Segmentation Policy: Items

0.1 Nomenclature:

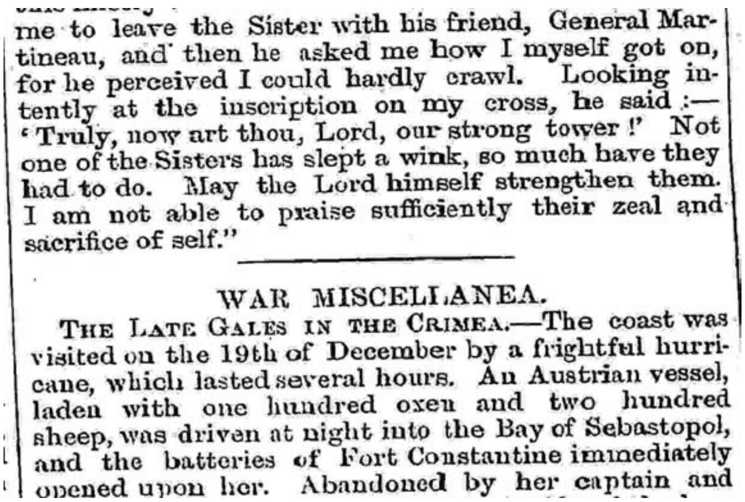
Term	Definition
Title	The title of the whole run of the periodical
Volume	A set of numbers, usually grouped annually and numbered sequentially.
Number	An individual issue.
Department	A division within a number such as correspondence or news.
Item	An article or component within a department.
Wrapper	An advertising wrapper that surrounds a number. It is usually paginated separately to distinguish it from the paginated letterpress (although not in the case of the *Leader*).
Volume Preliminaries	It was common practice in the C19th to issue material to be bound in with the numbers at the end of the volume. These might include a frontispiece and a preface.
Index	It was also common practice to issue an index to the volume, to be bound in at the back.
Masthead	The name of the title, usually printed in a fancy typeface.
Dateline	The strapline that contains information such as volume, number, and date.

These are arranged in a hierarchical structure:

Edition
- Title
    - Volume #1
        - Volume Preliminaries
        - Number #1
            - wrapper [if present]
            - number contents
                - department #1
                    - item #1
                    - item #2
                    - etc…
                - department #2
                - department #3
                - etc…
        - Number #2
        - Number #3
        - Etc…
        - Index
    - Volume #2
    - Etc…

1. 0 Segmentation Policy

1.1 Items in the *Leader* are separated by horizontal lines like this:



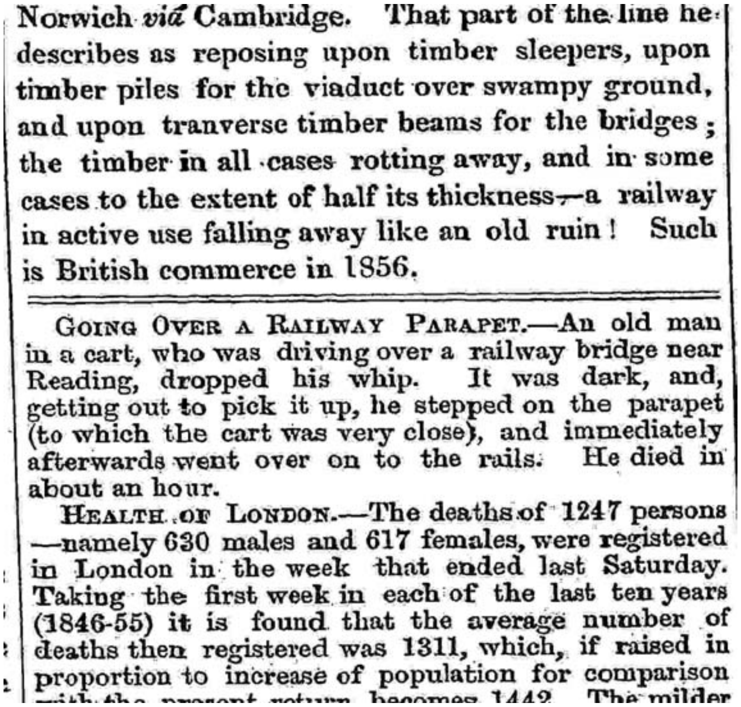
me to leave the Sister with his friend, General Martineau, and then he asked me how I myself got on, for he perceived I could hardly crawl. Looking intently at the inscription on my cross, he said :—'Truly, now art thou, Lord, our strong tower !' Not one of the Sisters has slept a wink, so much have they had to do. May the Lord himself strengthen them. I am not able to praise sufficiently their zeal and sacrifice of self."

