



# Introduction to the OpenMinTeD platform

15 May 2018

This manual contains exercises that guide you through your first exploration of the OpenMinTeD platform for text and data mining. It is part of the online course 'Introduction to Text and Data Mining' (<https://www.fosteropenscience.eu/learning/introduction-to-text-and-data-mining/>).



OpenMinTeD is funded by the European Union's Horizon 2020 research and innovation programme under Grant Agreement No. 654021.



OpenMinTeD is an open and sustainable Text and Data Mining (TDM) platform and infrastructure where researchers can discover, collaboratively create, share and re-use knowledge from a wide range of text based scientific and scholarly related sources. The platform was developed by a consortium of 16 partners between May 2015 and May 2018, and will be maintained, updated and further developed for your use. One of the main challenges in text and data mining is to make different text mining sources and applications interoperable. OpenMinTeD has successfully overcome this challenge, and made the platform open for third parties to build on. The aim was also to make it easier for non-computer scientists to run text mining applications on text corpora.

Before you continue, please remember that the platform is still being developed further, and that occasionally some functions may not be fully working yet. This guide is meant to show you the platform and some of the basic functions.

## 1. Go to [services.openminted.eu](https://services.openminted.eu) and register

OpenMinTeD Platform

https://services.openminted.eu/home

openMIN7ED  
Open Mining Infrastructure for Text & Data

HOME SEARCH ADD PROCESS SUPPORT

SIGN IN | REGISTER

Text and data mining for scholarly works on the cloud

SEARCH FOR SCIENTIFIC LITERATURE CORPORA OR TEXT MINING SERVICES... SEARCH

20 TOOLS & SERVICES | 6 MI FULLTEXTS | 10 MI ABSTRACTS

Discover TDM applications > Retrieve OA content > Run on the cloud

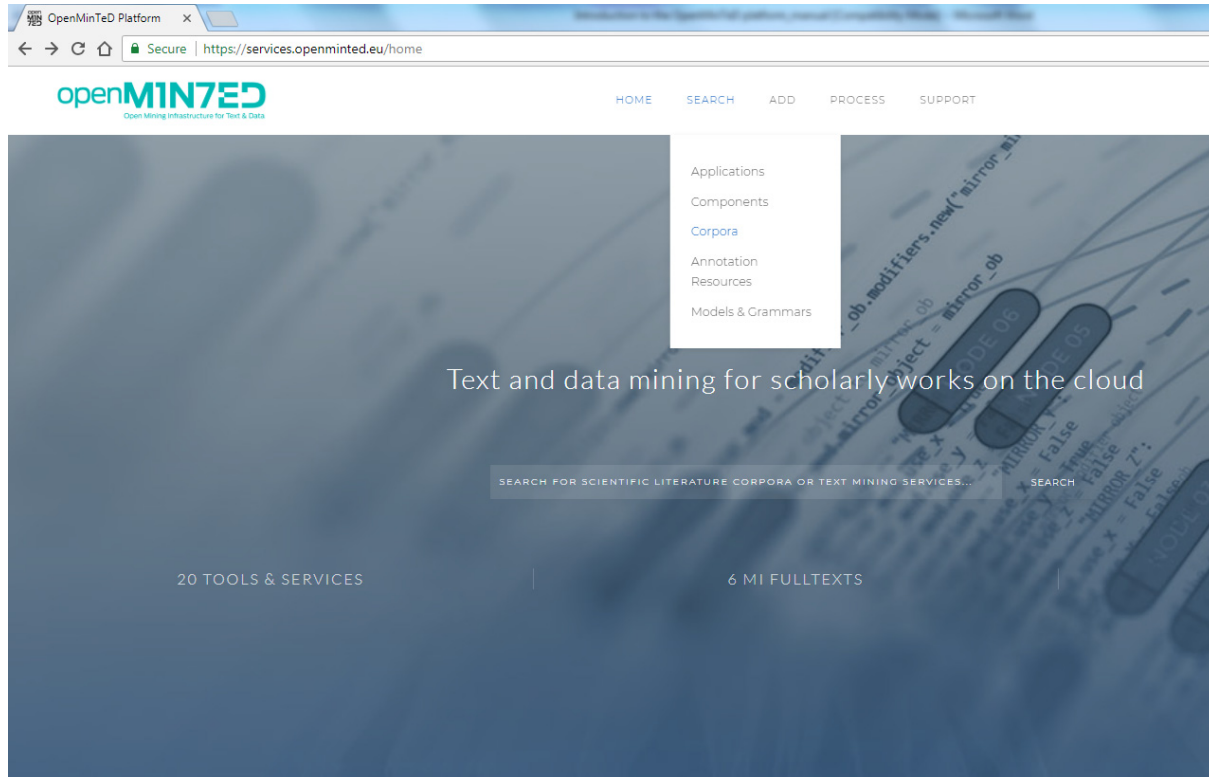
- 1 A REGISTRY OF TEXT AND DATA MINING APPLICATIONS
- 2 A BRIDGE TO OA SCIENTIFIC AND SCHOLARLY LITERATURE
- 3 A CLOUD COMPUTING ENVIRONMENT

