

Datathon 2025

One of the datasets of the 2025 datathon consists of a collection of war-time posters provided by the *KU Leuven bibliotheken*. It consists of images of these posters combined with some OCR extracted text files. The text files can be linked to the posters based on the titles of the files.

Since most posters contain text in multiple languages, the text files are organized in folders corresponding to Dutch, English, and French. However, note that there can be mistakes. Sometimes these mistakes can be the result of mistakes in the posters (e.g. example 2 below) but sometimes it can also be due to the processing of the images.

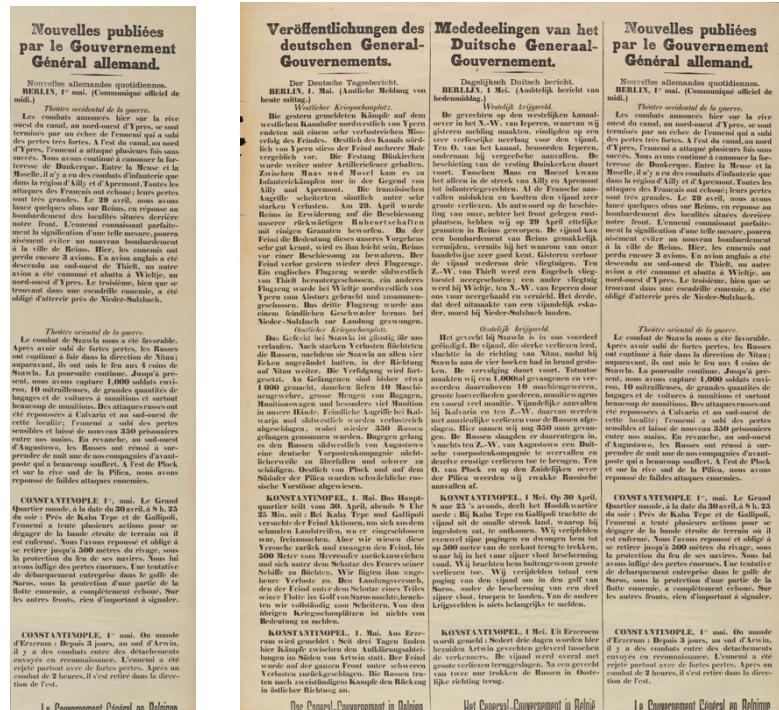
For each poster, both a **raw** file and a **processed** file are provided. The raw files are organized according to the layout of the page. The processed files are organized per language. To see the kind of processing that was performed, see example 3.

Example 1

Example is a standard poster that was processed correctly by the algorithm. It is a poster that consists of three languages. This image is split into three processed images that correspond to a cropped version of the original layout that only contains a single language (see left-hand image). This image is then passed into an OCR engine and converted into a text file:

[extract]

sVouvelles publiées par le Gouvernement | Général allemand. Nouvelles allemandes quotidiennes.
BERLIN, 1^o mai. (Communiqué officiel de



Théâtre occidental de la guerre. Les combats annoncés hier sur la rive ouest du canal, au nord-ouest d'Ypres, se sont terminés par un échec de l'ennemi qui a subi des pertes très fortes. A l'est du canal, au nord d'Ypres, l'ennemi a attaqué plusieurs fois sans succès. Nous avons continué à bombarder la forteresse de Dunkerque. Entre la Meuse et la Moselle, il n'y a eu des combats d'infanterie que dans la région d'Ailly et d'Apremont. Toutes les attaques des Français ont échoué; leurs pertes sont très grandes. Le 29 avril, nous avons lancé quelques assauts sur Reims, en réponse au bombardement de la ville par nos ennemis. Nous avons également mis en place une mesure, pour faire cesser l'ennemi de nos positions, pour empêcher l'ennemi de faire des assauts nocturnes contre nos positions. L'ennemi a aussi été contraint de faire des assauts nocturnes contre la ville de Reims. Hier, les canonniers ont perdu leur temps à bombarder la ville. Ils ont descendu au sud-ouest de Thiep, un autre assaut nocturne contre la ville de Wulff, au nord-ouest d'Ypres. Le résultat de ces attaques, qui étaient faites avec une escadre canonne, a été une victoire pour les deux batailles.

Example 2 – Mistake in poster

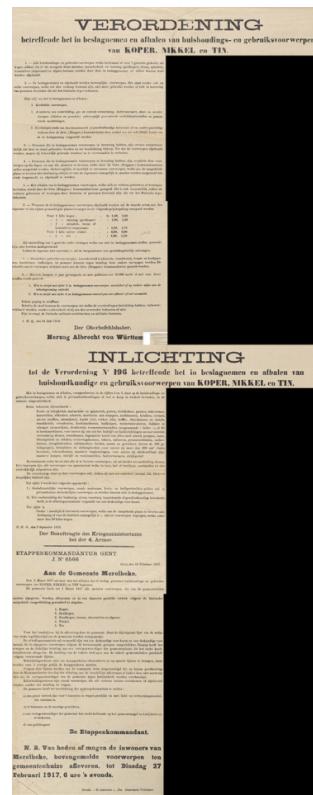
Sometimes, posters can have mistakes. For example, the first poster in the dataset has the French text twice:



In that case, the text will also be processed incorrectly. In this case, given the German title of the section on the left, the French text will be classified as German.

Example 3: More complicated formats

Sometimes the formats are more complicated compared to the two examples above. In these cases, we perform some more advanced preprocessing to construct a linear text flow into the processed files:



The file on the left has a non-linear flow of text that does not flow from top to bottom. This confuses the OCR tools. Therefore, these texts are converted first to linear text flows (on the right). These can be used as input for the OCR.