Test- and validation corpus

Test corpus

In order to develop the software, we agreed upon a set of articles that are freely available under a `CC-BY-NC-ND` license or less restrictively licensed. This facilitated open cooperation and development of the software via Github. We used the test corpus not to test the algorithms in the traditional machine learning sense, but in order to develop the software and work out problems in the process of extracting a first set of tables (i.e., also called development corpus). The validation set will then be used to estimate the retrieval rate of various characteristics such as pubstyle identification, header recognition, column identification, character stream recognition (in that order, see also this Github issue).

```
test <- read.csv('../data/metadata.csv')
# Print number of papers per publisher
table(test$publisher)
```

```
##
##              BMC              BMJ         Elsevier           Nature
##                6                5               14                5
##             PLOS         Springer Taylor & Francis            Wiley
##                5                4                6                6
##   Wolters Kluwer
##                3
```

```
write.csv(table(test$journal, test$publisher), '../data/test-pubs-jrnls.csv')
```

In order to catch multiple PubStyles, we included 26 journals from 9 publishers. Not all journals by one publisher necessarily include the same PubStyle because housestyles sometimes vary within a publisher (note: PubStyles can also vary over time, when the layout of a journal changes). The amount of tables per journal is available in the `data/test-pubs-jrnls.csv` file.

Validation corpus