

ProQuest Historical Newspapers: Page-level and Article-level Compared

Background

When the ProQuest Historical Newspapers program was launched, all newspapers were digitized at the article-level. This means that every page of newspaper content was “zoned” into its distinct articles and other component parts (editorials, advertisements, cartoons, etc.), and each of those component parts was then run through OCR and treated as an individual entity in the database. Beginning in 2016, some historical newspaper titles in the ProQuest Historical Newspapers program were digitized at the page-level. This means that the full-page image is run through OCR, and the full page of content is stored in its entirety in the database. This document examines the differences between article-level and page-level digitization in searching, search results, and content display.



Article-level Zoning



High-Resolution Page-level Image

Differences in Searching

The basic searching function is identical for both article-level and page-level. **Every part of every page of ProQuest Historical Newspapers is full-text searchable, whether they are digitized at the article-level or page-level.** If you search for a term such as the name “John Kennedy” and it appears in the OCR text any place on a page, it will generate a hit for that page—whether it is in an

article title, in an article, in an advertisement, etc. The primary difference in searching article-level titles is in the Advanced Search: because they include article-level metadata, newspapers digitized at the article-level provide users with the ability to restrict search results to different portions of the newspaper (articles, advertisements, cartoons, etc.).

Differences in Search Results

The search results interface has been designed so both article-level and page-level results are presented together in an integrated and intuitive way. Most users are unaware that there is any difference between the following results:

The screenshot displays the ProQuest search results interface for the query "john kennedy". The search bar at the top shows the query and a magnifying glass icon. Below the search bar, the results count is "120,973 results". To the right of the count are links for "Modify search", "Recent searches", and "Save search/alert". Below these are icons for "Cite", "Email", "Print", and "Save".

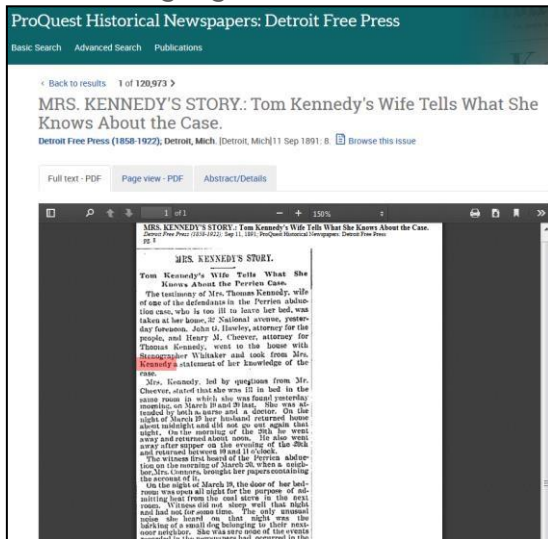
On the left side, there is a "Narrow results" section with a "Publication date" filter. It shows a bar chart for the years 1831-1999 (decades) and an "Update" button. The main results area shows three results:

- Result 1: "MRS. KENNEDY'S STORY.: Tom Kennedy's Wife Tells What She Knows About the Case. Detroit Free Press (1858-1922); Detroit, Mich. [Detroit, Mich]11 Sep 1891: 8. ...of Mrs. Thomas Kennedy, wife of one of the defendants ...avenue, yesterday forenoon. John G. Hawley, attorney for the people, and Abstract/Details Full text - PDF (44 KB) Preview"
- Result 2: "July 19, 1999 (Page 15 of 139) Detroit Free Press (1923-1999), General edition; Detroit [Detroit]19 July 1999: 15. ...John F. Kennedy Jr. had been lost in a plane crash. He was the one with the ...was a sleepy beach town in 'the 1950s. 1 John F. Kennedy Jr.'s grand- ...a, I. In 1963, President John F. Kennedy and a Naval aide help a Page view - PDF Preview"
- Result 3: "July 24, 1999 (Page 12 of 84) Detroit Free Press (1923-1999), General edition; Detroit [Detroit]24 July 1999: 12. ...Edward Kennedy, D-Mass., uncle of John F. Kennedy Jr.: Thank you, President ...to think, in that other Irish phrase, that this John Kennedy would live to ...up. "We dared to think ... that this John Kennedy would live to comb Page view - PDF Preview"

In the example above, the first result is article-level title, and the second and third results are page-level. In each case the user is presented with context and clues that help them quickly determine their interest level in the result. The search text is highlighted to show the keywords in context. The difference is that the article-level result includes the article name, while the page-level result includes the issue date and page.

Differences in Content Display

When a user selects an article-level search result, that article image is displayed with the search term highlighted. When a user selects a page-level result, the entire page of content is displayed with the search term highlighted¹:



Article-level Result



Page-level Result

The article-level result is a bitonal (black and white) image displayed at 300dpi (dots per inch). The page-level result is a high-resolution greyscale image displayed at 400dpi, which provides almost photographic-like quality of the microfilm images. In both cases the images may be scrolled and zoomed as needed, and saved to the user's local storage.

¹ Hit-term highlighting currently requires that users have the Adobe PDF plug-in installed as the default PDF viewer for their browser. Firefox and Internet Explorer 8 have the Adobe PDF plug-in installed by default. This link describes the use of the Adobe Plug-in with various browsers: <https://helpx.adobe.com/acrobat/using/display-pdf-in-browser.html>

Enhanced Page-Level Viewing Interface

The page-level interface provides an enhanced browsing experience. When a user selects "Browse this Issue" they are presented with a new newspaper browsing interface optimized for page-level newspaper content. This interface features a scrolling thumbnail section at the bottom of the screen that enables users to quickly skim through a newspaper issue, and a highly intuitive interface that allows a user to manipulate the page image:

The screenshot displays a web browser window with the address bar showing a ProQuest URL. The page is titled "ProQuest Historical Newspapers: Detroit Free Press" and "Page 15". The main article is titled "Saddest words: It might have been" by Ron Dzwonkowski. The article text is visible, discussing John F. Kennedy Jr. and his family. A large photograph of John F. Kennedy Jr. and his wife Carolyn Bessette Kennedy is shown on the right side of the article. Below the article, there is a scrolling thumbnail section labeled "FRIENDS" and "QUOTABLE". The interface includes various navigation and manipulation controls at the bottom of the page image.

Searchable PDF Images

The last difference between article-level and page-level digitization is with the image itself. The page-level images are all searchable PDF files. The downloaded PDF image can be searched using the PDF Reader search function and the OCR text can be copy/pasted into other documents:



He was the one with the potential to atone for all the troubled Kennedys, the one who handled all the Kennedy attention with the most grace, the one who seemed capable of even becoming the president that history revealed his father never really was and never had a chance to be.



Conclusion

While there are differences between the article-level and page-level digitization process, most users will not notice any difference in searching or search result interpretation. All portions of every page are full-text searchable regardless of the treatment. The page-level images are high-resolution 400dpi greyscale searchable PDF files that offer end users exciting new capabilities.

Learn more about Historical Newspapers

Visit <https://about.proquest.com/en/products-services/pq-hist-news/>