm, a news article's final score would be 500, a detailed list's final score would be 100, and an advertisement's final score would be 50. Therefore in the results list the news article appears first, the detailed list second and the advertising third. More details on the use of relevancy ranking in the search and delivery system are available at: [http://www.nla.gov.au/ndp/project\\_details/documents/ANDP\\_Relevancy\\_ranking.pdf](http://www.nla.gov.au/ndp/project_details/documents/ANDP_Relevancy_ranking.pdf)

## 5. Treatment of Illustrations

Illustrations are not listed as a separate article category. Articles containing illustrations or which are stand alone illustrations have a metadata illustration indicator added and are categorised using the same four categories. Use of the metadata indicator enables a user in the search and delivery system to refine their search by 'illustrated article'. Standalone illustrations, e.g. where there is no article text accompanying the illustration or only a brief caption, are categorised as News articles. Illustrations appearing in News articles also have the following illustration types applied to enable further refinement of searches:

- Illustration (default)
- Photo
- Cartoon
- Map
- Graph