ELO Textreader (gen. 2)

ELO Textreader (gen. 2)

Table of contents

Textreader (gen. 2)

General

ELO Textreader (gen. 2) extracts text from documents of specific formats, which is the basis for the full text search.

It also coverts documents into PDF format. The PDF files are displayed in ELO clients as document previews.

Another task is to provide OCR for the ELO Indexserver.

> **Information**
>
> ELO Textreader (gen. 2) is a new product and replaces the previous ELO Textreader (ELOtr), ELO Preview Converter (ELOpreview), and ELO OCR Service (ELOocr) starting with version ELO 21.4.

> **Please note**
>
> To allow ELO Textreader (gen. 2) to process documents containing special Microsoft fonts in a Linux operating system, you must install these fonts in advance. The installation process varies according to the Linux distribution you are using. You can find more information about this on the Internet if you search for the keyword *ttf-mscorefonts*.

Architecture

Standalone application

ELO Textreader (gen. 2) starts as a standalone application and does not require a Tomcat server. The ELO Server Setup installs it as a service in the operating system. It can also be executed from the command line via Java code. You will find a corresponding shell script in the *bin* folder of the download package.

Cluster capacity

Multiple instances of ELO Textreader (gen. 2) can be clustered with a load balancer. The load balancer must ensure that requests for a specific conversion job reach the same instance, so you need to configure it for *sticky sessions*. The ELO Indexserver assumes that the load balancer implements *sticky sessions* with HTTP cookies. It remembers the cookies transmitted from the server and sends them in all requests belonging to the conversion job.

This is a configuration example of a load balancer with the Apache2 web server:

```
<VirtualHost localhost:80>
        ServerName loadbalancer

        Header add Set-Cookie "ROUTEID=.%{BALANCER_WORKER_ROUTE}e; path=/" env=BALANCER_ROU

        <Proxy balancer://cluster>
                BalancerMember http://server1:9060 route=TR1
                BalancerMember http://server2:9050 route=TR2
                ProxySet lbmethod=byrequests
                ProxySet stickysession=ROUTEID
        </Proxy>

        ProxyPass / balancer://cluster/
        ProxyPassReverse / balancer://cluster/

        LogLevel info
        CustomLog ${APACHE_LOG_DIR}/access-loadbalancer.log combined
        ErrorLog ${APACHE_LOG_DIR}/error-loadbalancer.log
</VirtualHost>
```

The Apache2 modules *proxy*, *proxy_http*, *proxy_balancer*, and *lbmethod_byrequests* must be enabled.

Supported file types

The file types listed below can be processed by ELO Textreader (gen. 2).

Full text processing

File type	Converter ID
7z	com.elo.tr2.conv.zip
7zip	com.elo.tr2.conv.zip
ai	com.elo.tr2.conv.aspose.psd
bmp	com.elo.tr2.conv.ocr
csv	com.elo.tr2.conv.aspose.excel
dcx	com.elo.tr2.conv.ocr
djv	com.elo.tr2.conv.ocr
djvu	com.elo.tr2.conv.ocr
doc	com.elo.tr2.conv.aspose.word
docx	com.elo.tr2.conv.aspose.word
dxl	com.elo.tr2.conv.zip
eml	com.elo.tr2.conv.eml

File type	Converter ID
gif	com.elo.tr2.conv.ocr
gz	com.elo.tr2.conv.zip
html	com.elo.tr2.conv.aspose.word
j2k	com.elo.tr2.conv.ocr
jar	com.elo.tr2.conv.zip
jb2	com.elo.tr2.conv.ocr
jfif	com.elo.tr2.conv.ocr
jp2	com.elo.tr2.conv.ocr
jpc	com.elo.tr2.conv.ocr
jpeg	com.elo.tr2.conv.ocr
jpg	com.elo.tr2.conv.ocr
mhtml	com.elo.tr2.conv.aspose.word
msg	com.elo.tr2.conv.msg
odt	com.elo.tr2.conv.aspose.word
pcx	com.elo.tr2.conv.ocr
pdf	com.elo.tr2.conv.pdf
png	com.elo.tr2.conv.ocr
ppt	com.elo.tr2.conv.aspose.ppt
pptx	com.elo.tr2.conv.aspose.ppt
rtf	com.elo.tr2.conv.aspose.word
tar	com.elo.tr2.conv.zip
tif	com.elo.tr2.conv.ocr
tiff	com.elo.tr2.conv.ocr
vsd	com.elo.tr2.conv.aspose.visio
vsdx	com.elo.tr2.conv.aspose.visio
war	com.elo.tr2.conv.zip
xls	com.elo.tr2.conv.aspose.excel
xlsx	com.elo.tr2.conv.aspose.excel
z	com.elo.tr2.conv.zip
zip	com.elo.tr2.conv.zip

