Archive Search

This document will describe search behavior as of June 2016 when searching a historic archive of newspapers or magazines made available to readers using Visiolink's solutions and search options. This includes a walk-through of search query possibilities and the expected outcome of such searches.

1 Archive material and search quality

After pages are scanned and saved digitally, an OCR (Optical Character Recognition) text extraction is performed, after which they are imported into Visiolink's system. The quality (physical state, storage method) of the physical archive material varies, and the quality of the material will affect quality of the OCR and the search performance in the Visiolink solution. Different OCR scan methods from different vendors will also impact search behavior.

PDF files of current publications that are processed as part of a daily flow are also search indexed. This succeeds without use of OCR scanning. Therefore, differences in search behavior when searching in an archive of both historic and new publications can occur from old to new content.

Description of search behavior in this document states how search functionality is programmed and on what basis it is tested by Visiolink.

2 Solr search

When searching in a historic archive through the Visiolink search API, in the HTML Desktop WebApp or from the starting page/navigation drawer of the app, a search via an intelligent and powerful Solr search engine in the OCR material of the archive and the indexed PDF pages of current (non-archive) publications is performed. Following a walk-through of search behavior for searches performed through Solr. Please be aware that the Solr search mechanism is detached from the highlight mechanism in the app/web reader. The highlight mechanism is described separately.

2.1 General notes on Solr searching

- The Solr search index works from the concept of pages, not articles. This means that when a search is performed, results will be returned with a page number and a publication ID.
- Searches are performed online, thus internet connection is required in order to search.
- Words consisting of less than three characters are not indexed
    - Searches consisting of less than three characters will not be performed.
    - Words of less than three characters contained in a search phrase will not be part of the search
- In each language a list of stopwords[1] are not indexed. Words like *the, he, she, but, for* in English and *er, sie, es, ist, für, und* in German are not possible to find.
- Hyphenated words continuing from one column or line to the next are indexed as they are written in the publication. This means that the hyphenated word *horse-radish* is indexed as the two words *horse-* (including hyphen) and *radish*

2.2 Unquoted search

[1] See the full list of stopwords in Danish, Norwegian, German and English in appendix 1, page 5

Unquoted searches are more comprehensive and less precise than quoted searches, which will be expressed in the return of a big number of results if search term is too general. Unquoted searches are case insensitive.

2.2.1 Single unquoted word search

A single unquoted word search will perform a prefixed search in order to return both the exact word and also other forms of the word, e.g. plural.

Examples:

- Searching for the word *horse*, will return pages containing *horse, horses, horseradish, horseshoe, horsemen* etc.
- Words that change endings in plural will be found. A search for the word *city* will also return pages containing the plural form *cities,* or searching for the German word *Anzug* will also return pages containing the plural form *Anzüge/Anzügen*

2.2.2 Multiple unquoted word search

Searching for multiple words unquoted will perform a prefixed AND-search. This means that results returned are pages that contain all the words searched for in some form, e.g. plural.

Examples:

- Searching for *the mansion on the hill* will return pages containing a form of all unique words that consist of more than two letters and are not stopwords, in this case:
    - *mansion, mansions* and
    - *hill, hillside*, *hillary*

2.3 Quoted search

Performing quoted searches allows for searching exact expressions or phrases. For quoted searches, case insensitivity is simulated, which will affect some searches, see example below.

2.3.1 Single quoted word search

A single quoted word search performs an exact search for the word with case insensitivity.

Examples:

- A search for the word *"horse"* will return pages containing the exact word *horse*, including the two case versions *Horse* and *HORSE*

2.3.2 Multiple quoted word search

A multiple quoted word search performs an exact case insensitive search returning pages with all words present

Examples:

- A search for “horse” “whisperer” “petrol” will return pages containing all of the exact words, including the two case versions Horse, Whisperer, Petrol and HORSE, WHISPERER, PETROL

