Interreg
EUROPEAN UNION
North-West Europe
Di-Plast
European Regional Development Fund

Di-Plast Matrix Data Extractor

Di-Plast Matrix Data Extractor (MDE) is a web-based application, which can be deployed on personal computer. It identifies document table regions on PDF documents using *Computer Vision based Deep Learning, especially Transfer Learning and Object Detection* algorithm. Then it extracts all textual data into text files by applying *Optical Character Recognition (OCR)* and also extracts tabular data separately in excel files using *Camelot* python package. It supports to transfer manufacturer names and corresponding technical datasheets names (or PDF filenames) to *MongoDB* database table for further processing.

The code can be downloaded from *GitHub* (https://github.com/cslab-hub/MatrixDataExtractor). **Only open-source software or library are used** in this application. MDE is primarily divided into 2 sections-

1. **Table Detection**: It provides code to train and test deep learning object detection model for document table detection task. The model is built mainly on *PyTorch Detectron2* library. The official support of Detectron2 library is available only on *Linux* OS (operating system). *Linux Desktop* version is recommended for normal users. Advanced users can use *Linux Server* version according to their choice.
2. **Backend**: It is the code of *Django* based web application, which provides a basic user interface to access the application functionalities for normal users.

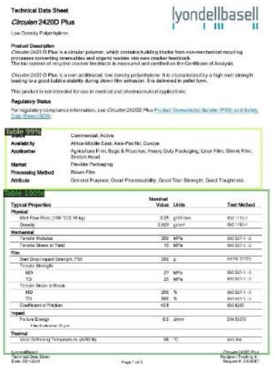
First, build table detection model weight *(model_final.pth*). Next, incorporate the model weight along with corresponding model description XML file (*faster_rcnn_R_101_FPN_3x_config.yaml or uos_dip_config.yaml*) within Django application. The pre-requisite of MDE is given below-

1. For <u>*normal users*</u>, it is recommended to use *Linux Desktop* version (which comes with nice user interface), e.g. *Ubuntu OS*. The user interface is recommended for normal users
    - to create *Manufacturer* sub-folders
    - to store *Technical datasheets* (or PDF files) within Manufacturer sub-folders
    - check the model inference results, if table detection model correctly identifies document table regions on unseen document images or not.

   <u>*Advanced users*</u> can use *Linux Server* version.

2. Please remember the difference between two terms-

   <u>***Document image***</u>: Each PDF page converted into image format.



Technical Data Sheet

Circulen 2420D Plus

Low Density Polyethylene

Product Description

Circulen 2420 D Plus is a circular polymer, which contains building blocks from non-mechanical recycling processes converting renewables and organic wastes into new cracker feedstock.
The bio content of recycled cracker feedstock is measured and certified on the Certificate of Analysis.

Circulen 2420 D Plus is a non additivated, low density polyethylene. It is characterized by a high melt strength leading to a good bubble stability during blown film extrusion. It is delivered in pellet form.

This product is not intended for use in medical and pharmaceutical applications.

Regulatory Status

For regulatory compliance information, see *Circulen* 2420D Plus Product Stewardship Bulletin (PSB) and Safety Data Sheet (SDS).

Status	Commercial: Active
Availability	Africa-Middle East; Asia-Pacific; Europe
Application	Agriculture Film; Bags & Pouches; Heavy Duty Packaging; Liner Film; Shrink Film; Stretch Hood
Market	Flexible Packaging
Processing Method	Blown Film
Attribute	General Purpose; Good Processability; Good Tear Strength; Good Toughness

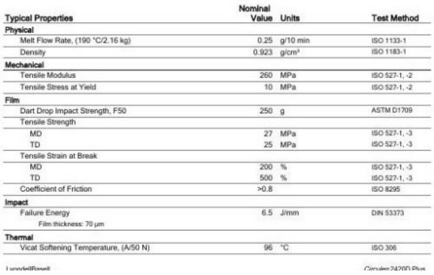
Typical Properties	Nominal Value	Units	Test Method
Physical			
Melt Flow Rate, (190 °C/2.16 kg)	0.25	g/10 min	ISO 1133-1
Density	0.923	g/cm³	ISO 1183-1
Mechanical			
Tensile Modulus	260	MPa	ISO 527-1, -2
Tensile Stress at Yield	10	MPa	ISO 527-1, -2
Film			
Dart Drop Impact Strength, F50	250	g	ASTM D1709
Tensile Strength			
MD	27	MPa	ISO 527-1, -3
TD	25	MPa	ISO 527-1, -3
Tensile Strain at Break			
MD	200	%	ISO 527-1, -3
TD	500	%	ISO 527-1, -3
Coefficient of Friction	>0.8		ISO 8295
Impact			
Failure Energy Film thickness: 70 µm	6.5	J/mm	DIN 53373
Thermal			
Vicat Softening Temperature, (A/50 N)	96	°C	ISO 306

LyondellBasell
Technical Data Sheet
Date: 5/21/2021

Page 1 of 3

Circulen 2420D Plus
Recipient: Trading X
Request #: 3308087

   <u>***Document Table Image***</u>: Each document table on each document image.



Typical Properties	Nominal Value	Units	Test Method
Physical			
Melt Flow Rate, (190 °C/2.16 kg)	0.25	g/10 min	ISO 1133-1
Density	0.923	g/cm³	ISO 1183-1
Mechanical			
Tensile Modulus	260	MPa	ISO 527-1, -2
Tensile Stress at Yield	10	MPa	ISO 527-1, -2
Film			
Dart Drop Impact Strength, F50	250	g	ASTM D1709
Tensile Strength			
MD	27	MPa	ISO 527-1, -3
TD	25	MPa	ISO 527-1, -3
Tensile Strain at Break			
MD	200	%	ISO 527-1, -3
TD	500	%	ISO 527-1, -3
Coefficient of Friction	>0.8		ISO 8295
Impact			
Failure Energy Film thickness: 70 µm	6.5	J/mm	DIN 53373
Thermal			
Vicat Softening Temperature, (A/50 N)	96	°C	ISO 306

LyondellBasell *Circulen* 2420D Plus

3. Please check README file at GitHub page for installation. Install *MongoDB* and *Elastic Search* on Linux along with *Anaconda* environment and other libraries. An interface of MongoDB database (e.g. *MongoDB*

Compass) is recommended to access data from MongoDB. Elastic Search can be useful to search results based on textual query, which could be incorporated through future code development.

4. To train and test deep learning model, GPU enabled computer is recommended. Install *PyTorch 1.8.0 GPU* version and relevant *Detectron2* library for Table Detection section. For Backend section, you can use *PyTorch 1.8.0 CPU* version and relevant Detectron2 library.

The primary code structure is shown below-

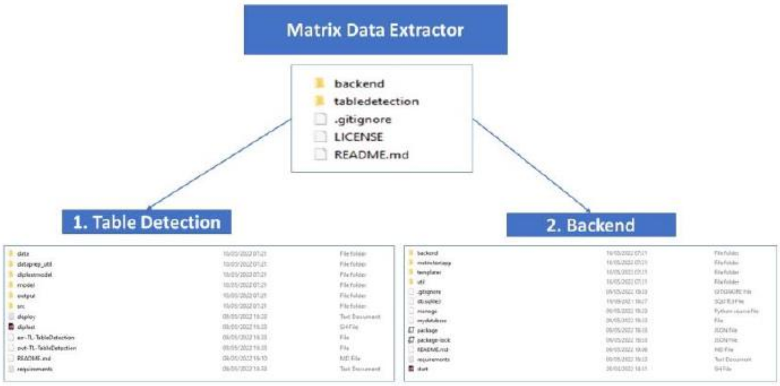


Fig. Primary code structure

Backend - Web Application

Installation:

Please check README file at GitHub page for installation. Optionally you can install *MongoDB Compass* tool.

Folder Structure Overview:

Normal users need to access *MatrixDataExtractor/backend/util* folder. Folder */util/prop* contains *MDE.xml* file, which is used for Django application configuration management. Folder */util/data/tabledet/modelweight* contains

- Faster R-CNN based object detection model description file: *faster_rcnn_R_101_FPN_3x_config.yaml* or *uos_dip_config.yaml*
- Table detection model weight: *model_final.pth*, which is built after table detection model training.