WAR MISCELLANEA.

THE LATE GALES IN THE CRIMEA.—The coast was visited on the 19th of December by a frightful hurricane, which lasted several hours. An Austrian vessel, laden with one hundred oxen and two hundred sheep, was driven at night into the Bay of Sebastopol, and the batteries of Fort Constantine immediately opened upon her. Abandoned by her captain and

1.1.1 Items divided by a horizontal line

This means these can be used to mark the end of one item and the beginning of the one that follows it. However, there are a number of different types of horizontal line in the run of the *Leader*. For instance, on the opposite page as the above is:



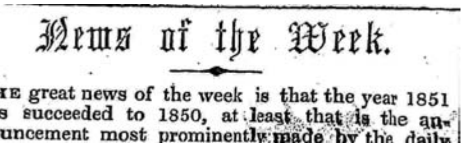
Norwich *viâ* Cambridge. That part of the line he describes as reposing upon timber sleepers, upon timber piles for the viaduct over swampy ground, and upon tranverse timber beams for the bridges ; the timber in all cases rotting away, and in some cases to the extent of half its thickness,—a railway in active use falling away like an old ruin ! Such is British commerce in 1856.

GOING OVER A RAILWAY PARAPET.—An old man in a cart, who was driving over a railway bridge near Reading, dropped his whip. It was dark, and, getting out to pick it up, he stepped on the parapet (to which the cart was very close), and immediately afterwards went over on to the rails. He died in about an hour.

HEALTH OF LONDON.—The deaths of 1247 persons —namely 630 males and 617 females, were registered in London in the week that ended last Saturday. Taking the first week in each of the last ten years (1846-55) it is found that the average number of deaths then registered was 1311, which, if raised in proportion to increase of population for comparison with the present return becomes 1442. The milder

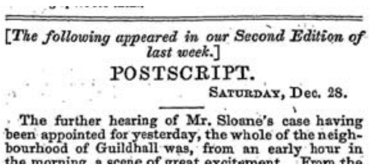
1.1.2. Items divided by a double horizontal line

There are also other examples of line:



News of the Week.

THE great news of the week is that the year 1851 has succeeded to 1850, at least that is the announcement most prominently made by the daily



[*The following appeared in our Second Edition of last week.*]

POSTSCRIPT.

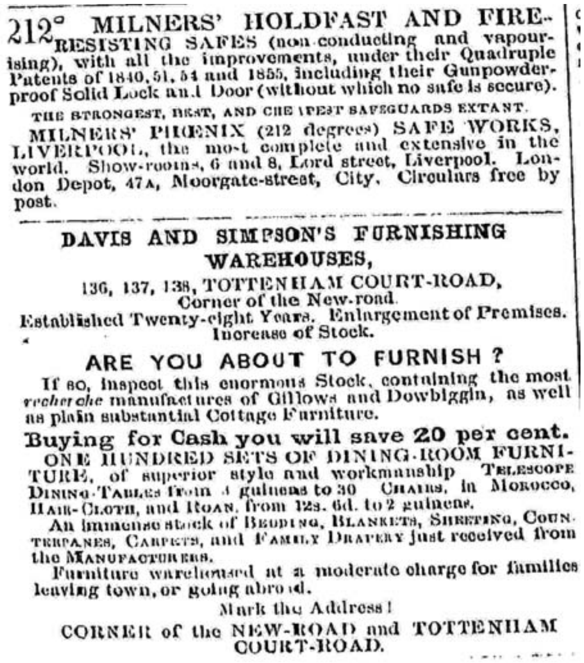
SATURDAY, Dec. 28.

