NAME

swish++.index − SWISH++ index file format

SYNOPSIS

```
long            num_words;
off_t           word_offset[ num_words ];
long            num_stop_words;
off_t           stop_word_offset[ num_stop_words ];
long            num_files;
off_t           file_offset[ num_files ];
long            num_meta_names;
off_t           meta_name_offset[ num_meta_names ];
                word index
                stop-word index
                file index
                meta-name index
```

DESCRIPTION

The index file format used by SWISH++ is as shown above. Every `word_offset` is an offset into the *word index* pointing at the first character of a word entry; similarly, every `stop_word_offset` is an offset into the *stop-word index* pointing at the first character of a stop-word entry; similarly, every `file_offset` is an offset into the *file index* pointing at the first character of a file entry; finally, every `meta_name_offset` is an offset into the *mete-name index* pointing at the first character of a meta-name entry.

The index file is written as it is so that it can be mapped into memory via the **mmap**(2) Unix system call enabling ‘‘instantaneous’’ access.

Word Entries

Every word entry in the *word index* is of the form:

word`0`{*data*}**...**`0`

that is: a null-terminated word followed by one or more *data* entries followed by a null byte where a *data* entry is:

*file-index***[**`2`*meta-ID***]...**`1`*rank*`1`

that is: a file-index followed by zero or more meta-IDs (each preceeded by the ASCII 2 character) followed by a rank (both preceeded and followed by the ASCII 1 character). The integers are in ASCII, not binary. Currently, only ASCII characters 1 and 2 (the binary values 1 and 2, not the digits `1` and `2` having ASCII codes of 49 and 50 decimal, respectively) are used. (Other low ASCII characters are reserved for future use.) The *file-index* is an index into the `file_offset` table; the *meta-IDs*, if present, are unique integers identifying which meta name(s) a word is associated with in the meta-name index.

Stop-Word Entries

Every stop-word entry in the *stop-word index* is of the form:

stop-word`0`

that is: every word is null-terminated.

File Entries

Every file entry in the *file index* is of the form:

path-name file-size file-title`0`

that is: the pathname for a file relative to where the indexing was performed (unless absolute paths were used) followed by the file's size in bytes followed by the file's title followed by a null byte. All the information is in ASCII.

For an HTML file, the title is what is between `<TITLE>` ... `</TITLE>` pairs. If a file is not an HTML file, or is but does not have a title, the title is simply the file (not path) name.

Meta-Name Entries

Every meta-name entry in the *meta-name index* is of the form:

meta-name`\0`

that is: every word is null-terminated.

CAVEAT

Generated index files are machine-dependent (size of data types and byte-order).

SEE ALSO

index(1), **search**(1)

AUTHOR

Paul J. Lucas *<pjl@best.com>*