The *MatrixDataExtractor/backend/util* folder structure is shown below-

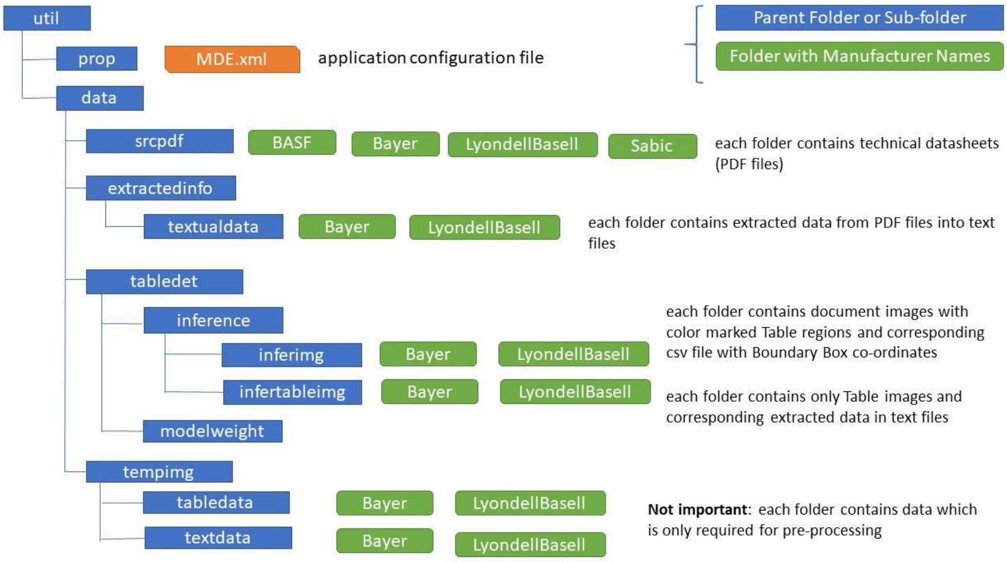

util
prop
MDE.xml
application configuration file
Parent Folder or Sub-folder
Folder with Manufacturer Names
data
srcpdf
BASF
Bayer
LyondellBasell
Sabic
each folder contains technical datasheets (PDF files)
extractedinfo
textualdata
Bayer
LyondellBasell
each folder contains extracted data from PDF files into text files
tabledet
inference
inferimg
Bayer
LyondellBasell
each folder contains document images with color marked Table regions and corresponding csv file with Boundary Box co-ordinates
infertableimg
Bayer
LyondellBasell
each folder contains only Table images and corresponding extracted data in text files
modelweight
tempimg
tabledata
Bayer
LyondellBasell
textdata
Bayer
LyondellBasell
Not important: each folder contains data which is only required for pre-processing

Fig. Utility folder structure overview

Pre-requisite for MDE web Application :

1. Make sure */util/prop* folder contains *MDE.xml* file, which is used for MDE web application configuration.
2. Folder */util/data/tabledet/modelweight* contains
   - Model description file: *faster_rcnn_R_101_FPN_3x_config.yaml* or *uos_dip_config.yaml*
   - *Model weight file: model_final.pth*

   These files are taken after Deep Learning Table Detection model training. Please refer *Model Training* sub-section of *Table Detection* section for more details.

Run web application:

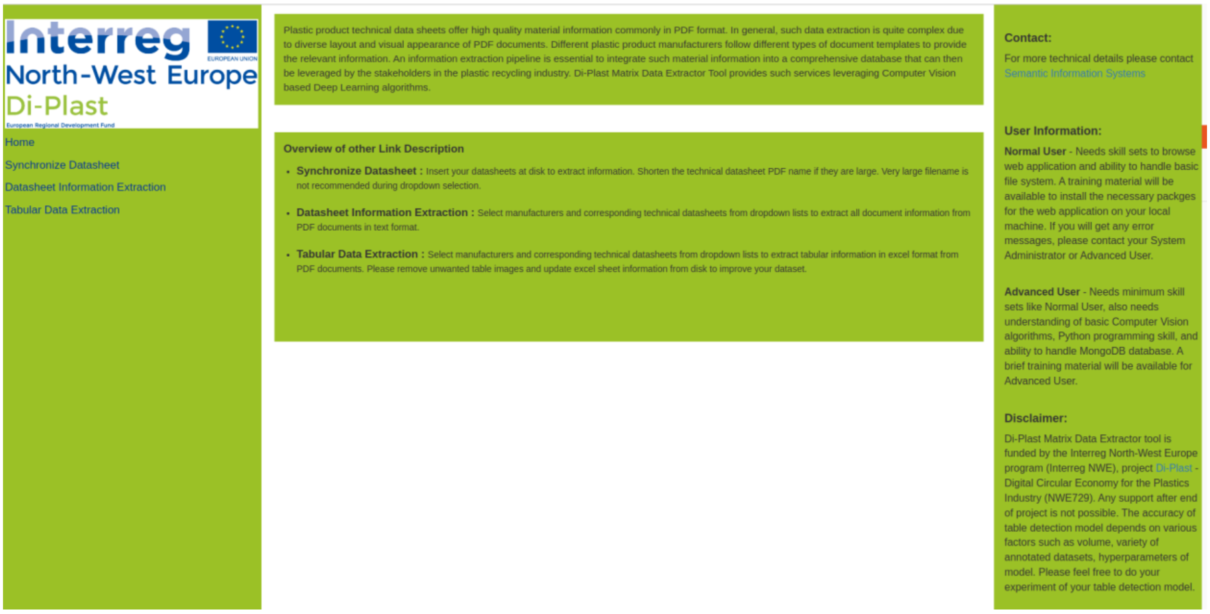
1. Make sure *MongoDB and Elastic Search* services are installed on your Linux OS. You will find *start.sh* (shell) file in *MatrixDataExtractor/backend* folder. Execute the shell file with below command-

   $ bash -i start.sh

2. The web application starts on Anaconda environment *env_mde* by running the shell file. You can browse the web application by accessing **localhost:8000** url on your personal computer. If you want, you can give your preferred URL name at ALLOWED_HOSTS of *MatrixDataExtractor/backend/backend* folder's *setting.py* file.

User guide of MDE web application:

1. When the web application is running, you can see homepage as **Home** link (at left panel) along with other link descriptions. Several instructions are mentioned on webpage for simplicity.



Home

Synchronize Datasheet

Datasheet Information Extraction

Tabular Data Extraction

Plastic product technical data sheets offer high quality material information commonly in PDF format. In general, such data extraction is quite complex due to diverse layout and visual appearance of PDF documents. Different plastic product manufacturers follow different types of document templates to provide the relevant information. An information extraction pipeline is essential to integrate such material information into a comprehensive database that can then be leveraged by the stakeholders in the plastic recycling industry. Di-Plast Matrix Data Extractor Tool provides such services leveraging Computer Vision based Deep Learning algorithms.

Overview of other Link Description

- **Synchronize Datasheet :** Insert your datasheets at disk to extract information. Shorten the technical datasheet PDF name if they are large. Very large filename is not recommended during dropdown selection.
- **Datasheet Information Extraction :** Select manufacturers and corresponding technical datasheets from dropdown lists to extract all document information from PDF documents in text format.
- **Tabular Data Extraction :** Select manufacturers and corresponding technical datasheets from dropdown lists to extract tabular information in excel format from PDF documents. Please remove unwanted table images and update excel sheet information from disk to improve your dataset.

Contact:

For more technical details please contact Semantic Information Systems

User Information:

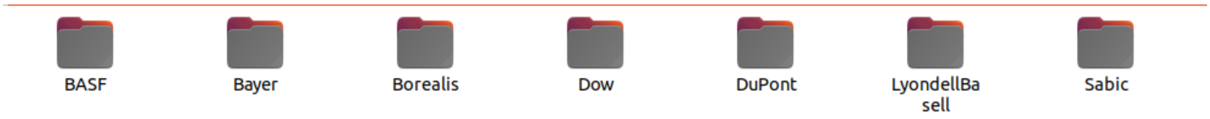
Normal User - Needs skill sets to browse web application and ability to handle basic file system. A training material will be available to install the necessary packges for the web application on your local machine. If you will get any error messages, please contact your System Administrator or Advanced User.

Advanced User - Needs minimum skill sets like Normal User, also needs understanding of basic Computer Vision algorithms, Python programming skill, and ability to handle MongoDB database. A brief training material will be available for Advanced User.

