



How you can download DocVQA dataset

Step-1 → Open any browser.

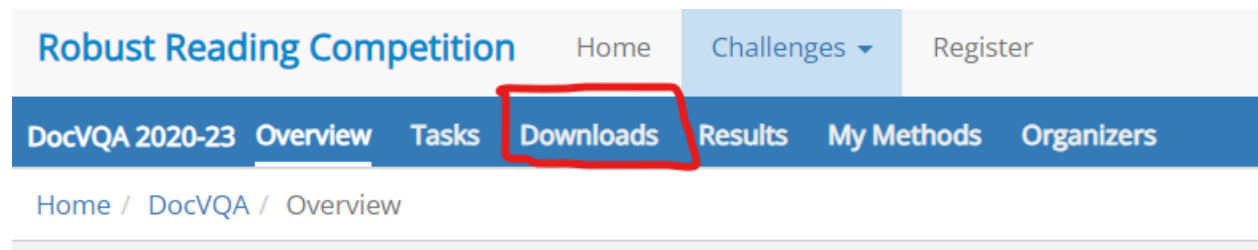
Step-2 → Head over to <https://www.docvqa.org/datasets/docvqa>

Step-3 → Scroll Down and Click on "challenge page"

Download

The dataset can be downloaded from the challenge page in RRC portal, Go to the "Download" tab in the [challenge page](#) and use the links under "Single Document Visual Question Answering"

Step-4 → Head over to Downloads



Step-5 → Now scroll down and navigate to Task 4

Task 4 - Multipage Document Visual Question Answering (MP-DocVQA)

The dataset for Multipage DocVQA is available to download from the following URLs:

- V1.0 uploaded on 19 February 2023. The dataset consists of 46436 questions posed over 5929 documents with 47952 pages in total. Besides the questions and answers, we also provide the document page images and the OCR extracted with Amazon Textract OCR of all the pages (64057) that belong to the documents within MP-DocVQA dataset, skipping the limitation of 20 pages. Hence, you can use longer versions of the documents with the same image and OCR quality if required.
 - [Questions and Answers](#)
 - [Images](#)
 - [OCR results \(Amazon Textract\)](#)
 - [IMDBs \(processed dataset for MP-DocVQA framework\)](#)

Step-6 → Click on "images" to start the download of the DocVQA dataset. The download will start and the extract all the files after the download is complete in

the file explorer.

Step-7 → You can also download the questions by clicking on the "Questions and Answers", also extract it and you will end up with json files

CONGRATS! You have successfully downloaded the DocVQA dataset.