Before you can start doing things, you first have to register by clicking **'register'** on the top right. You can get access through eduGAIN or one of the mentioned accounts (Facebook, Google, LinkedIn, ORCID).



## 2. Explore the platform

First you may want to get familiar with what the platform has to offer. Under **'SEARCH'** you can find all the content in the platform, including corpora (large structured sets of text) and applications.



*Discover TDM applications > Retrieve OA content > Run on the cloud*



If you click on **'Corpora'**, you get an overview of the sets of texts that are available in the platform. This includes sets of texts from [OpenAIRE](#) and [CORE](#), and corpora uploaded by users directly. If you click on a specific corpus, you will get a description of that corpus.

If you want to get an idea of the applications on the platform, go to **'SEARCH'** and click on **'Applications'**. If you click on a specific application, you will get information on the license, output format etc.

## 3. Adding tools and apps

We will not go into detail right now, but it is good to know that it is possible to connect a new text mining application to the platform, or build a new one from 'components' (pieces of software). Have a



look by clicking **'Applications'** under **'ADD'**.

Want to share a new application?

Applications are software programs intended for the end-user addressing one or multiple related user needs.

REGISTER AN EXISTING APPLICATION | BUILD AN APPLICATION WITH EXISTING COMPONENTS

OMTD EDITOR | UPLOAD XML

Describe your application using the OMTD editor

GO

HELP

You can register a TDM application in the OpenMinTeD catalog.

- registering an existing TDM application ; already possess. can be done by using the OMTD editor uploading an XML with the description of your application according to the guidelines for providers of software
- creating a new application by combining TDM components already available in the OpenMinTeD catalogue

In order to register software, it has to be compatible with the platform. Information about this is available at: <https://guidelines.openminded.eu/>.

## 4. Building a corpus

The OpenMinTeD platform includes publications from OpenAIRE (<https://www.openaire.eu/>) and CORE (<https://core.ac.uk/>). You can build your own corpus from the publications available through these two channels.

Go to **'ADD'** and **'Corpora'**.

You see two tabs **'UPLOAD YOUR CORPUS'** and **'BUILD A NEW CORPUS'**. Click on **'BUILD A NEW CORPUS'**.

Click on the button **'BUILD CORPUS'**.

You can now build your own corpus by searching for a certain keyword, and then refining it by publication year, open access availability, etc.

We are not going into detail here for now, but it is good to know this option exists for future use. If you build a corpus, you have to register it by filling in all the metadata required.



## 5. Running an application on a predefined corpus

Now you are going to run an application on a corpus, to get a feeling of how this works. Go to **'PROCESS'** in the main menu. You get the following screen:

Secure | https://services.openminted.eu/runApplication

openMIN7ED  
Open Mining Infrastructure for Text & Data

HOME SEARCH ADD PROCESS SUPPORT

RUN AN APPLICATION

SELECT AN INPUT CORPUS + SELECT AN APPLICATION → CLICK TO RUN THE APPLICATION! (YOU NEED TO SELECT BOTH AN INPUT AND AN APPLICATION FIRST)

**HELP**

In order to run a TDM application in OpenMinTeD you need to:

**STEP 1**

Select corpus available in OpenMinTeD catalogue.