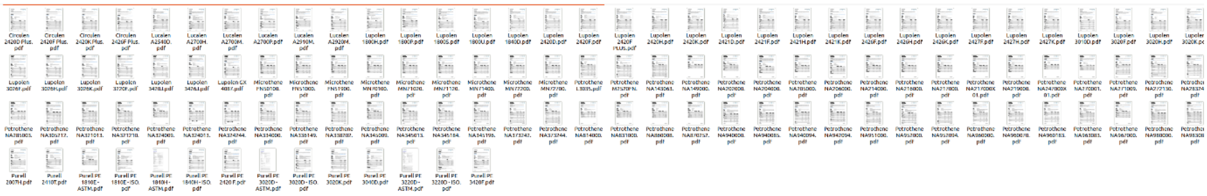
Disclaimer:

Di-Plast Matrix Data Extractor tool is funded by the Interreg North-West Europe program (Interreg NWE), project Di-Plast - Digital Circular Economy for the Plastics Industry (NWE729). Any support after end of project is not possible. The accuracy of table detection model depends on various factors such as volume, variety of annotated datasets, hyperparameters of model. Please feel free to do your experiment of your table detection model.

2. Create sub-folders with *Manufacturer* names under */util/data/srcpdf* folder (s.g. BASF, Bayer, LyondellBasell). Keep corresponding PDF files within each folder. The sub-folders are created under */util/data/srcpdf* folder as below-

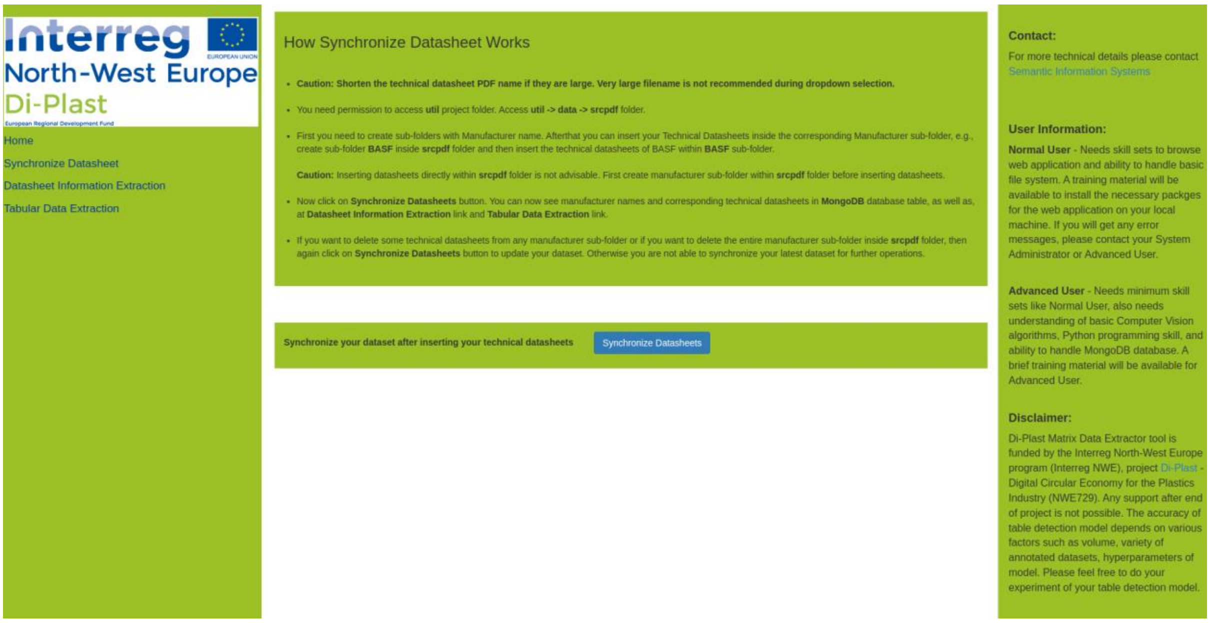


The PDF files are kept within each sub-folder (e.g. LyondellBasell sub-folder) as below-



3. Go to ***Synchronize Datasheet*** link and click on ***Synchronize Datasheets*** button. It will synchronize all PDF files and corresponding sub-folders under */util/data/srcpdf* folder. It synchronizes

manufacturer names and corresponding PDF filenames in ***matrixtextapp_datasheet*** table in *MongoDB* database.



How Synchronize Datasheet Works

- **Caution: Shorten the technical datasheet PDF name if they are large. Very large filename is not recommended during dropdown selection.**
- You need permission to access **util** project folder. Access **util -> data -> srcpdf** folder.
- First you need to create sub-folders with Manufacturer name. Afterthat you can insert your Technical Datasheets inside the corresponding Manufacturer sub-folder, e.g., create sub-folder **BASF** inside **srcpdf** folder and then insert the technical datasheets of BASF within **BASF** sub-folder.

  Caution: Inserting datasheets directly within **srcpdf** folder is not advisable. First create manufacturer sub-folder within **srcpdf** folder before inserting datasheets.
- Now click on **Synchronize Datasheets** button. You can now see manufacturer names and corresponding technical datasheets in **MongoDB** database table, as well as, at **Datasheet Information Extraction** link and **Tabular Data Extraction** link.
- If you want to delete some technical datasheets from any manufacturer sub-folder or if you want to delete the entire manufacturer sub-folder inside **srcpdf** folder, then again click on **Synchronize Datasheets** button to update your dataset. Otherwise you are not able to synchronize your latest dataset for further operations.

Synchronize your dataset after inserting your technical datasheets Synchronize Datasheets

Contact:

For more technical details please contact Semantic Information Systems

User Information:

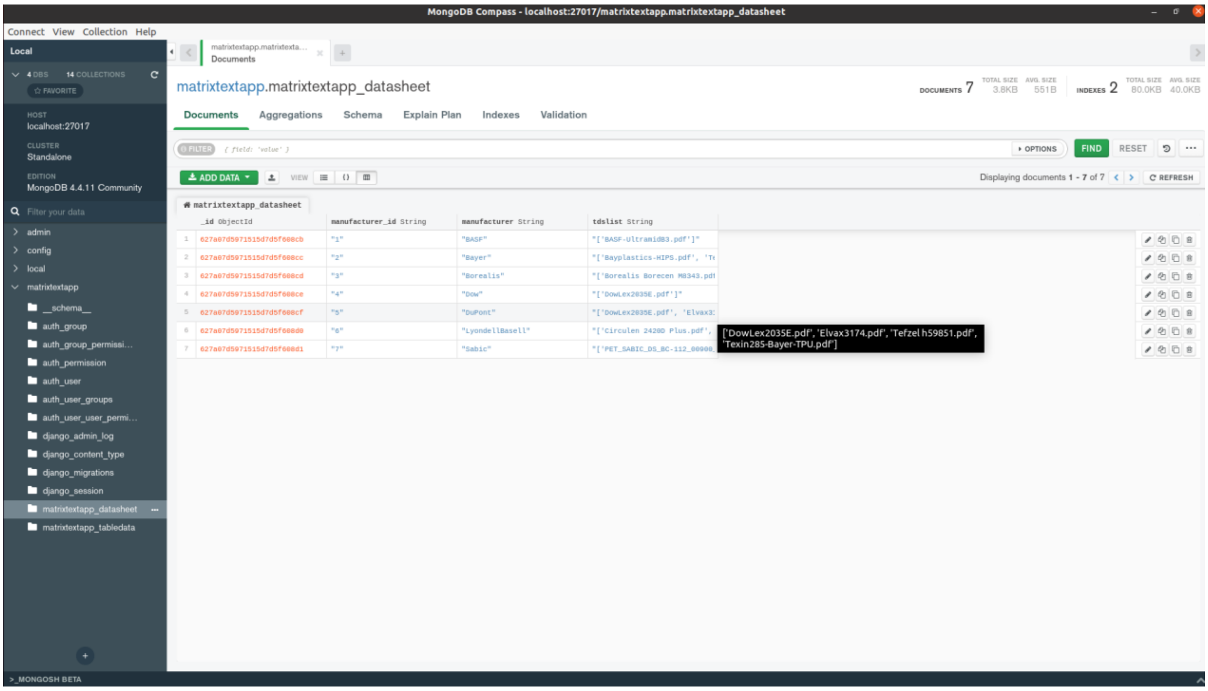
Normal User - Needs skill sets to browse web application and ability to handle basic file system. A training material will be available to install the necessary packges for the web application on your local machine. If you will get any error messages, please contact your System Administrator or Advanced User.

Advanced User - Needs minimum skill sets like Normal User, also needs understanding of basic Computer Vision algorithms, Python programming skill, and ability to handle MongoDB database. A brief training material will be available for Advanced User.

Disclaimer:

Di-Plast Matrix Data Extractor tool is funded by the Interreg North-West Europe program (Interreg NWE), project Di-Plast - Digital Circular Economy for the Plastics Industry (NWE729). Any support after end of project is not possible. The accuracy of table detection model depends on various factors such as volume, variety of annotated datasets, hyperparameters of model. Please feel free to do your experiment of your table detection model.

