



Matrix Data Extractor (MDE) Tool User Interface

Arnab Ghosh Chowdhury, Prof. Dr. Martin Atzmueller

Semantic Information Systems, Institute of Computer Science
Osnabrueck University, Germany

For Di-Plast project specific use only

14th September, 2022



MDE Tool Functionality

- Homepage
- Synchronize Datasheet
- Datasheet Information Extraction
- Tabular Data Extraction
- Detailed information mentioned in User Guide at Knowledge Hub wiki and in GitHub

Table Detection Deep Learning Model

- Primarily Computer Vision based approach
- Transfer learning based Object Detection problem for document table detection task
- Use TableBank pre-trained model
- Manually create domain specific annotated dataset for table detection (e.g Labelling tool)
- Train table detection deep learning model and test the model
- Incorporate model description file (XML file) and model weight file into Django based MDE web application
- Use model description file and model weight file to infer new data
- Map image pixel values to PDF co-ordinate values using Camelot python package to extract tabular data

Document Image vs Document Table Image

- Document Image - Each PDF page converted into image format.
- Document Table Image - Each document table from each document image.

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-activated, 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 \(SDS\)](#).

Table 99%

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

Table 100%

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/cm ³	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: 02/10/201

Page 1 of 3

Circulen 2420D Plus
Recipient Tracking #
Request #: 300897

Fig. Document Image

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/cm ³	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

Fig. Document Table Image

Homepage

- Describe information about basic functionalities of MDE tool



European Regional Development Fund

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 packages for the web application on your local machine. If you will get any error messages, please contact your System Administrator or Advanced User.

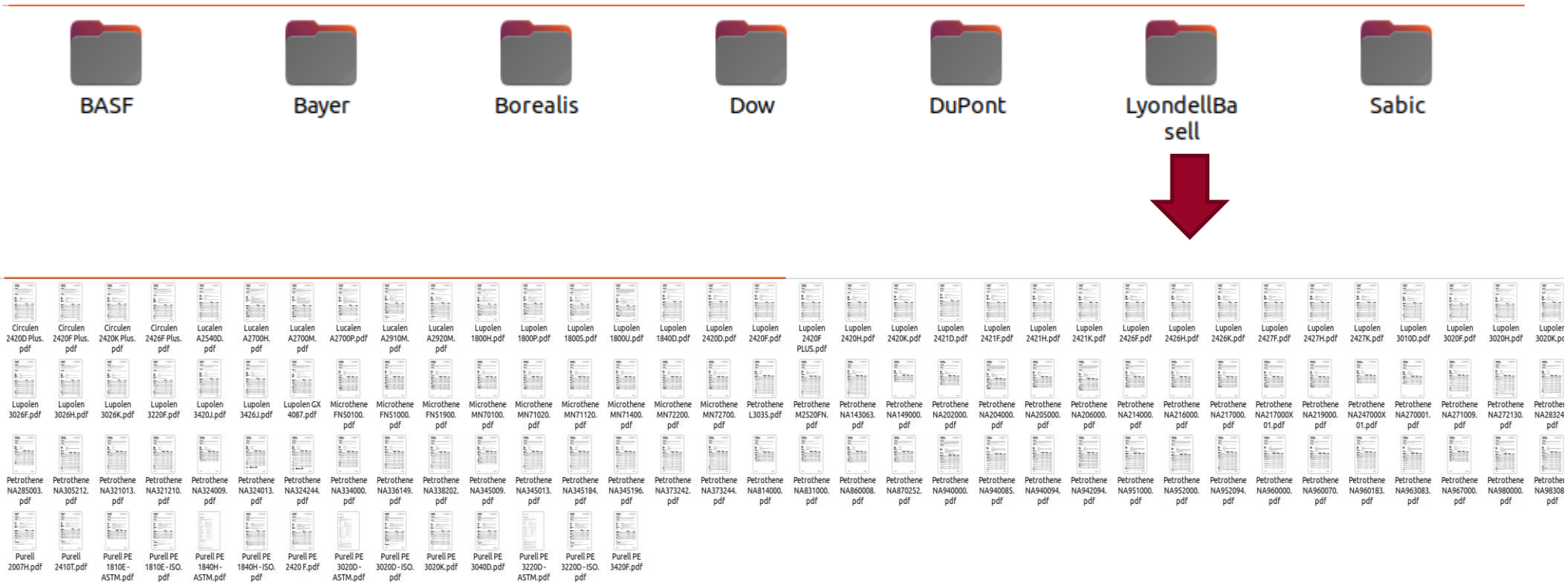
Advanced User - Needs minimum skill sets like Normal User, also needs understanding of basic Computer Vision algorithms, Python programming skill, and ability to handle MongoDB database. A brief training material will be available for Advanced User.

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.