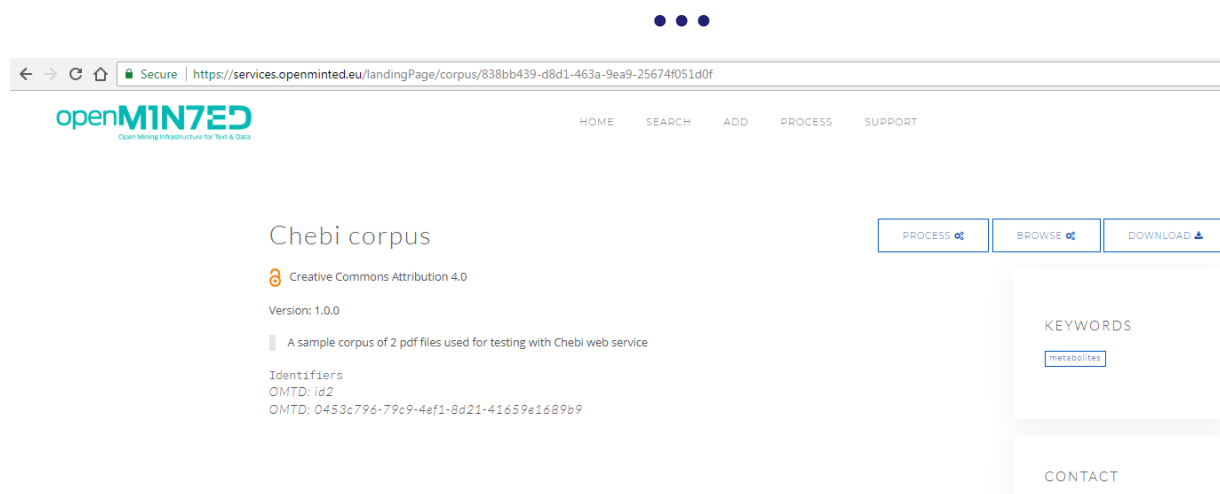
**Important:** To be able to use your private content, you should set it **public** first thus make it visible in the OpenMinTeD catalogue. How? Go to "My Corpora" under your profile menu.

**STEP 2**


Select an application available in the OpenMinTeD catalogue. Make sure that the application you select fits to your needs and can be used on the right content.

**Important:** To be able to use your private content, you

Click on **'SELECT AN INPUT CORPUS'** and search for the Chebi corpus. This is a small corpus with 2 PDFs that is meant for testing.



Chebi corpus

 Creative Commons Attribution 4.0

Version: 1.0.0

A sample corpus of 2 pdf files used for testing with Chebi web service

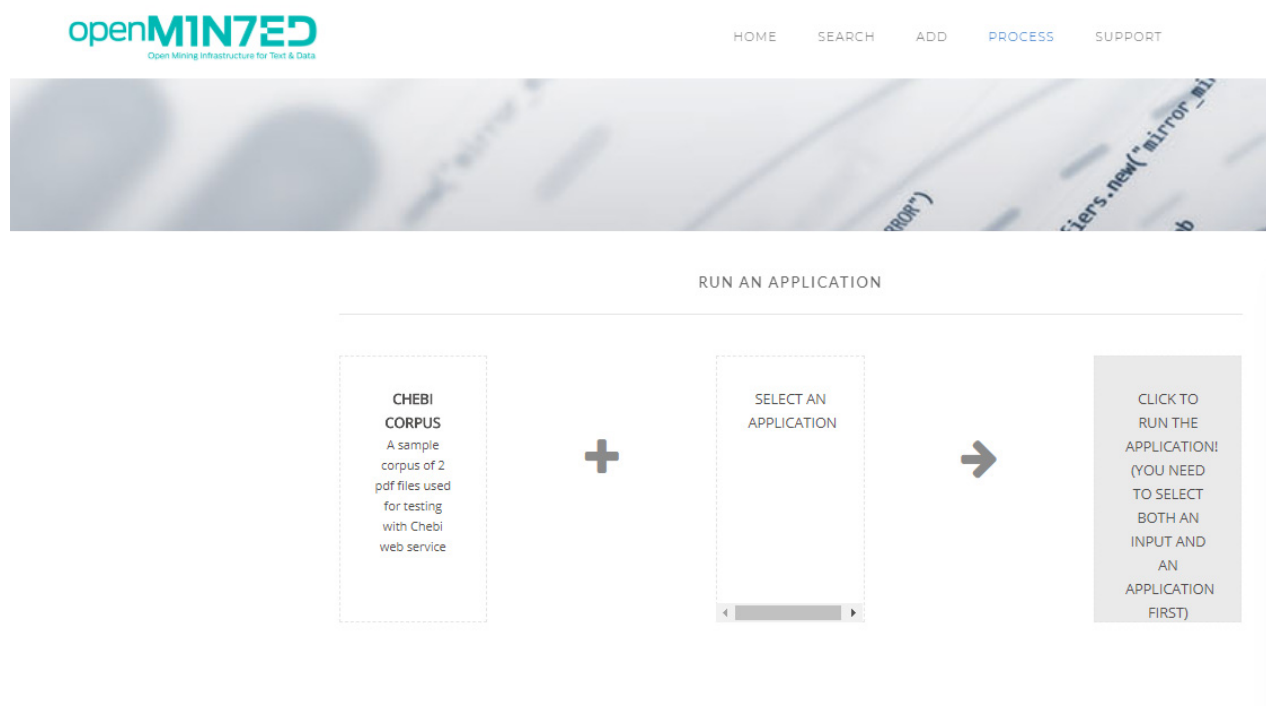
Identifiers  
OMTD: id2  
OMTD: 0453c796-79c9-4ef1-8d21-41659e1689b9