4. You can check manufacturer names and corresponding PDF filenames in ***matrixtextapp_datasheet*** table in *MongoDB* database. You can use *MongoDB Compass* tool to access the data.

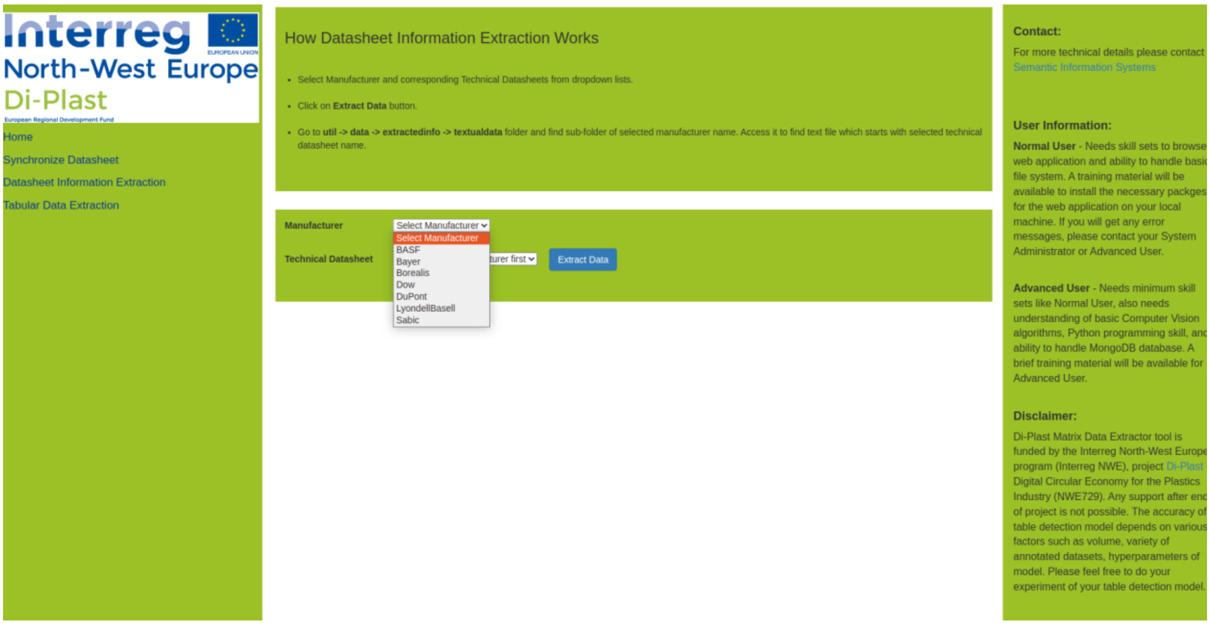


matrixtextapp_datasheet

	_id ObjectId	manufacturer_id String	manufacturer String	tdslist String
1	627a07d5971515d7d5f608cb	"1"	"BASF"	"['BASF-UltramidB3.pdf']"
2	627a07d5971515d7d5f608cc	"2"	"Bayer"	"['Bayplastics-HIPS.pdf', 'Te
3	627a07d5971515d7d5f608cd	"3"	"Borealis"	"['Borealis Borecen M8343.pdf
4	627a07d5971515d7d5f608ce	"4"	"Dow"	"['DowLex2035E.pdf']"
5	627a07d5971515d7d5f608cf	"5"	"DuPont"	"['DowLex2035E.pdf', 'Elvax3:
6	627a07d5971515d7d5f608d0	"6"	"LyondellBasell"	"['Circulen 2420D Plus.pdf',
7	627a07d5971515d7d5f608d1	"7"	"Sabic"	"['PET_SABIC_DS_BC-112_00908_

['DowLex2035E.pdf', 'Elvax3174.pdf', 'Tefzel h59851.pdf', 'Texin285-Bayer-TPU.pdf']

5. Go to ***Datasheet Information Extraction*** link to verify manufacturer names and corresponding technical datasheet names (or PDF filenames) available in dropdown menu. Select *Manufacturer* name first to access PDF files.



Interreg North-West Europe Di-Plast

Home

Synchronize Datasheet

Datasheet Information Extraction

Tabular Data Extraction

How Datasheet Information Extraction Works

- Select Manufacturer and corresponding Technical Datasheets from dropdown lists.
- Click on **Extract Data** button.
- Go to **util -> data -> extractedinfo -> textualdata** folder and find sub-folder of selected manufacturer name. Access it to find text file which starts with selected technical datasheet name.

Manufacturer Select Manufacturer

Select Manufacturer
BASF
Bayer
Borealis
Dow
DuPont
LyondellBasell
Sabic

Technical Datasheet turer first

Contact:

For more technical details please contact Semantic Information Systems

User Information:

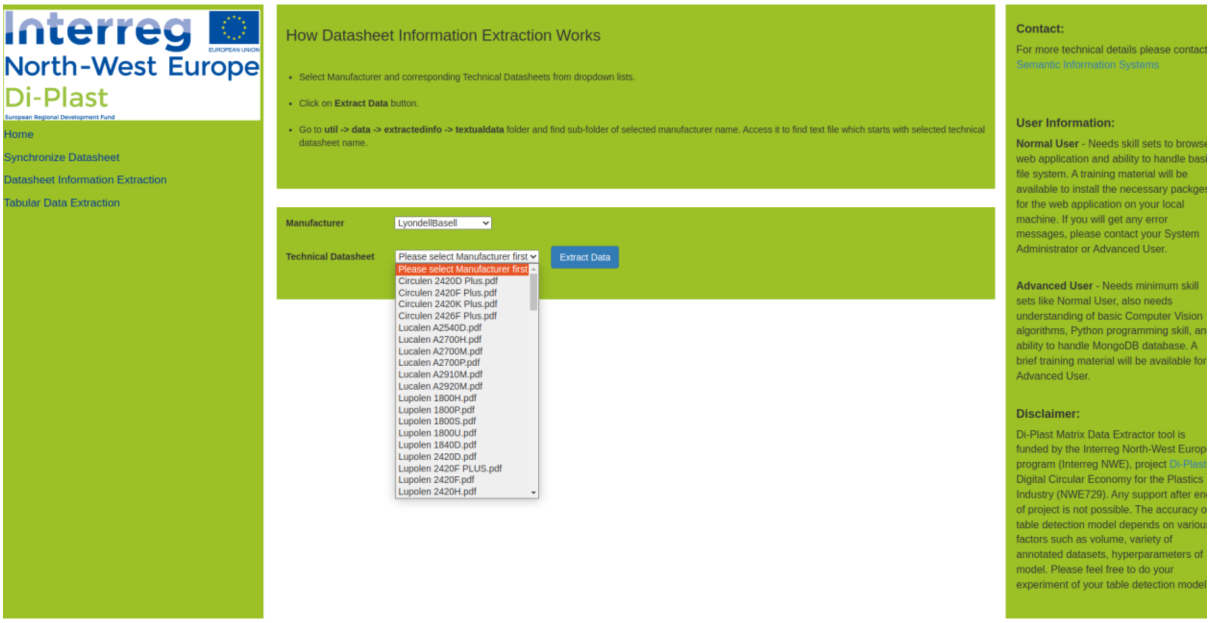
Normal User - Needs skill sets to browse web application and ability to handle basic file system. A training material will be available to install the necessary packges for the web application on your local machine. If you will get any error messages, please contact your System Administrator or Advanced User.

Advanced User - Needs minimum skill sets like Normal User, also needs understanding of basic Computer Vision algorithms, Python programming skill, and ability to handle MongoDB database. A brief training material will be available for Advanced User.

Disclaimer:

Di-Plast Matrix Data Extractor tool is funded by the Interreg North-West Europe program (Interreg NWE), project Di-Plast - Digital Circular Economy for the Plastics Industry (NWE729). Any support after end of project is not possible. The accuracy of table detection model depends on various factors such as volume, variety of annotated datasets, hyperparameters of model. Please feel free to do your experiment of your table detection model.

6. Upon selection of *Manufacturer* name, you can get corresponding *Technical datasheet* names (or PDF filenames) in dropdown menu.



Interreg North-West Europe Di-Plast

Home

Synchronize Datasheet

Datasheet Information Extraction

Tabular Data Extraction

How Datasheet Information Extraction Works

- Select Manufacturer and corresponding Technical Datasheets from dropdown lists.
- Click on **Extract Data** button.
- Go to **util -> data -> extractedinfo -> textualdata** folder and find sub-folder of selected manufacturer name. Access it to find text file which starts with selected technical datasheet name.