Preview generation

File type	Converter ID
ai	com.elo.tr2.conv.aspose.psd
bmp	com.elo.tr2.conv.ocr
csv	com.elo.tr2.conv.aspose.excel
dcx	com.elo.tr2.conv.ocr

File type	Converter ID
djv	com.elo.tr2.conv.ocr
djvu	com.elo.tr2.conv.ocr
doc	com.elo.tr2.conv.aspose.word
docx	com.elo.tr2.conv.aspose.word
dxl	com.elo.tr2.conv.zip
eml	com.elo.tr2.conv.aspose.mail
gif	com.elo.tr2.conv.ocr
html	com.elo.tr2.conv.aspose.word
j2k	com.elo.tr2.conv.ocr
jb2	com.elo.tr2.conv.ocr
jfif	com.elo.tr2.conv.ocr
jp2	com.elo.tr2.conv.ocr
jpc	com.elo.tr2.conv.ocr
jpeg	com.elo.tr2.conv.ocr
jpg	com.elo.tr2.conv.ocr
mhtml	com.elo.tr2.conv.aspose.word
msg	com.elo.tr2.conv.aspose.mail
odt	com.elo.tr2.conv.aspose.word
pcx	com.elo.tr2.conv.ocr
pdf	com.elo.tr2.conv.pdf
png	com.elo.tr2.conv.ocr
ppt	com.elo.tr2.conv.aspose.ppt
pptx	com.elo.tr2.conv.aspose.ppt
rtf	com.elo.tr2.conv.aspose.word
tif	com.elo.tr2.conv.ocr
tiff	com.elo.tr2.conv.ocr
vsd	com.elo.tr2.conv.aspose.visio
vsdx	com.elo.tr2.conv.aspose.visio
xls	com.elo.tr2.conv.aspose.excel
xlsx	com.elo.tr2.conv.aspose.excel

Configuration

The configuration is done in a file consisting of name-value pairs called *config.properties* in the configuration directory of the ELO installation.

> **Important**
>
> Only ELO Textreader (gen. 2) and the responsible administrators may access the configuration file.

Resource requirements

ELO Textreader (gen. 2) consumes significant CPU time and hard disk space for conversion operations in addition to a large amount of RAM. Resource requirements can best be monitored and controlled when run in a separate environment (e.g. VM or Docker container). You should allocate several GB of RAM and several GB of disk space for production environments, especially because of the resource-intensive OCR process (see `com.elo.tr2.conv.ocr.workerCount`).

Configuration file example

```
http.port=8100
com.elo.tr2.textreader.instanceName=Instance1
com.elo.tr2.textreader.maxWorkers=10
com.elo.tr2.textreader.workDir=/tmp/textreader
com.elo.tr2.textreader.authUser=admin
com.elo.tr2.textreader.authPassword=elo
com.elo.tr2.textreader.authKeys=*
com.elo.tr2.textreader.fileType.pdf=com.elo.tr2.conv.pdf
com.elo.tr2.conv.ocr.minFileSizeKB=4
com.elo.tr2.conv.ocr.instDir=/opt/elo/prog/ocr10
com.elo.tr2.conv.ocr.workDir=/tmp/textreader/ocr-conv

com.elo.tr2.conv.ocr.workerCount=4
```

Comments

Comment lines start with a #.

Connection settings

> **Important**

ELO Textreader (gen. 2) accepts requests from all computers by default. This entails the risk of a DoS attack. With a firewall rule, you can allow requests from servers running an ELO Indexserver only.

URL of ELO Textreader (gen. 2) in the ELO Indexserver configuration

You must enter the URL of ELO Textreader (gen. 2) as the `textreaderUrl` option in the ELO Indexserver configuration. For more instructions, see the ELO Indexserver options page (in the ELO Application Server, select */ix-<Name> > Login > Configure Options > Help*).

```
_ALL | textreaderUrl | http://localhost:9060/tr
```

If multiple ELO Indexservers are active for the repository, they negotiate among themselves which one will perform conversion. If only one specific ELO Indexserver is to be responsible for this, enter the option value only for the instance name of this ELO Indexserver. However, this is only worthwhile if converting a large number of documents. It is not usually necessary to run an ELO Indexserver solely to manage the conversion process.