PROCESS BROWSE DOWNLOAD

KEYWORDS  
metabolites

CONTACT

Click on 'PROCESS'. You will see the following screen:



openMIN7ED

HOME SEARCH ADD PROCESS SUPPORT

RUN AN APPLICATION

**CHEBI CORPUS**  
A sample corpus of 2 pdf files used for testing with Chebi web service

+

**SELECT AN APPLICATION**

→

**CLICK TO RUN THE APPLICATION!**  
(YOU NEED TO SELECT BOTH AN INPUT AND AN APPLICATION FIRST)

Now click on 'SELECT AN APPLICATION' and look for 'CHEBI CURATION WEB SERVICE - A MACHINE LEARNING-BASED WORKFLOW'. If you click on this application, you will get information on what this application can do. It can identify metabolites, chemicals, species, proteins, biological information and chemical structural information in the texts. If you want a bit more background information: the team that built this application also wrote a blogpost on their work and the context: <http://openminted.eu/text-mining-discovery-small-molecules/>.

If you want to run this application click on 'RUN' and in the next screen on 'CLICK TO RUN THE APPLICATION'.



After a couple of minutes you see the following screen:

The screenshot shows a web browser window with the URL `https://services.openminted.eu/runApplication?input=838bb439-d8d1-463a-9ea9-25674f051d0f;application=579d1336-c0f4-46ee-a629-770fe8af06`. The page header includes the OpenMinTeD logo and navigation links: HOME, SEARCH, ADD, PROCESS, and SUPPORT. Below the header is a blurred image of a document. The main content area is titled "RUN AN APPLICATION" and features a green notification bar stating "Application run finished successfully". A workflow diagram is displayed, showing a sequence of steps: 1. "CHEBI CORPUS" (A sample corpus of 2 pdf files used for testing with Chebi web service). 2. A plus sign (+). 3. "CHEBI CURATION WEB SERVICE - A MACHINE LEARNING-BASED WORKFLOW" (This web service can detect entities in six categories: Metabolite, Chemical, Protein, Species, Biological Activity, and Spectral Data. The workflow underlying the web service consists of nine Argo components: Lingpipe, Sentence, etc.). 4. A right-pointing arrow (→). 5. A blue button labeled "CLICK HERE TO VIEW THE OUTPUT!".

This means the application has run successfully and you have built an annotated version of the text corpus, in which metabolites, chemicals, species, proteins, biological information and chemical structural information have been identified. Unfortunately you cannot see this, as the annotation viewer has not been implemented in the platform yet - the OpenMinTeD team is working on it. You can also download the results, they are in a XMI format.

## 6. Final remark

You should now have a sense of the OpenMinTeD platform and some of its possibilities. If you need help, have a look at the **'SUPPORT'** page of the platform. Feel free to browse around and explore more.