Manufacturer LyondellBasell

Technical Datasheet Please select Manufacturer first

Please select Manufacturer first
Circulen 2420D Plus.pdf
Circulen 2420F Plus.pdf
Circulen 2420K Plus.pdf
Circulen 2426F Plus.pdf
Lucalen A2540D.pdf
Lucalen A2700H.pdf
Lucalen A2700M.pdf
Lucalen A2700P.pdf
Lucalen A2910M.pdf
Lucalen A2920M.pdf
Lupolen 1800H.pdf
Lupolen 1800P.pdf
Lupolen 1800S.pdf
Lupolen 1800U.pdf
Lupolen 1840D.pdf
Lupolen 2420D.pdf
Lupolen 2420F PLUS.pdf
Lupolen 2420F.pdf
Lupolen 2420H.pdf

Contact:

For more technical details please contact Semantic Information Systems

User Information:

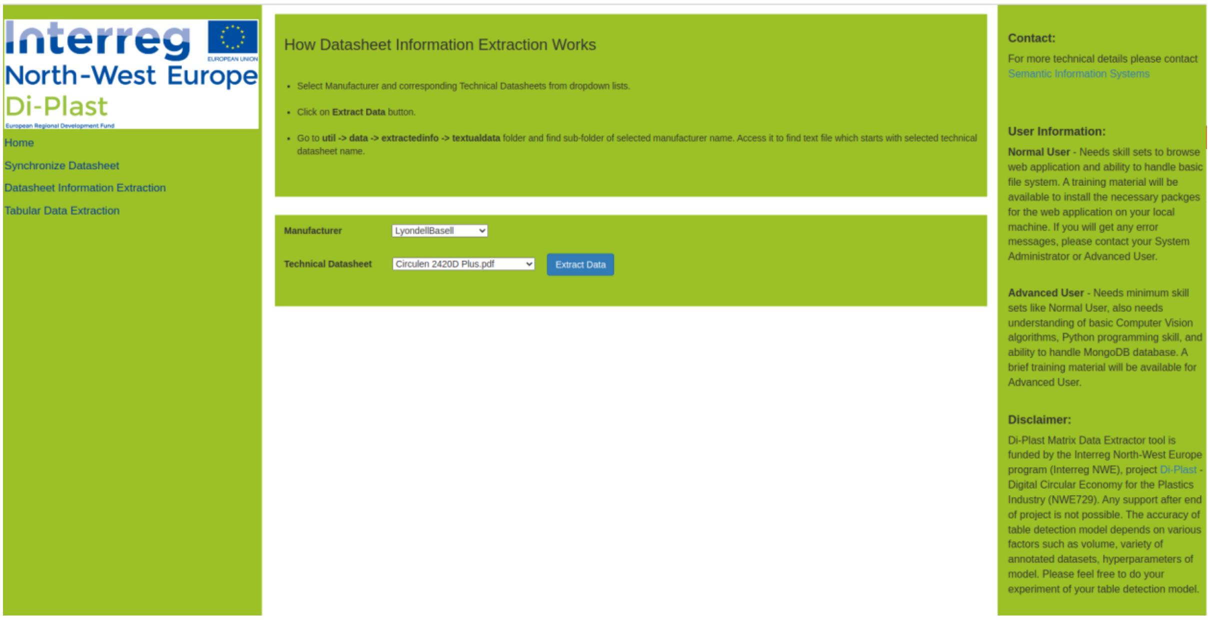
Normal User - Needs skill sets to browse web application and ability to handle basi file system. A training material will be available to install the necessary packges for the web application on your local machine. If you will get any error messages, please contact your System Administrator or Advanced User.

Advanced User - Needs minimum skill sets like Normal User, also needs understanding of basic Computer Vision algorithms, Python programming skill, an ability to handle MongoDB database. A brief training material will be available for Advanced User.

Disclaimer:

Di-Plast Matrix Data Extractor tool is funded by the Interreg North-West Europ program (Interreg NWE), project Di-Plast Digital Circular Economy for the Plastics Industry (NWE729). Any support after en of project is not possible. The accuracy o table detection model depends on variou factors such as volume, variety of annotated datasets, hyperparameters of model. Please feel free to do your experiment of your table detection model

7. Then select *Technical Datasheet* name (or PDF file) for further processing.



How Datasheet Information Extraction Works

- Select Manufacturer and corresponding Technical Datasheets from dropdown lists.
- Click on **Extract Data** button.
- Go to **util -> data -> extractedinfo -> textualdata** folder and find sub-folder of selected manufacturer name. Access it to find text file which starts with selected technical datasheet name.

Manufacturer LyondellBasell

Technical Datasheet Circulen 2420D Plus.pdf

Contact:

For more technical details please contact Semantic Information Systems

User Information:

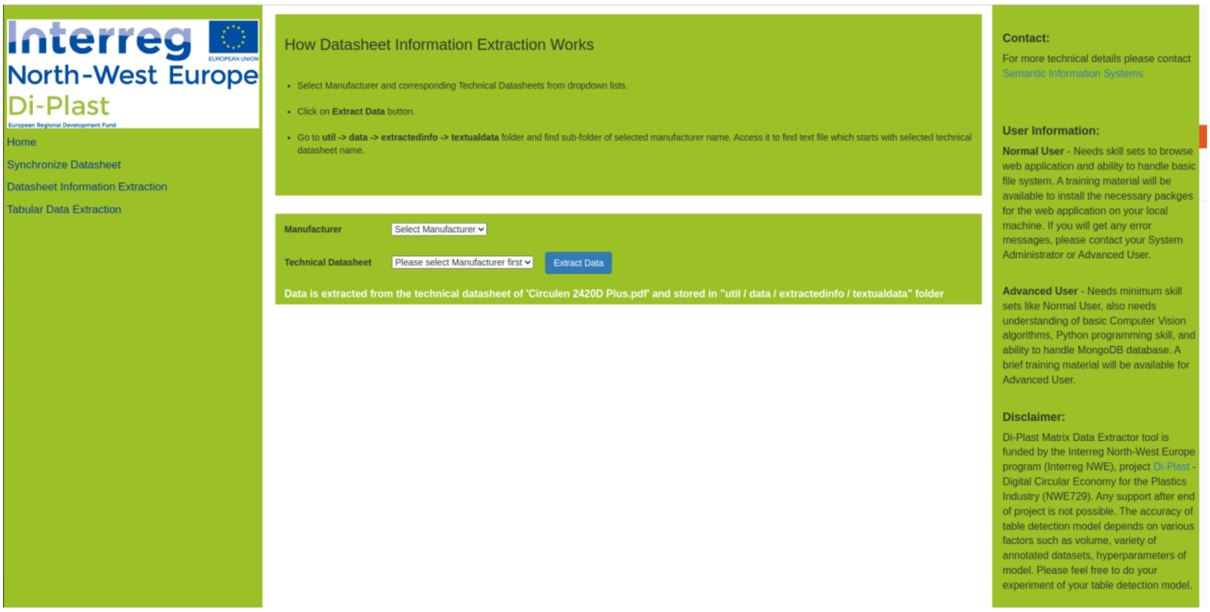
Normal User - Needs skill sets to browse web application and ability to handle basic file system. A training material will be available to install the necessary packges for the web application on your local machine. If you will get any error messages, please contact your System Administrator or Advanced User.

Advanced User - Needs minimum skill sets like Normal User, also needs understanding of basic Computer Vision algorithms, Python programming skill, and ability to handle MongoDB database. A brief training material will be available for Advanced User.

Disclaimer:

Di-Plast Matrix Data Extractor tool is funded by the Interreg North-West Europe program (Interreg NWE), project Di-Plast - Digital Circular Economy for the Plastics Industry (NWE729). Any support after end of project is not possible. The accuracy of table detection model depends on various factors such as volume, variety of annotated datasets, hyperparameters of model. Please feel free to do your experiment of your table detection model.

8. You can extract data from PDF files into textual format by clicking on ***Extract Data*** button. Data will be stored under */util/data/extractedinfo/textualdata* folder. You will get successful response (in white color) on webpage, if you have extracted data from PDF files as below-



How Datasheet Information Extraction Works

- Select Manufacturer and corresponding Technical Datasheets from dropdown lists.
- Click on **Extract Data** button.
- Go to **util -> data -> extractedinfo -> textualdata** folder and find sub-folder of selected manufacturer name. Access it to find text file which starts with selected technical datasheet name.