TCP port for HTTP

```
http.port=9060
```

ELO Textreader (gen. 2) accepts HTTP requests on this TCP port. If only encrypted connections should be allowed, leave out this value and only enter `https.port`.

TCP port for HTTPS

```
https.port=9063
```

ELO Textreader (gen. 2) accepts HTTPS requests on this TCP port. If only unencrypted connections should be used, leave out this value and enter `https.port`. If an HTTPS port is specified, the `https.keystore`, `https.keystoreType`, and `https.keystorePassword` settings must also be specified.

TCP host

```
http.host=localhost
```

Server name or server address on which ELO Textreader (gen. 2) is running and whose TCP ports accept connections. With the value `0.0.0.0`, ELO Textreader (gen. 2) listens to all server addresses. Default value: `localhost`.

Certificate file

```
https.keystore
```

File path to the certificate file. This is only required for HTTPS connections.

Type of certificate file

```
https.keystoreType
```

Type of certificate file: *PKCS12*, *PKCS11*, *JKS*. This is only required for HTTPS connections.

Password of certificate file

```
https.keystorePassword
```

The certificate file is protected with this password. This is only required for HTTPS connections.

Manage the conversion process

Instance name

```
com.elo.tr2.textreader.instanceName=Instance1
```

You can install multiple ELO Textreader instances by using a load balancer to group them under a single address. Each instance must have a unique name. The name should consist of the characters *A-Z*, *a-z*, and *0-9*.

> **Information**
>
> If you run a load balancer in front of multiple text readers, make sure that it is configured for *sticky sessions*.

Parallel conversion

```
com.elo.tr2.textreader.maxWorkers=10
```

It is possible to perform multiple conversion operations simultaneously. The more CPU cores the server has, the more parallel operations can be performed. Because the CPU pauses until the data is read or written, the number of operations can be higher than the number of CPU cores. Default value: 10

Working directory

```
com.elo.tr2.textreader.workDir=/tmp/textreader
```

Temporary files are stored in this directory. Default value: Temporary directory of the environment.

> **Important**
>
> Only ELO Textreader (gen. 2) needs full access in this directory. The directory must not be visible to users. No other programs need access to the directory.

> **Information**
>
> Directory separators have to be entered with a double backslash in Windows. Example: `C:\\ELO\\data\\textreader...`.

Minimum free disk space

```
com.elo.tr2.textreader.minFreeDiskSpaceMB=100
```

ELO Textreader (gen. 2) will abort processing if there are less than the specified amount of MB free in the working directory. Because built-in third-party components do not take this setting into account, the value should be chosen generously. Default value: 100 MB.

Maximum processing time

```
com.elo.tr2.textreader.maxFileProcessingSeconds=600
```

Processing of a document will be aborted if it takes longer than the specified number of seconds. For container documents, the specification also includes the processing time of all documents in the container. However, not all conversion operations in ELO Textreader (gen. 2) can be aborted after exactly this amount of time. Some processes must run at least until parts of them are completed. Default value: 600 s.

Wait time for result

```
com.elo.tr2.textreader.waitForResultSeconds=10
```

A conversion request waits the maximum specified number of seconds for the result. If it is available within this time, the result is returned. If the conversion takes longer, the ELO Indexserver receives a *handle* with which it can repeatedly request the result (*polling*). This request waits for the number of seconds specified here. Default value: 10 s.

User authentication for status page

```
com.elo.tr2.textreader.authUser=admin
com.elo.tr2.textreader.authPassword=geheim
```

These settings specify the authentication credentials for the status page. The status page opens in a reduced state by default. Additional information can only be viewed after logging on.

Authentication for client application

```
com.elo.tr2.textreader.authKeys=
```

The value of this configuration entry should be left empty if a license key is available (new as of ELO 21.4). Otherwise, you need to enter *.

The ELO Textreader (gen. 2) functions may only be used by ELO programs. This is due to licensing requirements of several commercial components that are integrated in ELO Textreader. Especially in the case of cloud installations, you should pay attention to the license terms if an ELO Textreader instance is made publicly available to multiple clients on the Internet.

This is why ELO Textreader (gen. 2) requires an authentication key in every request. If a license key is stored in the ELO installation, the public part of the license key is transmitted as authentication. The public part of the key is checked when the request is sent to the ELO Indexserver on *license.elo.com*.