The further hearing of Mr. Sloane's case having been appointed for yesterday, the whole of the neighbourhood of Guildhall was, from an early hour in the morning a scene of great excitement. From the

1.1.3. Items divided by other types of line.

However, the line still divides the items, regardless of its format.

1.2. This is a consistent rule throughout the run of the *Leader*, and is a better guide than attempting to use headlines. Although the first example has a headline, and so could be used as a marker, the second does not. Although the block capitals might be used, the second paragraph also begins with block capitals, even though it is the part of the same item. Typographical indicators also cause problems when it comes to the advertisements:



212° MILNERS' HOLDFAST AND FIRE-RESISTING SAFES (non-conducting and vapourising), with all the improvements, under their Quadruple Patents of 1840, 51, 54 and 1855, including their Gunpowder-proof Solid Lock and Door (without which no safe is secure).

THE STRONGEST, BEST, AND CHEAPEST SAFEGUARDS EXTANT.

MILNERS' PHŒNIX (212 degrees) SAFE WORKS, LIVERPOOL, the most complete and extensive in the world. Show-rooms, 6 and 8, Lord street, Liverpool. London Depot, 47A, Moorgate-street, City. Circulars free by post.

DAVIS AND SIMPSON'S FURNISHING WAREHOUSES,

136, 137, 138, TOTTENHAM COURT-ROAD,
Corner of the New-road.
Established Twenty-eight Years. Enlargement of Premises.
Increase of Stock.

ARE YOU ABOUT TO FURNISH ?

If so, inspect this enormous Stock, containing the most *recherche* manufactures of Gillows and Dowbiggin, as well as plain substantial Cottage Furniture.

Buying for Cash you will save 20 per cent.

ONE HUNDRED SETS OF DINING-ROOM FURNITURE, of superior style and workmanship TELESCOPE DINING-TABLES from 4 guineas to 30 CHAIRS, in MOROCCO, HAIR-CLOTH, and ROAN, from 12s. 6d. to 2 guineas.

An immense stock of BEDDING, BLANKETS, SHEETING, COUNTERPANES, CARPETS, and FAMILY DRAPERY just received from the MANUFACTURERS.

Furniture warehoused at a moderate charge for families leaving town, or going abroad.

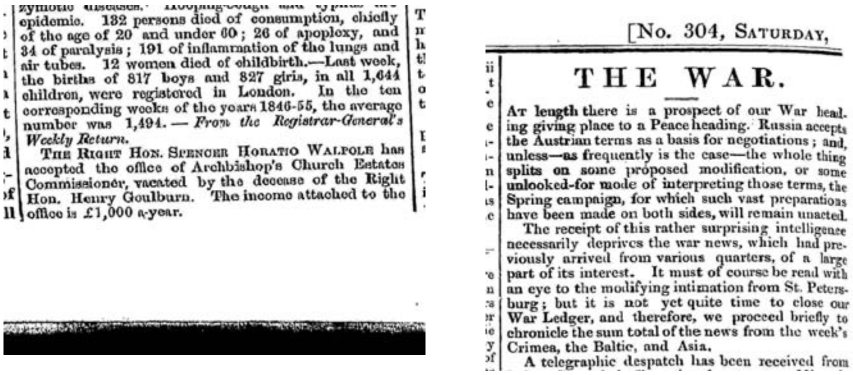
Mark the Address!

CORNER of the NEW-ROAD and TOTTENHAM COURT-ROAD.

1.2.1. Advertisements showing different text sizes, but divided by a horizontal line.

As the above example shows, advertisements employ a range of different sizes of types, making it impossible to use them to determine where items begin and end. However, as you can see, a horizontal line divides the advertisements, and so is a much more reliable indicator.