Manufacturer Select Manufacturer

Technical Datasheet Please select Manufacturer first

Data is extracted from the technical datasheet of 'Circulen 2420D Plus.pdf' and stored in "util / data / extractedinfo / textualdata" folder

Contact:

For more technical details please contact Semantic Information Systems

User Information:

Normal User - Needs skill sets to browse web application and ability to handle basic file system. A training material will be available to install the necessary packges for the web application on your local machine. If you will get any error messages, please contact your System Administrator or Advanced User.

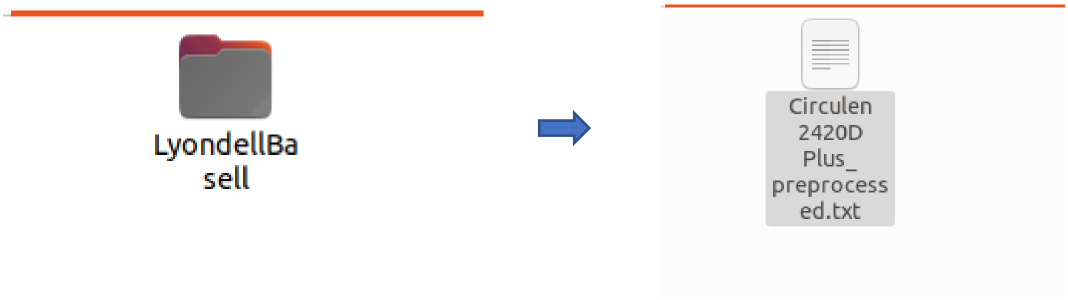
Advanced User - Needs minimum skill sets like Normal User, also needs understanding of basic Computer Vision algorithms, Python programming skill, and ability to handle MongoDB database. A brief training material will be available for Advanced User.

Disclaimer:

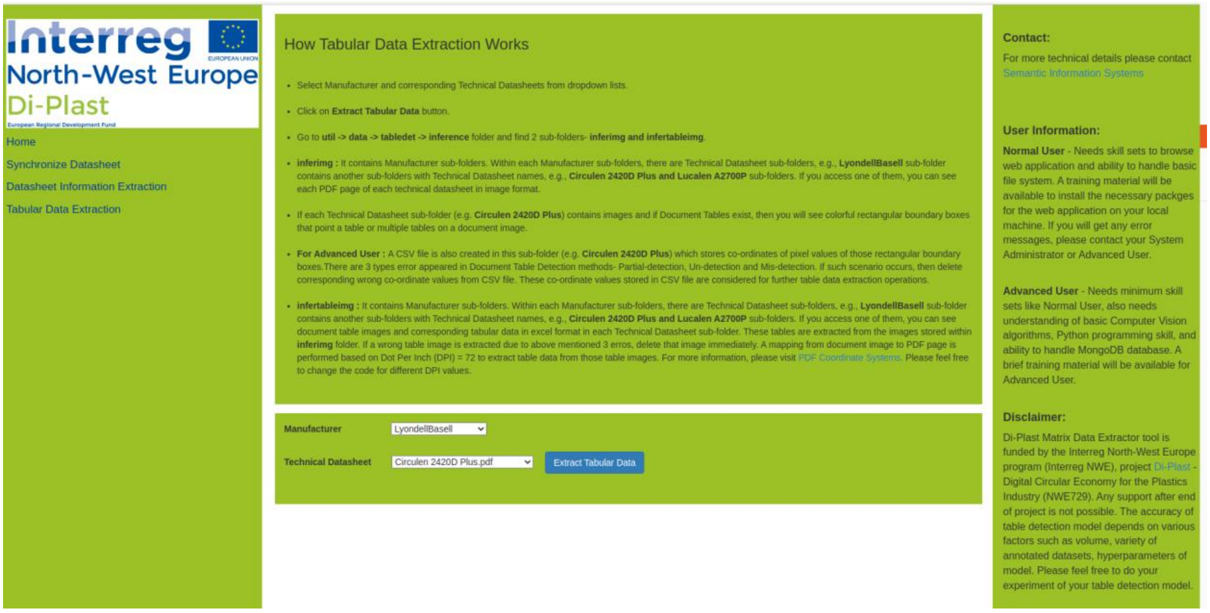
Di-Plast Matrix Data Extractor tool is funded by the Interreg North-West Europe program (Interreg NWE), project Di-Plast - Digital Circular Economy for the Plastics Industry (NWE729). Any support after end of project is not possible. The accuracy of table detection model depends on various factors such as volume, variety of annotated datasets, hyperparameters of model. Please feel free to do your experiment of your table detection model.

9. Please look above carefully that a sample notification is shown on web page after extracting 1 PDF document (e.g., Circulen 2420D Plus.pdf from LyondellBasell sub-folder is extracted) and stored PDF information in a text files.

10. Folder */util/data/extractedinfo/textualdata* contains *LyondellBasell* sub-folder, which also contains *Circulen 2420D Plus_preprocessed.txt* file as below. This unstructured textual information can be interesting for Elastic Search, Natural Language Processing (NLP) and Big Data technologies.



11. Go to ***Tabular Data Extraction*** link to click on ***Extract Tabular Data*** button.



How Tabular Data Extraction Works

- Select Manufacturer and corresponding Technical Datasheets from dropdown lists.
- Click on **Extract Tabular Data** button.
- Go to **util -> data -> tabledet -> inference** folder and find 2 sub-folders- **inferimg and infertableimg**.
- **inferimg :** It contains Manufacturer sub-folders. Within each Manufacturer sub-folders, there are Technical Datasheet sub-folders, e.g., **LyondellBasell** sub-folder contains another sub-folders with Technical Datasheet names, e.g., **Circulen 2420D Plus and Lucalen A2700P** sub-folders. If you access one of them, you can see each PDF page of each technical datasheet in image format.
- If each Technical Datasheet sub-folder (e.g. **Circulen 2420D Plus**) contains images and if Document Tables exist, then you will see colorful rectangular boundary boxes that point a table or multiple tables on a document image.
- **For Advanced User :** A CSV file is also created in this sub-folder (e.g. **Circulen 2420D Plus**) which stores co-ordinates of pixel values of those rectangular boundary boxes.There are 3 types error appeared in Document Table Detection methods- Partial-detection, Un-detection and Mis-detection. If such scenario occurs, then delete corresponding wrong co-ordinate values from CSV file. These co-ordinate values stored in CSV file are considered for further table data extraction operations.
- **infertableimg :** It contains Manufacturer sub-folders. Within each Manufacturer sub-folders, there are Technical Datasheet sub-folders, e.g., **LyondellBasell** sub-folder contains another sub-folders with Technical Datasheet names, e.g., **Circulen 2420D Plus and Lucalen A2700P** sub-folders. If you access one of them, you can see document table images and corresponding tabular data in excel format in each Technical Datasheet sub-folder. These tables are extracted from the images stored within **inferimg** folder. If a wrong table image is extracted due to above mentioned 3 erros, delete that image immediately. A mapping from document image to PDF page is performed based on Dot Per Inch (DPI) = 72 to extract table data from those table images. For more information, please visit PDF Coordinate Systems. Please feel free to change the code for different DPI values.

Manufacturer LyondellBasell

Technical Datasheet Circulen 2420D Plus.pdf

Contact:

For more technical details please contact Semantic Information Systems

User Information:

Normal User - Needs skill sets to browse web application and ability to handle basic file system. A training material will be available to install the necessary packges for the web application on your local machine. If you will get any error messages, please contact your System Administrator or Advanced User.

Advanced User - Needs minimum skill sets like Normal User, also needs understanding of basic Computer Vision algorithms, Python programming skill, and ability to handle MongoDB database. A brief training material will be available for Advanced User.