For installations with a license file instead of a license key, ELO Textreader (gen. 2) must not be publicly accessible. It is the responsibility of the network administrator to ensure that only ELO programs can access the ELO Textreader, e.g. through an appropriate firewall configuration. Currently, only the ELO Indexserver requires access to the text reader.

Map file extension to converter

```
com.elo.tr2.textreader.fileType.<file-extension>=<converter-id>

com.elo.tr2.textreader.fileType.pdf=com.elo.tr2.conv.pdf
```

With the default settings, the converters provided in ELO Textreader (gen. 2) are already mapped to all file extensions that they can process. However, this mapping is not unique for PDF files. This setting can be used to specify which converter should be used to process the file type.

Check uploaded files for malicious content

```
com.elo.tr2.textreader.verifyFileUploadCommand=clamdscan --fdpass $1
```

You can execute a command line that invokes a virus scanner to check uploaded files for malicious content. In the call line, the placeholder `$1` specifies the position at which the uploaded file is inserted. The command must return the exit code `0` if the file is OK. Any other value will be read as an error.

Whitelist for file types

```
com.elo.tr2.textreader.allowedFileTypes=txt;pdf;docx;xlsx
```

If you only want certain file types to be processed, they can be specified as a semicolon separated list of file extensions or content types (mime types). If the setting is not made or is left empty, all file types are allowed.

Exclusion list for file types

```
com.elo.tr2.textreader.disallowedFileTypes=exe;bin;dat
```

If you do not want certain file types to be processed, they can be specified as a semicolon separated list of file extensions or content types (mime types). If the setting is not made or is left empty, all file types are allowed.

OCR converter settings

Maximum file size

```
com.elo.tr2.conv.ocr.maxFileSizeMB=10
```

Maximum size of a file sent to the OCR service. This is only relevant with `com.elo.tr2.conv.ocr.ocrUrls`. Default value: 10 MB.

OCR component installation directory

```
com.elo.tr2.conv.ocr.instDir=/opt/elo/prog/ocr10
```

If you want ELO Textreader (gen. 2) to load the OCR component, this value specifies the installation directory of the OCR component.

> **Information**
>
> Directory separators have to be entered with a double backslash in Windows. Example: `C:\\ELO\\prog\\ocr10`.

Working directory

```
com.elo.tr2.conv.ocr.workDir=/tmp/textreader/ocr-conv
```

This specifies the working directory where temporary files are stored. This is only relevant with `com.elo.tr2.conv.ocr.instDir`.

> **Information**
>
> Directory separators have to be entered with a double backslash in Windows. Example: `C:\\ELO\\data\\textreader\\ocr-conv`.

Number of OCR worker processes

```
com.elo.tr2.conv.ocr.workerCount=4
```

Number of worker processes that will be started for OCR. At least 1 GB must be provided by the environment. The number of workers multiplied by 400 MB RAM is added to this. This is only relevant with `com.elo.tr2.conv.ocr.instDir`.

Minimum size for documents

```
com.elo.tr2.conv.ocr.minFileSizeKB=5
```

OCR is only performed for files that might have relevant content based on their size. For example, OCR should not be performed on icons because they are not expected to contain text. Default value: 5 KB.

Timeout for a document page

```
com.elo.tr2.conv.ocr.pageTimeoutSeconds=600
```

The analysis of a page should not take longer than the specified number of seconds. This value is multiplied by the number of pages to calculate the maximum total duration. The analysis of a document will be aborted if it takes longer than the total duration. Default value: 600 s.

Timeout for a document

```
com.elo.tr2.conv.ocr.fileTimeoutSeconds=600
```

The analysis of a document should not take longer than the specified number of seconds. If the timeout is exceeded, the process is aborted. Default value: 600 s.

Foreground and background processing

```
com.elo.tr2.conv.all.reservedForForegroundPercent=60
```

Foreground and background tasks are managed with separate task pools. This value specifies how the `com.elo.tr2.textreader.maxWorkers` and `com.elo.tr2.conv.ocr.workerCount` are distributed. If you set 60, for example, 60 percent of these settings are reserved for background tasks. The default value 0 means that all tasks are treated equally as background tasks.

Settings for generating preview files

Disable preview generation

A preview is automatically created for each document added to the full text. If you do not want a preview to be generated, for example to save storage space, you will have to make the following setting in the ELO Indexserver options, and not in the ELO Textreader (gen. 2) configuration file:

Instance	Name	Value
_ALL	previewGenerationDisabled	true

> **Information**
>
> Previews are helpful for displaying documents in the ELO Web Client if they cannot be shown in the browser in their original format.