1.3 The only drawback with using horizontal lines to identify where items end is that those items at the top or bottom of the page do not have them. For instance, figue 1.3.1 gives two examples: the one on the left shows an item that ends at the bottom of the column. As you can see, there is no line to mark this. Equally, the example on the right begins an item, and it has no line above it (apart from the line dividing the letterpress from the dateline).



zymotic diseases. Hooping-cough and typhus ... epidemic. 132 persons died of consumption, chiefly of the age of 20 and under 60; 26 of apoplexy, and 34 of paralysis; 191 of inflammation of the lungs and air tubes. 12 women died of childbirth.—Last week, the births of 817 boys and 827 girls, in all 1,644 children, were registered in London. In the ten corresponding weeks of the years 1846-55, the average number was 1,494. — *From the Registrar-General's Weekly Return.*

The Right Hon. Spencer Horatio Walpole has accepted the office of Archbishop's Church Estates Commissioner, vacated by the decease of the Right Hon. Henry Goulburn. The income attached to the office is £1,000 a-year.

[No. 304, Saturday,

THE WAR.

At length there is a prospect of our War heading giving place to a Peace heading. Russia accepts the Austrian terms as a basis for negotiations; and, unless—as frequently is the case—the whole thing splits on some proposed modification, or some unlooked-for mode of interpreting those terms, the Spring campaign, for which such vast preparations have been made on both sides, will remain unacted.

The receipt of this rather surprising intelligence necessarily deprives the war news, which had previously arrived from various quarters, of a large part of its interest. It must of course be read with an eye to the modifying intimation from St. Petersburg; but it is not yet quite time to close our War Ledger, and therefore, we proceed briefly to chronicle the sum total of the news from the week's Crimea, the Baltic, and Asia.

A telegraphic despatch has been received from

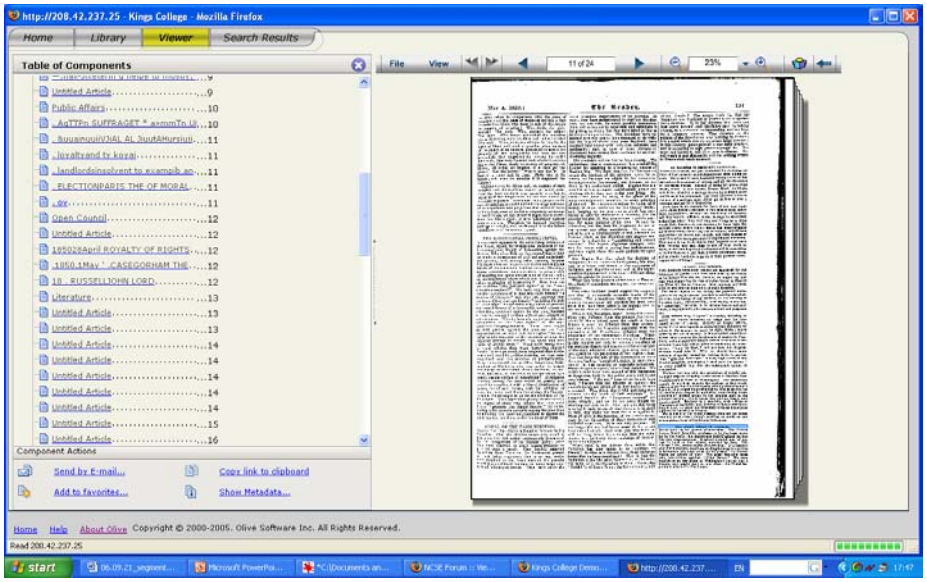
1.3.1. Items that end at the bottom and begin at the top of a column have no dividing lines.

In case where an item appears at the top of the page, it is better to use typographical features such as gothic typeface or larger type to signal the beginning of an item. This is not a problem for advertisements or tables, as they never run from one column to another, so when segmented they will always end at the bottom of a column, and a new item begin at the top of the next column.