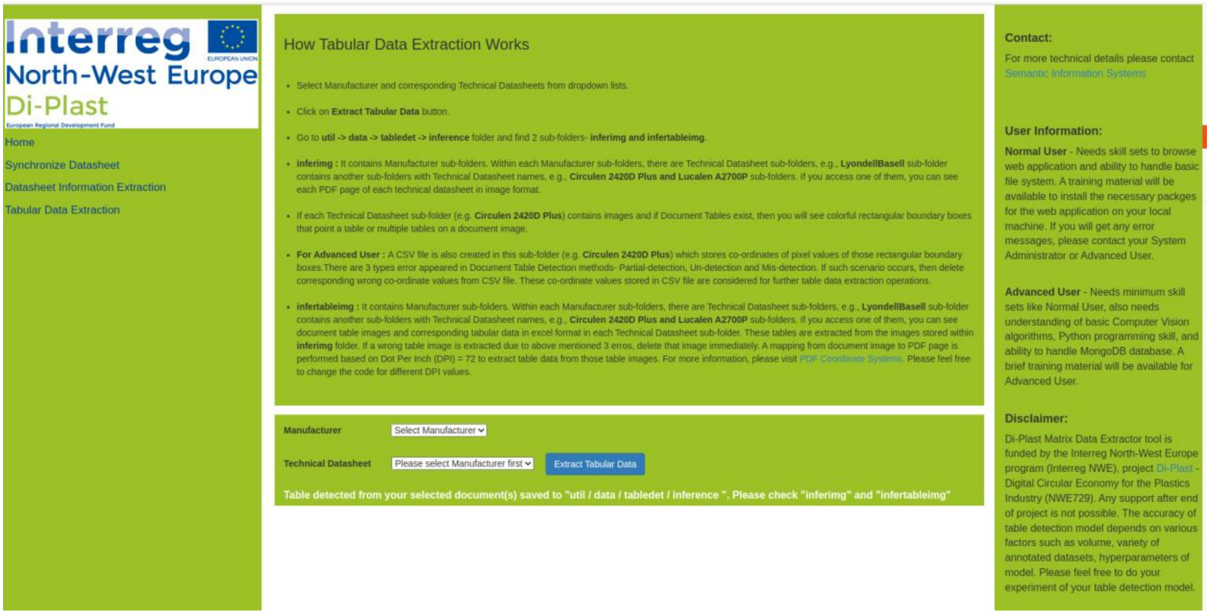
Disclaimer:

Di-Plast Matrix Data Extractor tool is funded by the Interreg North-West Europe program (Interreg NWE), project Di-Plast - Digital Circular Economy for the Plastics Industry (NWE729). Any support after end of project is not possible. The accuracy of table detection model depends on various factors such as volume, variety of annotated datasets, hyperparameters of model. Please feel free to do your experiment of your table detection model.

12. It identifies table regions of document images in a Rectangular Boundary Box (BBox) format under */util/data/tabledet/inference/inferimg* folder and stores BBox pixel information of document images in CSV files.

13. Simultaneously, you will also extract only document table images (cropped rectangular BBox region from document images) under /util/data/tabledet/inference/infertableimg folder and corresponding tabular data in excel files.

14. After clicking ***Extract Tabular Data*** button, you get successful response (in white color) on webpage. Now you access *document images and document table images both* along with tabular data in excel files.



How Tabular Data Extraction Works

- Select Manufacturer and corresponding Technical Datasheets from dropdown lists.
- Click on **Extract Tabular Data** button.
- Go to **util -> data -> tabledet -> inference** folder and find 2 sub-folders- **inferimg and infertableimg**.
- **inferimg :** It contains Manufacturer sub-folders. Within each Manufacturer sub-folders, there are Technical Datasheet sub-folders, e.g., **LyondellBasell** sub-folder contains another sub-folders with Technical Datasheet names, e.g., **Circulen 2420D Plus and Lucalen A2700P** sub-folders. If you access one of them, you can see each PDF page of each technical datasheet in image format.
- If each Technical Datasheet sub-folder (e.g. **Circulen 2420D Plus**) contains images and if Document Tables exist, then you will see colorful rectangular boundary boxes that point a table or multiple tables on a document image.
- **For Advanced User :** A CSV file is also created in this sub-folder (e.g. **Circulen 2420D Plus**) which stores co-ordinates of pixel values of those rectangular boundary boxes.There are 3 types error appeared in Document Table Detection methods- Partial-detection, Un-detection and Mis-detection. If such scenario occurs, then delete corresponding wrong co-ordinate values from CSV file. These co-ordinate values stored in CSV file are considered for further table data extraction operations.
- **infertableimg :** It contains Manufacturer sub-folders. Within each Manufacturer sub-folders, there are Technical Datasheet sub-folders, e.g., **LyondellBasell** sub-folder contains another sub-folders with Technical Datasheet names, e.g., **Circulen 2420D Plus and Lucalen A2700P** sub-folders. If you access one of them, you can see document table images and corresponding tabular data in excel format in each Technical Datasheet sub-folder. These tables are extracted from the images stored within **inferimg** folder. If a wrong table image is extracted due to above mentioned 3 erros, delete that image immediately. A mapping from document image to PDF page is performed based on Dot Per Inch (DPI) = 72 to extract table data from those table images. For more information, please visit PDF Coordinate Systems. Please feel free to change the code for different DPI values.

Manufacturer Select Manufacturer

Technical Datasheet Please select Manufacturer first

Table detected from your selected document(s) saved to "util / data / tabledet / inference ". Please check "inferimg" and "infertableimg"

Contact:

For more technical details please contact Semantic Information Systems

User Information:

Normal User - Needs skill sets to browse web application and ability to handle basic file system. A training material will be available to install the necessary packges for the web application on your local machine. If you will get any error messages, please contact your System Administrator or Advanced User.

Advanced User - Needs minimum skill sets like Normal User, also needs understanding of basic Computer Vision algorithms, Python programming skill, and ability to handle MongoDB database. A brief training material will be available for Advanced User.

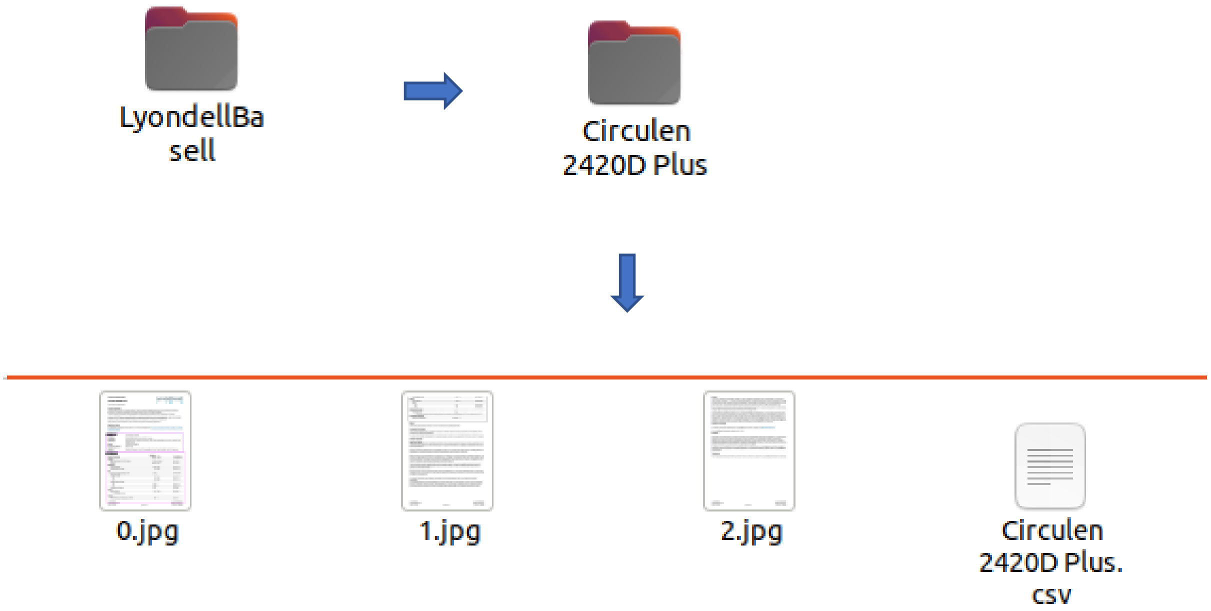
Disclaimer:

Di-Plast Matrix Data Extractor tool is funded by the Interreg North-West Europe program (Interreg NWE), project Di-Plast - Digital Circular Economy for the Plastics Industry (NWE729). Any support after end of project is not possible. The accuracy of table detection model depends on various factors such as volume, variety of annotated datasets, hyperparameters of model. Please feel free to do your experiment of your table detection model.

