

Technical Bulletin #22

This technical bulletin describes how to use the **style.lex** file to permit searching for words containing non-alphanumeric characters.

Issue

By default, Astoria's search functionality does not allow you to search for words that contain non-alphanumeric characters (e.g., &, /, and ") as search criteria; however, users may need to specify these characters as part of their search criteria.

Resolution

You can create a **style.lex** file to override the system defaults and include those non-alphanumeric characters for which you want to be able to search. Entries in this file identify the patterns that the search engine interprets as valid characters in words.

The collection word index is created based on the specifications made in the **style.lex** file, when this file is present in the default **style** directories (c:\astoria\verity\common\style on Windows and /opt/astoria/verity/common/style on Solaris). As a result, the words stored in the index can contain the non-alphanumeric characters specified in the file. For example, if the non-alphanumeric character "/" is recognized as a valid character, users can search for the word "OS/2".

After you edit the default **style.lex** file as described in the following section, you can then run the Reindex command from Database Administrator on all cabinets. When a cabinet is reindexed, the existing cab directory is deleted (e.g., c:\astoria\database\cab00000.db.col). The new cab directory is then populated with the contents of the default style directory (among other things).

Sample style.lex file

This sample **style.lex** file has been created to accommodate the following characters as search criteria: forward slash (/) to accommodate words such as OS/2, and ampersand (&) to accommodate words such as AT&T, and apostrophe (') for possessive words (in English).

Edit the default **style.lex** file to contain the following syntax:

```
$control: 1
lex:
{
  define: WHT      "[ \t]"          #whitespace macro
  define: NL       "{WHT}*\n"      #newline macro

  token:  WORD    "[A-Za-z0-9/&' ]+" #word
  token:  WORD    "[0-9]+\.\.[0-9]+" #word
  token:  EOS     "[.?!]"          #end of sentence
  token:  NEWLINE "{NL}"          #single end-of-line
  token:  PARA    "{NL}{NL}"      #end of paragraph
  token:  WHITE   "{WHT}"         #whitespace
  token:  PUNCT   "."             #all other text
}
$$
```

Your **style.lex** file must specify token statements for all the tokens you want the search engine to match. The first "token: WORD" statement identifies the three non-alphanumeric characters to be included (/ , & , and '). Customize your **style.lex** file to include any additional characters for which you would like to search.

After updating your **style.lex** file, test the file by creating a cabinet containing only a small XML test document. Edit the **style.lex** file, reindex the cabinet, and confirm that your searches work as expected.