In Manual Editing Option 1, circulated as part of this pilot:

http://208.42.237.25/Olive/FileCabinet/KC/welcome.htm

A segmentation policy has been implemented that seems to cope with this problem. For instance:



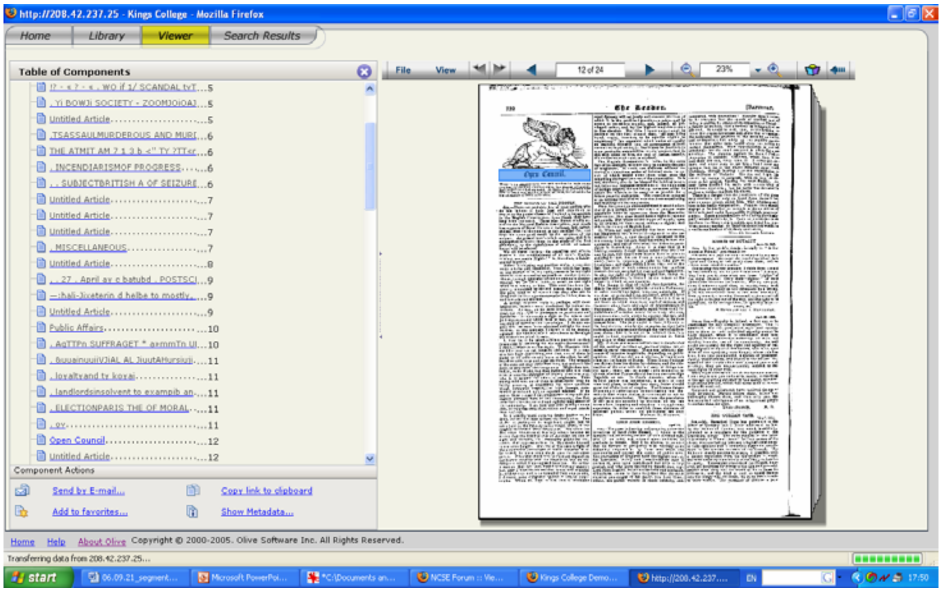
Untitled Article....9
Public Affairs....10
AgTTPn SUFFRAGET * asmmTo U....10
SuuamuuiVJIAL AL JuutAHursiuli....11
Joyaltyand ty koyal....11
landlordsinsolvent to exampib an....11
ELECTIONPARIS THE OF MORAL....11
.ox....11
Open Council....12
Untitled Article....12
185028April ROYALTY OF RIGHTS....12
1850.1May ' CASEGORHAM THE....12
18 . RUSSELLJOHN LORD....12
Literature....13
Untitled Article....13
Untitled Article....13
Untitled Article....14
Untitled Article....14
Untitled Article....14
Untitled Article....14
Untitled Article....14
Untitled Article....15
Untitled Article....16

Copyright © 2000-2005. Olive Software Inc. All Rights Reserved.

Read 208.42.237.25

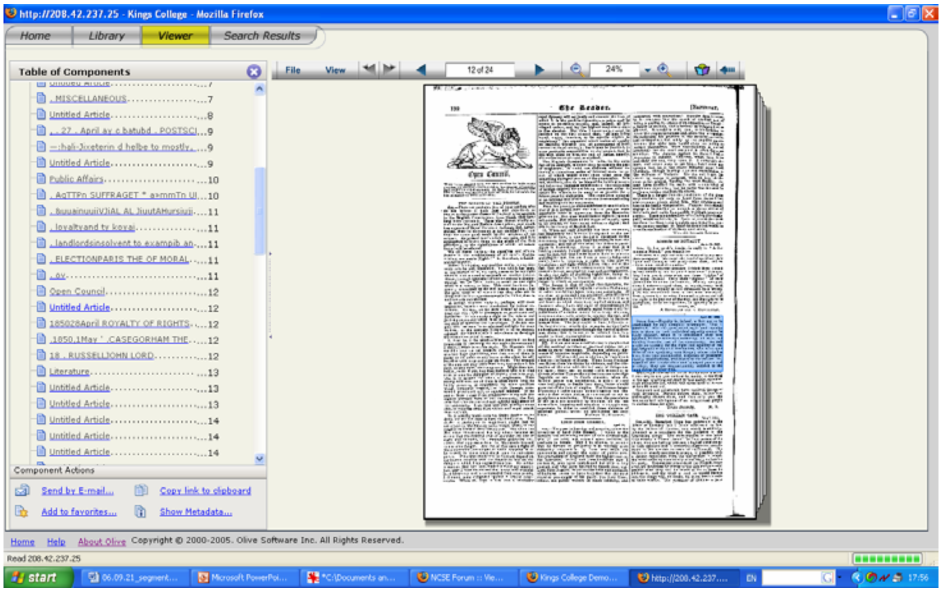
1.3.2. Screenshot of Manual Editing 1 showing correct segmentation.