2.3.3 Quoted phrase search

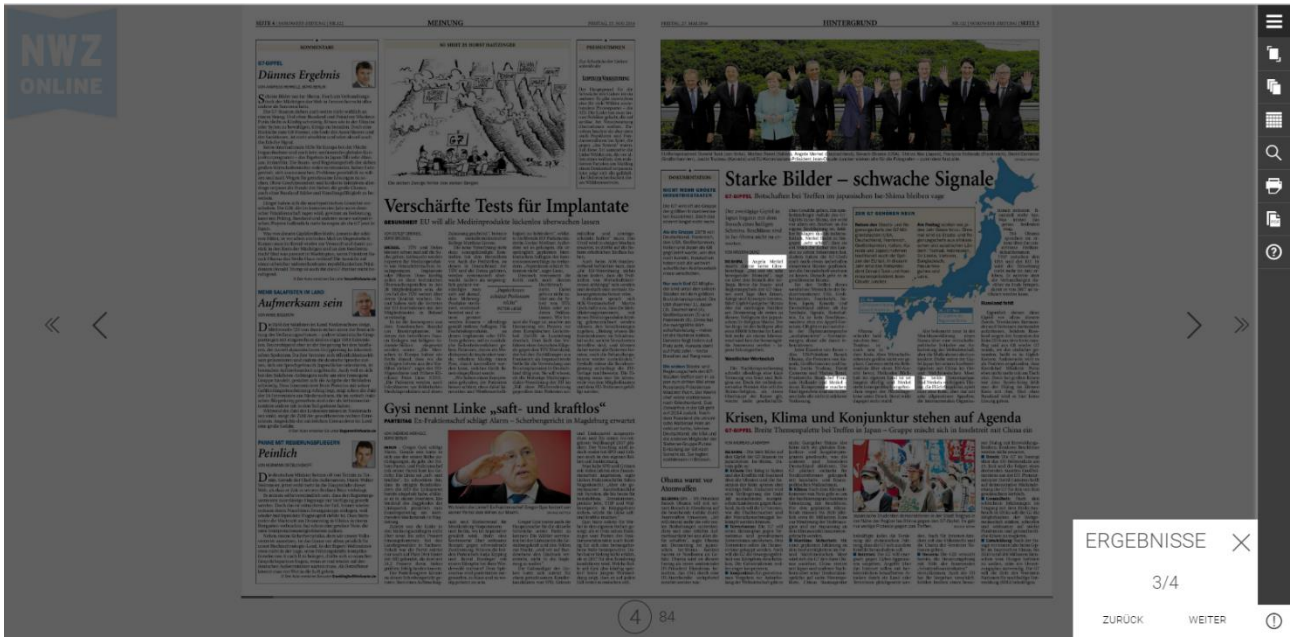
A quoted phrase search performs an exact search for the phrase with case insensitivity to some extent.

Examples:

- Searching the phrase “tony blair” will return pages containing the exact phrase *tony blair*, including the two case versions *Tony Blair* and *TONY BLAIR*
- Searching the phrase *“the day he left office”* will return the exact phrase “*the day he left office”*, including the two case versions “*The Day He Left Office”* and “*THE DAY HE LEFT OFFICE”*
- Searching the phrase *“the day tony blair left office”* will probably not return results, as it would search for the exact phase “*the day tony blair left office”* and the two case versions “*The Day Tony Blair Left Office”* and “*THE DAY TONY BLAIR LEFT OFFICE”*. The real phrase in this case would most likely consist of both capitalized and non-capitalized words: “*the day Tony Blair left office”.*

3 Highlights in the app/web

When a search in the Solr-index is performed, a list of pages containing results is presented in the reader (or returned via the search API for your presentation). When the user selects a page, technically a new search is performed in a highlight index that contains all words of the page (minus words less than two characters and stopwords) including geometric coordinates of word location on the page. After selecting a page containing search results, the publication will open on the selected page. An overlay is active and all occurrences of the searched word(s) are highlighted, see illustration below.



If more pages of the opened publication contain results, the user can jump between those pages by swiping right/left (apps) or pressing “forward/back” in the bottom right corner (web). The search overlay can be dismissed and the publication appears in normal reading mode.

3.1 General notes on highlight search presentation

- The highlight index is based on individual words and has no concept of phrases, which will impact the page highlighting presented to the user. This will be apparent from the walkthrough below.
- The highlight mechanism aims on making sure that the searched word(s) are highlighted on the page. This means that even exact quoted phrase searches can cause highlights to appear that are not part of the phrase.
- If a Solr search returns results that cannot be found in the search highlight index, the right page will be shown, but without highlights.

3.2 Highlight behavior

Following a walkthrough including examples of cases that describes the highlight behavior where it differs from the Solr search.

- Words that change endings in plural will not be highlighted. Even if the Solr search has found the plural form *cities* from a search for the singular form *city,* the highlight mechanism is not able to make the same link between the two forms and will only highlight the originally searched word. When this happens the right page will open without highlights (unless the originally searched word or a form of the word that does not change the core word – e.g. *cityscape* – is present somewhere else on the page).
- When highlighting a quoted phrase, it will be handled like a multiple single quoted word search. This means that searching for “the day Tony Blair left office” will be highlighted like this:
    - the is a stopword and will be sorted out
    - “day”, “Tony”, “Blair”, “left” and “office” are handled as single quoted words

  This means that false positives can occur as more words than the ones making out a coherent phrase will be highlighted. Normally, the searched phrase is easy to spot on the page because highlights will occur close together.