Mode of operation

The ELO Indexserver FT plug-in is responsible for finding the documents to be processed. It searches for new conversion jobs in the *vtdocs* control table of the repository database approximately every 10 seconds.

If multiple ELO Indexservers are active for the repository and are responsible according to the configuration, they communicate with one another so that only one server executes these requests.

The documents that are found are sent to ELO Textreader (gen. 2) via HTTP/S.

If ELO Textreader (gen. 2) delivers the text content for a document within the default value of 10 seconds, it is returned as a response in the HTTP/S call. Otherwise, the ELO Indexserver first receives a *handle* with which it repeatedly requests the text content at intervals of a few seconds.

A preview file in PDF format is generated along with the full text content. The ELO Indexserver stores both these files in the repository along with the document.

> **Information**
>
> A preview file is only generated for documents that have been added to the full text search.

The ELO Indexserver also calls ELO Textreader (gen. 2) to process OCR tasks. Therefore, a separate connection to the OCR service is not required.

> **Information**
>
> Unlike the previous ELO Textreader, the connection is established by the ELO Indexserver. This allows you to use ELO Textreader (gen. 2) in different installations without having to list the URLs of the ELO Indexservers in its configuration.

Conversion for foreground and background tasks

To prioritize conversion tasks triggered by user actions, ELO Textreader (gen. 2) manages separate task pools for foreground and background tasks. In the configuration, you can specify the allocation of total resources to the two task pools. In addition, foreground tasks are executed with a higher priority (Java thread priority) than the background tasks.

The ELO Indexserver treats queries from the ELO Java Client, ELO Web Client, ELO Desktop Client, or the ELO apps for iOS and Android as foreground tasks. This currently only relates to OCR tasks.

Service tasks (e.g. from ELO Automation Services) such as full text and preview processing, and OCR, are considered background tasks.

Converter

ELO Textreader (gen. 2) delegates tasks to internal converter modules, which are selected according to document type and target type (text content or PDF preview). Temporary files may be created in the working directory in the process. These are deleted after conversion.

If the document being converted additionally contains documents (e.g. e-mails with attachments), there is a child folder for each of the contained documents. The conversion process is only completed when all contained documents have been converted. Afterwards, the result can be retrieved from the ELO Indexserver.

Logging

With the default settings, ELO Textreader (gen. 2) writes three log files:

- *tr2.log*: Log of processing steps
- *tr2-ocr.log*: Log of OCR processing
- *tr2-report.log*: For each conversion job, one line with information about success or error

The log outputs are configured in the *logback.xml* file of the ELO Textreader (gen. 2) configuration directory. You will find information on the file syntax here.

Every time ELO Textreader (Gen. 2) is called, it sends unique call tags that identify all associated actions. Extract from *tr2.log*:

```
21:33:11.192 INFO  XNIO-2 task-1 IX1-1355106364585954729 (TextreaderImpl.java:141) - Begin
21:33:11.194 INFO  XNIO-2 task-1 IX1-1355106364585954729 (TextreaderImpl.java:141) - Begin
21:33:12.326 INFO  tr-conv-4 1355106364585954729-685511-1-FT-8710415920948495412 (ConvertJo
21:33:12.583 INFO  tr-conv-5 1355106364585954729-685511-2-PV-2160424150872155649 (ConvertJo
21:33:12.584 INFO  tr-conv-5 1355106364585954729-685511-2-PV-2160424150872155649 (report.ja
21:33:14.341 INFO  tr-conv-6 1355106364585954729-685511-1-FT-8710415920948495412 (ConvertJo
21:33:14.352 INFO  tr-conv-6 1355106364585954729-685511-1-FT-8710415920948495412 (report.ja
```

In this example, the call identifier `IX1-1355106364585954729` is transmitted for the HTTP request. It consists of the instance name of the ELO Indexserver `IX1` and the random number `1355106364585954729`. There are two conversion tasks in the call, tagged `1355106364585954729-685511-1-FT` and `1355106364585954729-685511-2-PV`. They consist of the random number from the HTTP request, the object ID of the document `685511`, a consecutive number `1` or `2`, and the abbreviation `FT` for full text and `PV` for preview.

You can also see from the extract that a document was sent in XLSX format and has 150639 bytes: `Begin convert [file-extension/xlsx, #150639]`.