15. Above functionality involves *deep learning model inference* to identify table regions in rectangular BBox format and cropped that regions to save table images. Folder /util/data/tabledet/inference/*inferimg* contains-

- *Manufacturer* sub-folder

- *Technical Datasheet* sub-folder
- *Document images* along with BBox inference information in CSV file as below-



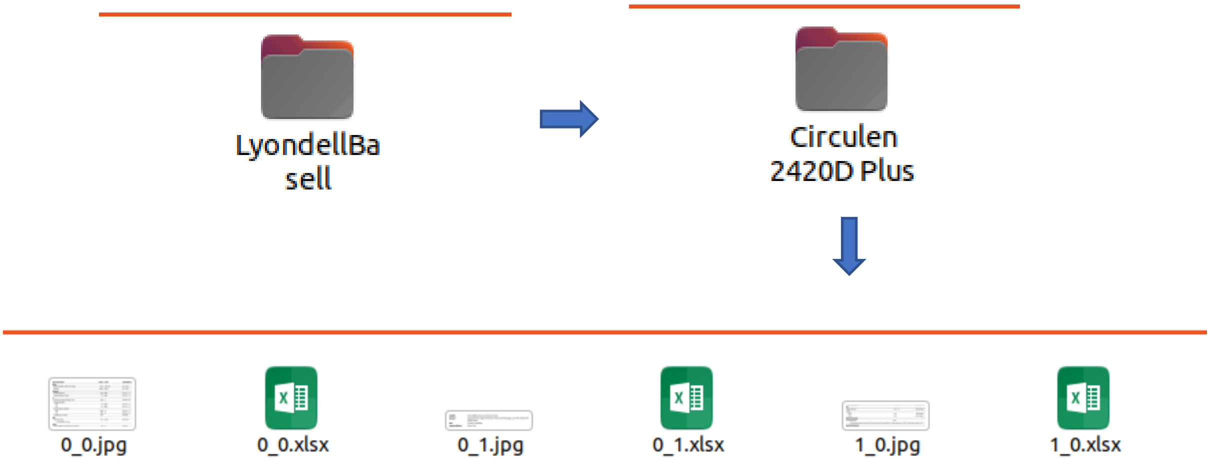
16. The document image pixels to PDF co-ordinates mapping is performed (https://www.pdfscripting.com/public/PDF-Page-Coordinates.cfm) to identify table regions on each PDF pages by considering DPI (dot per inch) value=72. DPI value is generally used to map digital images to physical pages (e.g. A4 page).

17. If you change DPI value other than 72, and your technical datasheets (or PDF files) are not A4 types, then feel free to adapt code changes to incorporate customized DPI value at *MatrixDataExtractor/backend/matrixtextapp/cv_basic_service.py*

18. When image pixel values to PDF co-ordinate values mapping is performed, *Camelot* python package is used with parameters *table_areas* and *flavor='stream'* to extract tabular data from PDF files in excel format.

19. Folder /util/data/tabledet/inference/*infertableimg* contains-

- *Manufacturer* sub-folder
- *Technical datasheet* sub-folder
- *Document table images* along with corresponding excel files as below-



20. Folder */util/data/tempimg* contains *tabledata* and *textdata* sub-folders for pre-processing purpose. Please delete all sub-folders and files within *tabledata* and *textdata* sub-folders when you finish your information extraction task.

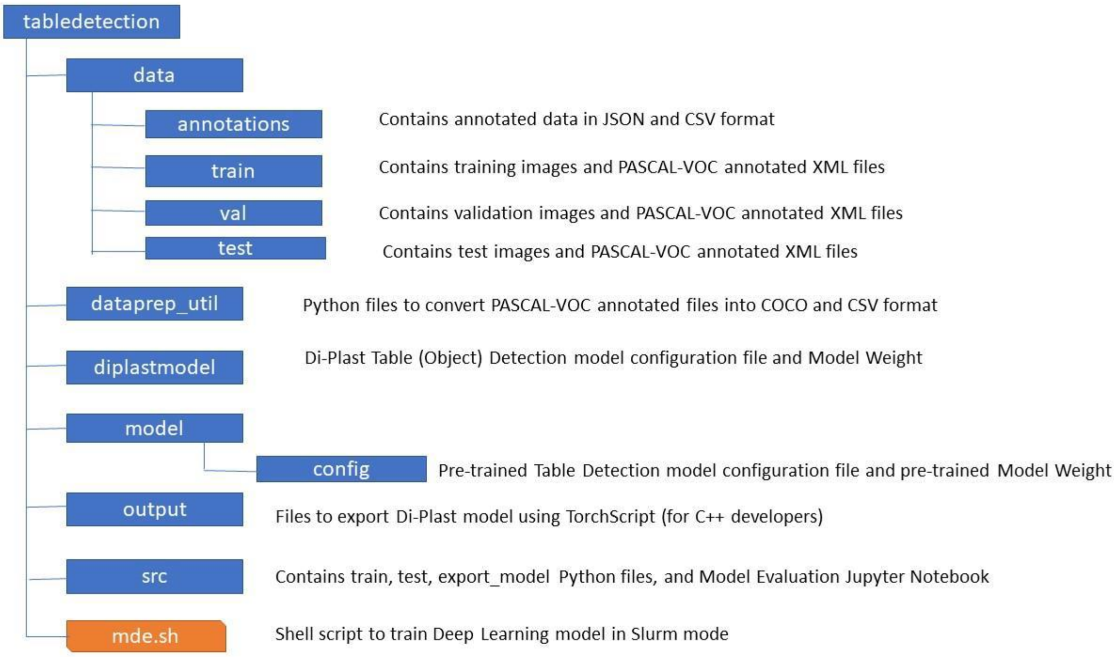
Table Detection- Deep Learning Model

Installation:

Please check GitHub page for installation. The Anaconda environment ***env_mde*** needs to be created to train and to evaluate model (preferably in GPU server).

Folder Structure:

The folder structure is shown below-



tabledetection	
data	
annotations	Contains annotated data in JSON and CSV format
train	Contains training images and PASCAL-VOC annotated XML files
val	Contains validation images and PASCAL-VOC annotated XML files
test	Contains test images and PASCAL-VOC annotated XML files
dataprep_util	Python files to convert PASCAL-VOC annotated files into COCO and CSV format
diplastmodel	Di-Plast Table (Object) Detection model configuration file and Model Weight
model	
config	Pre-trained Table Detection model configuration file and pre-trained Model Weight
output	Files to export Di-Plast model using TorchScript (for C++ developers)
src	Contains train, test, export_model Python files, and Model Evaluation Jupyter Notebook
mde.sh	Shell script to train Deep Learning model in Slurm mode

Configuration:

1. Download pre-trained *TableBank* (*faster_rcnn_R_101_FPN_3x*) model config file and pre-trained model weight from *Layout-Parser* GitHub page (mentioned in *catalog.py* python file) from below URL- https://github.com/Layout-Parser/layout-parser/tree/main/src/layoutparser/models/detectron2

2. Save those files in *MatrixDataExtractor/tabledetection/model/configs* folder.

Image Annotation:

1. You can use any *Image Annotation* tool (e.g. *LabelImg*) to annotate ***Table*** images for Supervised Learning. If you have store data in PASCAL-VOC (XML) format, then you can convert XML annotated files into COCO (JSON) format. Also you need to convert annotated image information into CSV format for Di-Plast table detection model. The utility functions are referred in *tabledetection/dataprep_util* folder.
2. You can store JSON and CSV annotated information in *tabledetection/data/annotations* folder.
3. Keep your images and corresponding annotated XML files (PASCAL-VOC) in *tabledetection/data/train, tabledetection/data/val, tabledetection/data/test* folders.

Model Training:

1. In training, Di-Plast Table Detection model configuration file (*faster_rcnn_R_101_FPN_3x_config.yaml or uos_dip_config.yaml*) and model training weight (*model_final.pth*) are saved in *tabledetection/diplastmodel* folder.
2. Run *tabledetection/src/train.py* script for model training. A shell script (*mde.sh*) is provided to train model in *Slurm* mode.

Important Note:

- You need to save *model_final.pth* and *faster_rcnn_R_101_FPN_3x_config.yaml or uos_dip_config.yaml* files into */util/data/tabledet/ modelweight* folder in MDE web application for model inference.

Model Evaluation:

- Evaluate model by running *tabledetection/src/test.py* script. The Jupyter Notebook (*Eval_DiPlast_TableDetection_AP75 .ipynb*) is provided to evaluate the model and to visualize the inferred images.

Optional- Export Model (for C++ Developers):

- *TorchScript*: Export model by running *tabledetection/src/export_model.py* script and saved model in *tabledetection/output* folder as *model.ts* format. This can be used in C++ development to infer Table images.

For more transfer learning based document table detection research work, please check below research paper-

Chowdhury, Arnab Ghosh, Nils Schut, and Martin Atzmüller. "A Hybrid Information Extraction Approach using Transfer Learning on Richly-Structured Documents." LWDA. 2021. (http://ceur-ws.org/Vol-2993/paper-02.pdf)