The item highlighted ends at the bottom of this page. There is a line above to signal where it starts, but whatever segmentation policy was used to segment this pilot has realized where this item ends. The next item, on the following page, has also been segmented correctly:


http://208.42.237.25 - Kings College - Mozilla Firefox
Home
Library
Viewer
Search Results
Table of Components
!? - « ? - « . WO if 1/ SCANDAL tvT...5
. Yi BOWJi SOCIETY - ZOOMJOIOAJ...5
Untitled Article...5
.TSASSAULMURDEROUS AND MURI...6
THE ATMIT AM 7 1 3 b <" TY ?TTer...6
. INCENDIARISMOF PROGRESS...6
. . SUBJECTBRITISH A OF SEIZURE...6
Untitled Article...7
Untitled Article...7
Untitled Article...7
. MISCELLANEOUS...7
Untitled Article...8
. . 27 . April ay c batubd . POSTSCI...9
—:hali-Jiveterin d helbe to mostly...9
Untitled Article...9
Public Affairs...10
. AqTTPn SUFFRAGET * axmmTn Ul...10
. 8uuainuuiiVJiAL AL JiuutAHursizli...11
. loyaltyand tv koyai...11
. landlordsinsolvent to exampib an...11
. ELECTIONPARIS THE OF MORAL...11
. gy...11
Open Council...12
Untitled Article...12
Component Actions
Send by E-mail...
Copy link to clipboard
Add to favorites...
Show Metadata...
File
View
12 of 24
23%
Home
Help
About Olive
Copyright © 2000-2005. Olive Software Inc. All Rights Reserved.
Transferring data from 208.42.237.25...
start
17:50

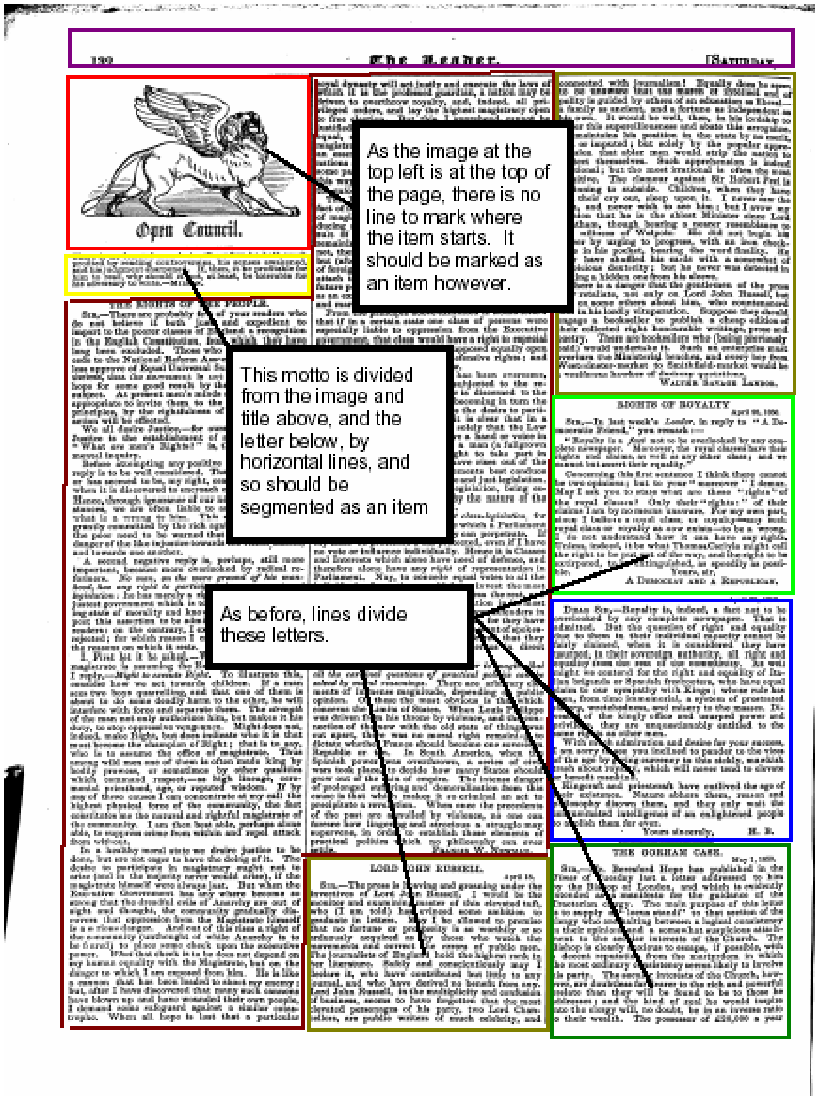
1.3.3. Screenshot from Manual Editing 1 that shows correct identification of an item at the top of a page.

