1. ICDAR OCR Datasets:
   * ICDAR 2019 MLT (Multi-Lingual Text) OCR Dataset (14.47 GB, images + annotations):  
     Download page: <https://rrc.cvc.uab.es/?ch=15>  
     Kaggle mirror: <https://www.kaggle.com/datasets/zubairalibhutto/mlt-19-ocr-dataset>
   * ICDAR 2019 Post-OCR Text Correction Dataset (approx 4.8 GB for Finnish Fraktur, others vary):  
     Zenodo link: <https://zenodo.org/records/3515403>
   * ICDAR 2021 Datasets (various, organized by challenge specifics):  
     General download portal: <https://tc11.cvc.uab.es/datasets/publ/24>
2. IIIT 5K-Word Dataset:
   * Official site with dataset and annotations:  
     <http://cvit.iiit.ac.in/projects/SceneTextUnderstanding/IIIT5K.html>
   * Direct dataset page (TC-11 site):  
     <https://tc11.cvc.uab.es/datasets/IIIT%205K-Word_1>
3. Google Research OCR Dataset (HierText with word, line, paragraph annotations):
   * GitHub repository with instructions and download command:  
     <https://github.com/google-research-datasets/hiertext>
   * Download images hosted on AWS S3 (requires AWS CLI):

text

aws s3 --no-sign-request cp s3://open-images-dataset/ocr/train.tgz .

aws s3 --no-sign-request cp s3://open-images-dataset/ocr/validation.tgz .

aws s3 --no-sign-request cp s3://open-images-dataset/ocr/test.tgz .

* + Decompress the tar files after downloading.

1. UW English Document Image Database:
   * Detailed description and some access details in this manual:  
     <https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=04746c4c5e546b818de1a6ad351ed0055f1d77c9>
   * Public digital collections at UW Libraries (may require specific access or scraping/documented download):  
     [https://content.lib.washington.edu](https://content.lib.washington.edu/)