The conversion to PDF format was completed after 1390 ms and returned a file size of 164841 bytes: `[1390] Converted=SUCCESS, in=[file-extension/xlsx, #150639], out=[application/pdf, #164841]`.

The text content was ready after 4145 ms and has 11678 bytes: `[3160] Converted=SUCCESS, in=[file-extension/xlsx, #150639], out=[text/plain; charset=utf-8, #11678].`

For a detailed analysis, you can filter out all lines from the logs relating to full text processing, for example. These lines contain the tag `1355106364585954729-685511-1-FT`.

```
21:33:11.192 INFO  XNIO-2 task-1 IX1-1355106364585954729 (TextreaderImpl.java:141) - Begin
21:33:12.326 INFO  tr-conv-4 1355106364585954729-685511-1-FT-8710415920948495412 (ConvertJo
21:33:14.341 INFO  tr-conv-6 1355106364585954729-685511-1-FT-8710415920948495412 (ConvertJo
21:33:14.352 INFO  tr-conv-6 1355106364585954729-685511-1-FT-8710415920948495412 (report.ja
```

Here, you can see that there is a PNG image in the XLSX document:

```
[2015] Job=SUCCESS, in=[file-extension/png, #168220, file=[/tmp/textreader/tr-work/2022-11-
```

In the *tr2-ocr.log*, a search for `1355106364585954729-685511-1-FT` finds the start and end of OCR processing for the PNG file.

```
21:33:12.327 INFO  tr-conv-6 1355106364585954729-685511-1-FT-8710415920948495412 (OcrServic
...
21:33:14.338 INFO  ocr-impl-17 1355106364585954729-685511-1-FT-8710415920948495412 (OcrConv
21:33:14.339 INFO  tr-conv-6 1355106364585954729-685511-1-FT-8710415920948495412 (WorkerCli
```

In the ELO Indexserver log, you can find the associated lines based on the random number `1355106364585954729` and the thread name `ft-send` in the call identifier:

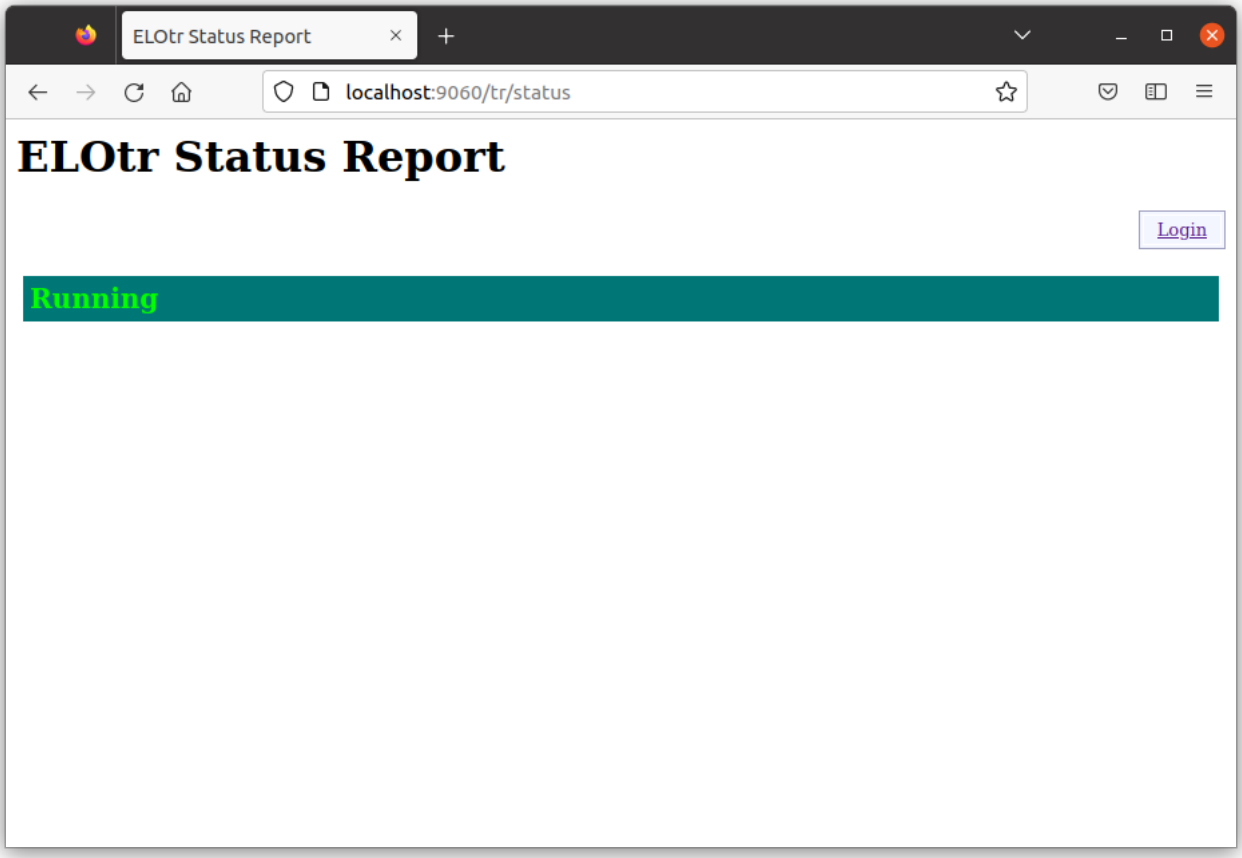
```
21:33:14.358 INFO  ft-send-2  (FtProcessorImpl.java:502)                        - Process c
...
21:33:14.372 INFO  ft-send-2 http-46 (ExtraStorage.java:404)                    - ft file w
...
21:33:14.381 INFO  ft-send-2  (FtProcessorImpl.java:502)                        - Process c
...
21:33:14.392 INFO  ft-send-2 http-48 (DocDb.java:1302)                          - preview f
...
21:33:14.398 INFO  ft-send-2  (TextreaderClient.java:376)                       - [3212] tr
```

