# Data

## Retrieving data

Data is being loaded from Event Registry (ER) using its Python based API. A class ERReader has been crated, that supports the following methods:

* \_\_init\_\_ (perpage, days)  
  *Setups class variables, reads username/password from settings.json file (corresponding to ER specifications), connects to Event Registry and logs in. Sets how many articles per page will be retrieved and how long will be the maximum retrieval interval.*
* get\_articles\_period(concept, startdate, enddate)

*Retrieves articles for a particular concept and for particular period. Uses get\_articles.*

* get\_articles(concept, startdate, enddate)  
  *Checks available articles for a particular concept between the dates. Starts loading of articles by pages using get\_articles\_page.*
* get\_articles\_page(concept, startdate, enddate, pagenum)

*Loads a particular page of articles for defined concept and start and end date.*

* print\_articles(list)  
  *Prints list of dates and times of articles to the stadard output.*
* save\_articles(file\_name)  
  *Saves retrieved articles to a JSON file. Articles are stored in ER format[[1]](#footnote-1), one article per line.*

## Test data

The articles (news) for the following concepts have been retrieved between 2014-01-01 and 2017-01-01.

* Borut Pahor (in Slovene), 8410 instances, 49Mb
* Borut Pahor (in English), 3160 instances, 17Mb
* Peter Prevc (in Slovene), 3127 instances, 16Mb
* Peter Prevc (in English), 742 instances, 3,5Mb
* Microsoft (in English), 155576 instances, 1.3Gb
* European Comission (in English), 221736 instances, 1.9Gb

1. <https://github.com/gregorleban/EventRegistry/wiki/Data-models#article-data-model> [↑](#footnote-ref-1)