Appendix 1: Stopwords

English

"about","above","across","after","afterwards","against","all","almost","alone","along","already", "also","although","always","am","among","amongst","amoungst","amount","an","and","another","any","anyhow","anyone","anything","anyway","anywhere","are","around","as","at","back","be","became","because","become","becomes","becoming","been","before","beforehand","behind","being","below","beside","besides","between","beyond","both","bottom","but","by","call","can","cannot","cant","co","con","could","couldnt","cry","de","describe","detail","do","done","down","due","during","each","eg","eight","either","eleven","else","elsewhere","empty","enough","etc","even","ever","every","everyone","everything","everywhere","except","few", "fifteen","fify","fill","find","first","five","for","former","formerly","forty","found","four","from","front", "full","further","get","give","go","had","has","hasnt","have","he","hence","her","here","hereafter","hereby", "herein","hereupon","hers","herself","him","himself","his","how","however","hundred","ie","if","in","inc", "indeed","interest","into","is","it","its","itself","keep","last","latter","latterly","least","less","ltd", "made","many","may","me","meanwhile","might","mill","mine","more","moreover","most","mostly","move","much", "must","my","myself","name","namely","neither","never","nevertheless","next","nine","no","nobody","none","noone", "nor","not","nothing","now","nowhere","of","off","often","on","once", "one","only","onto","or","other","others", "otherwise","our","ours","ourselves","out","over","own","part", "per","perhaps","please","put","rather","re", "same","see","seem","seemed","seeming","seems","serious", "several","she","should","show","side","since", "sincere","six","sixty","so","some","somehow","someone", "something","sometime","sometimes","somewhere","still", "such","system","take","ten","than","that","the","their","them","themselves","then","thence","there", "thereafter","thereby","therefore","therein","thereupon","these","they","thick","thin","third","this","those", "though","three","through","throughout","thru","thus","to","together","too","top","toward","towards","twelve", "twenty","two","un","under","until","up","upon","us","very","via","was","we","well","were","what","whatever", "when","whence","whenever","where","whereafter","whereas","whereby","wherein","whereupon","wherever","whether", "which","while", "whither","who","whoever","whole","whom","whose","why","will","with","within","without","would", "yet","you","your","yours","yourself","yourselves"

Danish

"af","alle","andet","andre","at","begge","da","de","den","denne","der","deres","det","dette","dig","din","dog", "du","ej","eller","en","end","ene","eneste","enhver","et","fem","fire","flere","fleste","for","fordi","forrige", "fra","få","før","god","han","hans","har","hendes","her","hun","hvad","hvem","hver","hvilken","hvis","hvor", "hvordan","hvorfor","hvornår","ikke","ind","ingen","intet","jeg","jeres","kan","kom","kommer","lav","lidt","lille", "man","mand","mange","med","meget","men","mens","mere","mig","ned","ni","nogen","noget","ny","nyt","nær","næste", "næsten","og","op","otte","over","på","er","kun","se","seks","ses","som","stor","store","syv","ti","til","to", "tre","ud","var", "vil", "vi"

Norwegian

"alle","andre","arbeid","av","begge","bort","bra","bruke","da","denne","der","deres","det","din","disse","du", "eller","en","ene","eneste","enhver","enn","er","et","folk","for","fordi","fra","gjorde","god","ha","hadde","han", "hans","hennes","her","hva","hvem","hver","hvilken","hvis","hvor","hvordan","hvorfor","jeg",

"ikke","inn","innen","kan", "kunne","lage","lang","lik","like","makt","mange","med","meg","meget","men", "mens","mer","mest","min","mye","må", "måte","navn","nei","ny","og","også","om","opp","oss","over","part","punkt","på","rett","riktig","samme","sant", "si","siden","sist","skulle","slik","slutt","som","start","stille","så","tid","til","tilbake","tilstand","under", "ut","uten","var","ved","verdi","vi","vil","ville","vite","vår","være","vört"

German

"aber","als","am","an","auch","auf","aus","bei","bin","bis","bist","da","dadurch","daher","darum","das","daß", "dass","dein","deine","dem","den","der","des","dessen","deshalb","die","dies","dieser","dieses","doch","dort", "du","durch","ein","eine","einem","einen","einer","eines","er","es","euer","eure","für","hatte","hatten", "hattest","hattet","hier","hinter","ich","ihr","ihre","im","in","ist","ja","jede","jedem","jeden","jeder", "jedes","jener","jenes","jetzt","kann","kannst","können","könnt","machen","mein","meine","mit","muß","mußt", "musst","müssen","müßt","nach","nachdem","nein","nicht","nun","oder","seid","sein","seine", "sich","sie","sind", "soll","sollen","sollst","sollt","sonst","soweit","sowie","und","unser","unsere","unter","vom","von","vor","wann", "warum","was","weiter","weitere","wenn","wer","werde","werden","werdet","weshalb","wie","wieder","wieso","wir", "wird","wirst","wo","woher","wohin","zu","zum","zur","über"