Status page

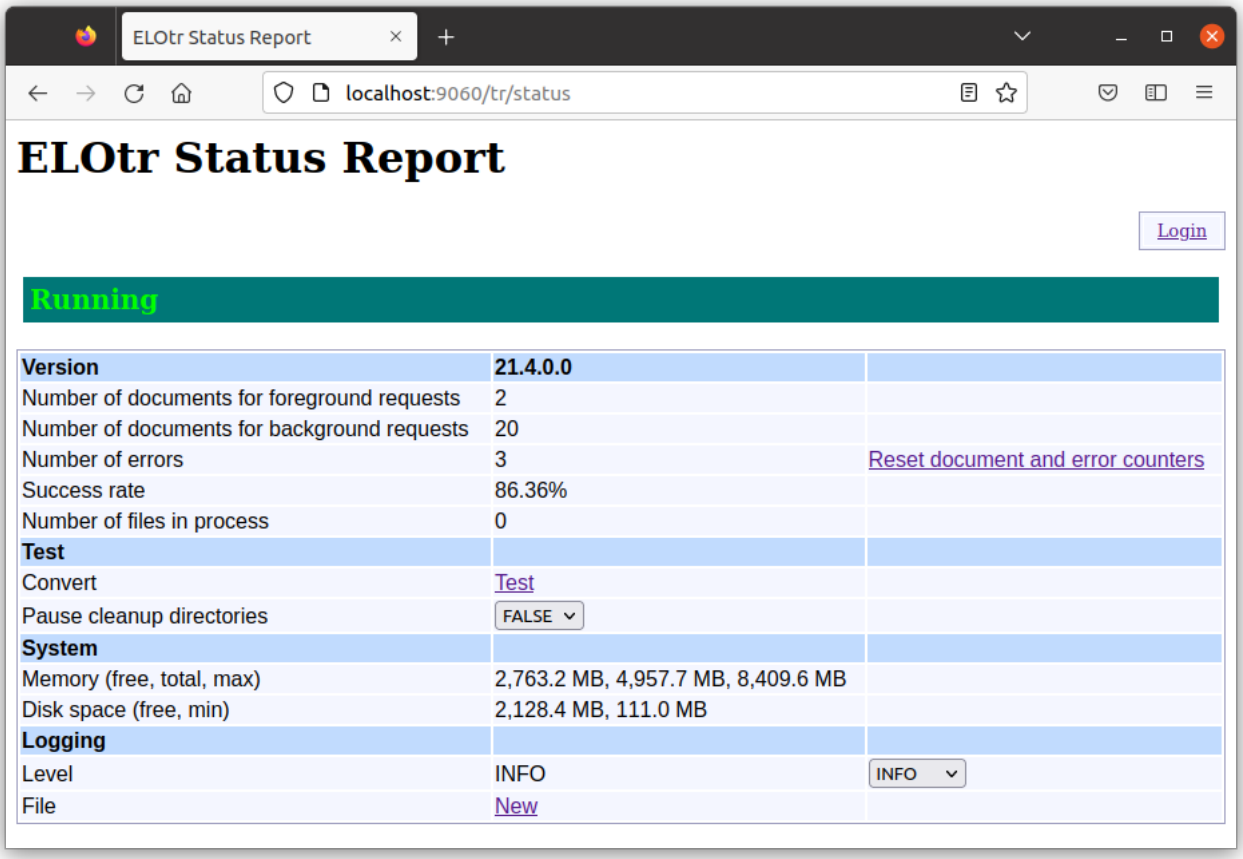
The status page of ELO Textreader (Gen. 2) can be accessed via `<host name>:<port number>/tr/status`.