In this respect, this segmentation policy works very well. However, on this page the item in the screenshot below is incorrectly highlighted, and the motto (immediately beneath the image) and the first letter have not been segmented.


http://208.42.237.25 - Kings College - Mozilla Firefox
Home
Library
Viewer
Search Results
Table of Components
. MISCELLANEOUS...7
Untitled Article...8
. . 27 . April ay c batubd . POSTSCI...9
—:hali-Jiveterin d helbe to mostly...9
Untitled Article...9
Public Affairs...10
. AqTTPn SUFFRAGET * axmmTn Ul...10
. 8uuainuuiiVJiAL AL JiuutAHursizli...11
. loyaltyand tv koyai...11
. landlordsinsolvent to exampib an...11
. ELECTIONPARIS THE OF MORAL...11
. gy...11
Open Council...12
Untitled Article...12
185028April ROYALTY OF RIGHTS...12
. 1850.1May ' .CASEGORHAM THE...12
18 . RUSSELLJOHN LORD...12
Literature...13
Untitled Article...13
Untitled Article...13
Untitled Article...14
Untitled Article...14
Untitled Article...14
Component Actions
Send by E-mail...
Copy link to clipboard
Add to favorites...
Show Metadata...
File
View
12 of 24
24%
Home
Help
About Olive
Copyright © 2000-2005. Olive Software Inc. All Rights Reserved.
Read 208.42.237.25
start
17:56

1.3.4 Screenshot from Manual Editing 1 that shows incorrect highlighting, and does not segment the motto and the first letter.

Below is a page taken from the pdf documents accompanying this guide. As you can see, a segmentation policy segmenting items according to horizontal lines would result in the following:



130 The Leader. [Saturday,

Open Council.

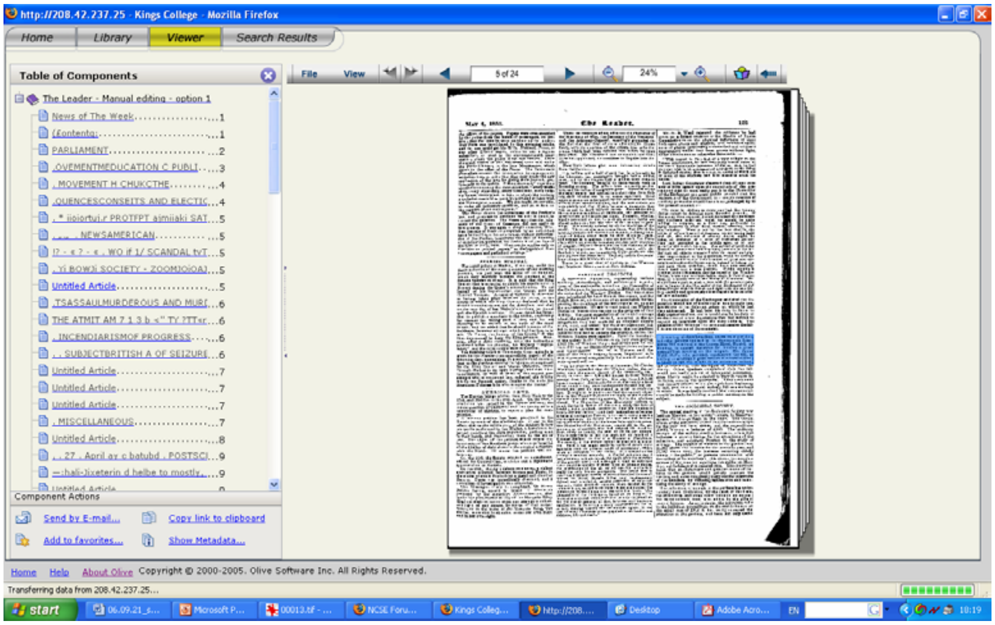
As the image at the top left is at the top of the page, there is no line to mark where the item starts. It should be marked as an item however.

