Steps:

Go to the TCGA (the cancer genome atlas) database website and do the following:

- https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga
- Click on access TCGA data or go directly to the: https://portal.gdc.cancer.gov/
- Click on “Repository”
- On the left panel, from “data type” select the “slide image” (all have .SVS format and open access)
- From “experimental strategy” select “tissue slide” (represent the frozen slides)- more than 18,000 slides at the moment

Now we need to select the organs from the “Cases” (on the left panel again). There are 52 options for organ selection. We should not use the organ from the MoNuSeg dataset to have a new dataet. So, we exclude 9 organs, namely breast, liver, kidney, prostate, bladder, colon, stomach, lung and brain. For more details read the paper (https://ieeexplore.ieee.org/document/8880654)

From the rest, select 5-6 organs such as ovary, Thyroid gland, skin, Pancreas, heart, floor of the mouth, small intestine, etc.

Then from each organ, select 2 slides (from different centers and from different patients) to have the most possible variability in the dataset. (i.e. pay attention to the second and third identifier in the slide name. See details from here:
https://docs.gdc.cancer.gov/Encyclopedia/pages/TCGA_Barcode/)

Download and install the QuPath software (https://qupath.github.io) to show the whole slide images (WSI).

Pay attention to the image magnification (should be 40x) and the access (should be open)

Select a tile from the WSI using Qupath. Then export it to ImageJ and save it as a .tif file.

- Use rectangular tool
- Extension→ ImageJ →send region to ImageJ (set the down sampling factor to one, we need 40x objective)→ in ImageJ save it as tif file. Then using a common software such as Irfan Viewer, Matlab or a simple python code extract a tile with size of 512 x 512 (e.g. from top left coner) of the image.

For each image fill up the following table (to save all metadata)

Image name	organ	Bit-depth	magnification	Actual size of the image	Pixel dimension	Staining	Disease type
TCGA-D3-A51F-06A-01-TSA.70FEEBE6-799F-47DE-85CA-ACFC7FA342BE.svs	Skin	8	40x	143752x41484	0.246 um	H&E	

Perform manual instance segmentation of nuclei using ImageJ software on the extracted tile with the size of 512 x 512.