Reduced and full status page

Without authentication, only a reduced status page is displayed:

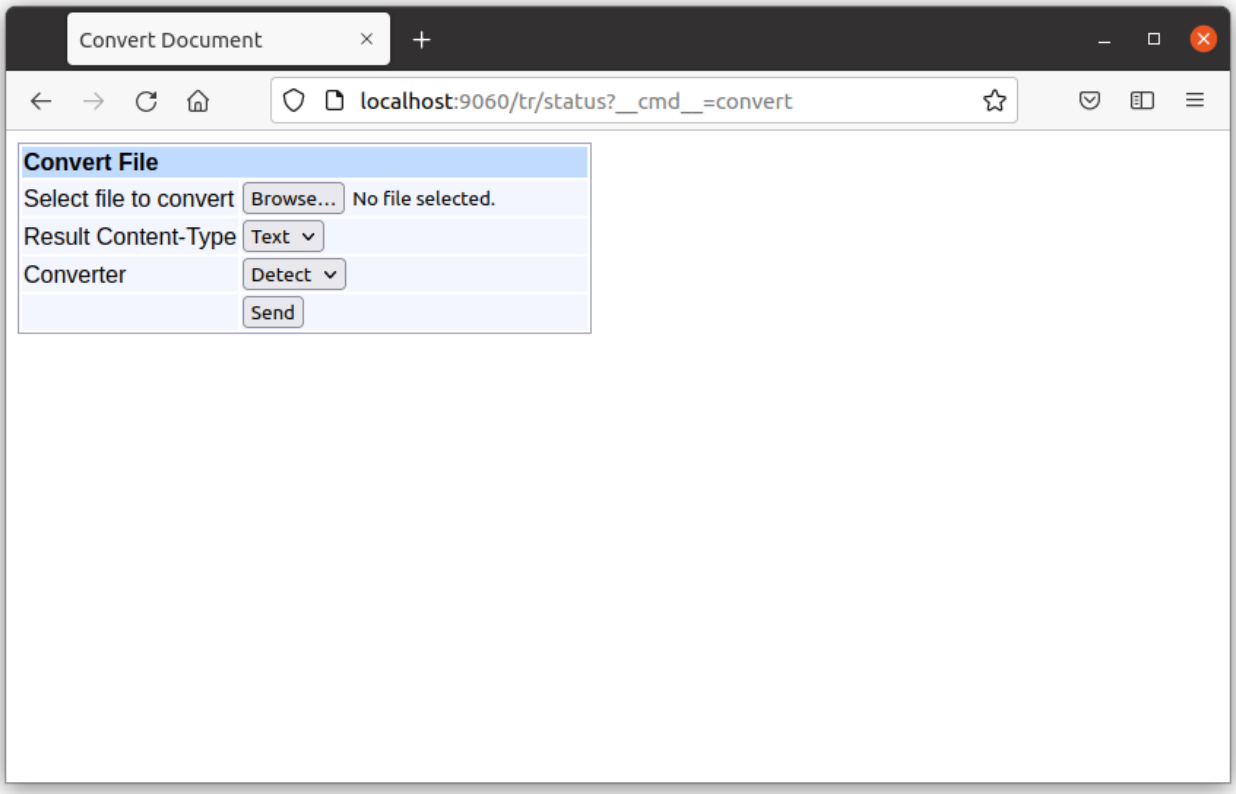


You will only see the whole page after logging on with the configured credentials.



Test page

Use the *Test* link to open a test page for converting a document.



After selecting a file, you can use the *Send* button to upload the document to ELO Textreader (gen. 2) and convert it. The result will appear in the same browser window. If conversion takes longer, you will get a new page with a link to get the result.

Logging level

You can select the logging level in the drop-down menu. The *Debug* setting should only be used for a short period for troubleshooting purposes because a lot of data is written to the report files over time.