This motto is divided from the image and title above, and the letter below, by horizontal lines, and so should be segmented as an item

As before, lines divide these letters.

1.3.5. Correct segmentation of this page, according to horizontal lines.

Notice with the above segmentation we have included the image as well as the text, and have divided items strictly according to the presence of horizontal lines. We are not sure under what segmentation policy the pilot was segmented, but figure 1.3.6 shows that there are existing strategies for those items that are not indicated by the presence of headlines. The horizontal line segmentation policy provides a much more accurate segmentation of items, and avoids the problem of what to do with untitled articles.


http://208.42.237.25 - Kings College - Mozilla Firefox
Home
Library
Viewer
Search Results
Table of Components
The Leader - Manual editing - option 1
News of The Week.....1
(Eontentg:.....1
PARLIAMENT.....2
.OVEMENTMEDUCATION C PUBLI.....3
. MOVEMENT H CHUKCTHE.....4
.QUENCESCONSEITS AND ELECTIC...4
. * iioiortui.r PROTFPT ajmiiaki SAT...5
. . . . NEWSAMERICAN.....5
I? - ε ? - ε . WO if 1/ SCANDAL tvT...5
. Yi BOWJi SOCIETY - ZOOMJOiOAJ...5
Untitled Article.....5
.TSASSAULMURDEROUS AND MURI...6
THE ATMIT AM 7 1 3 b <" TY ?TT«r...6
. INCENDIARISMOF PROGRESS.....6
. . SUBJECTBRITISH A OF SEIZURE...6
Untitled Article.....7
Untitled Article.....7
Untitled Article.....7
. MISCELLANEOUS.....7
Untitled Article.....8
. . 27 . April ay c batubd . POSTSCI...9
=:hali-Jixeterin d helbe to mostly....9
Component Actions
Send by E-mail...
Copy link to clipboard
Add to favorites...
Show Metadata...
File
View
5 of 24
24%
Home
Help
About Olive
Copyright © 2000-2005. Olive Software Inc. All Rights Reserved.
Transferring data from 208.42.237.25...
start

1.3.6. Screenshot from Manual Editing Option 1 showing segmentation of item even though it is not distinguished by a headline.

2.0 Important Notes

It is important for us that segmentation is accurate, and that all content on a page appears in a segment. Although we need segmentation to provide text strings to appear in the ToC, not all segments need appear there. For instance, we would like to segment the masthead on the first page, the image in the examples above, and the datelines at the top of every page etc… All these features need not appear in the ToC, but even if they do, it does not matter. We intend to use the uncorrected ToC as generated through Olive processing in order to generate an edited ToC for users, so the more things that appear there, the better.

3.0 Priorities

Our priorities in determining this segmentation policy are (in order of importance):

1. Attempt to segment all items on a page.
2. Ensure that each item is in its own segment, and not split into two etc…
3. Ensure that items that run from one column to another, or even from one page to another, are segmented as a single item.

As mentioned, for the further work we intend to on **ncse**, it is important that all items – whether articles, large headlines, or other non-textual features – are segmented.

Jim Mussell and Suzanne Paylor.