Synchronize Datasheet - I

- Create sub-folders of Manufacturer names and place relevant technical datasheets within each sub-folder
- For example, LyondellBasell folder contains PDF files (technical datasheets)



Synchronize Datasheet - II

- Synchronize manufacturer names and relevant datasheets with MongoDB database
- Get manufacturer names and relevant filenames in dropdown menu on web UI (User Interface)

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

- Home
- Synchronize Datasheet
- Datasheet Information Extraction
- Tabular Data Extraction

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. After that 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](#).

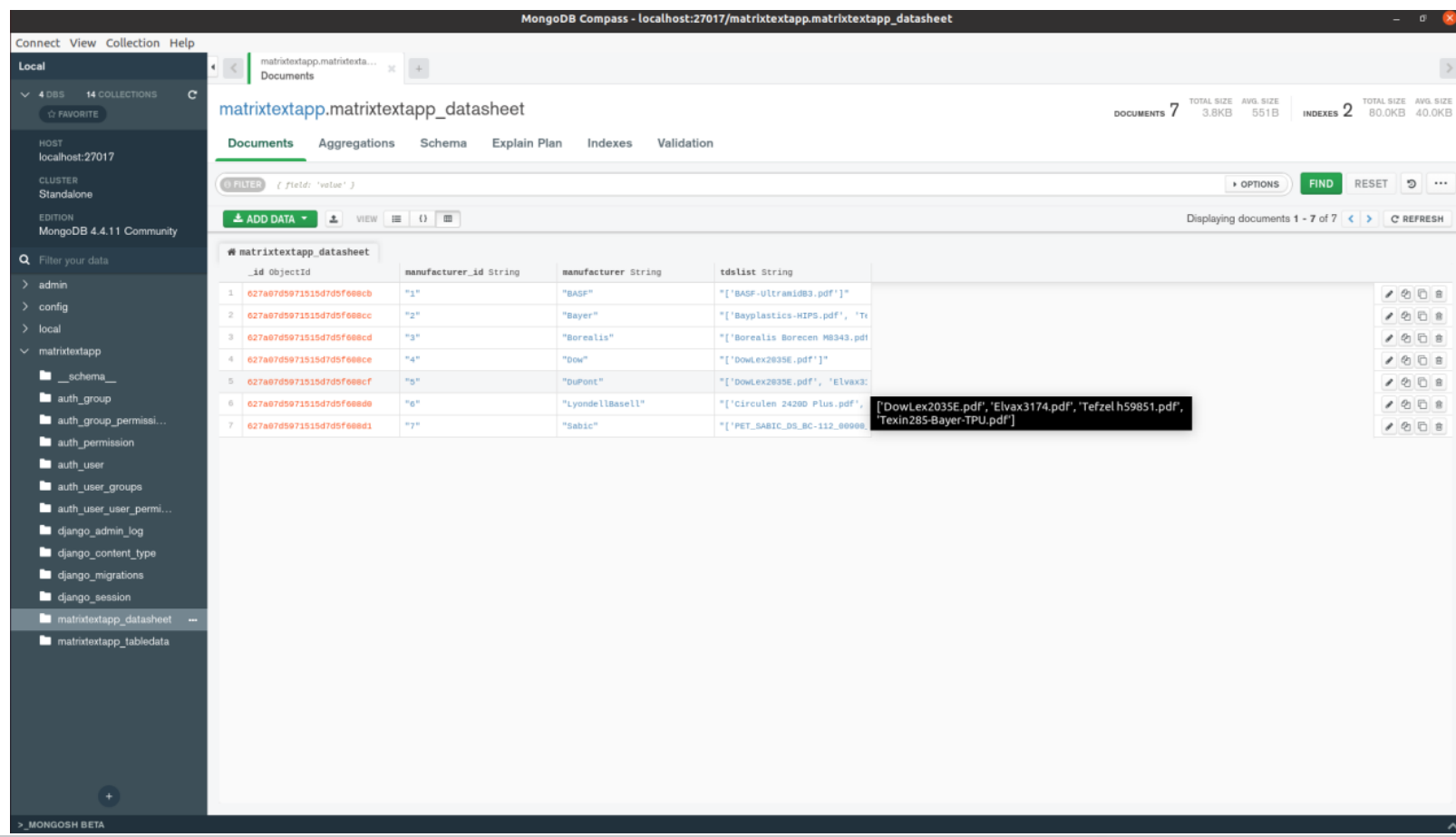
User Information:
Normal User - Needs skill sets to browse web application and ability to handle basic file system. A training material will be available to install the necessary packages for the web application on your local machine. If you will get any error messages, please contact your System Administrator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

Synchronize Datasheet - III

- Manufacturer names and relevant PDF filenames shown in MongoDB database UI (MongoDB Compass)



Datasheet Information Extraction - I

- Extract textual information (unstructured data) from PDF files into text files
- Can be useful for Big Data Analysis and Natural Language Processing (NLP)

How Datasheet Information Extraction Works

- Select Manufacturer and corresponding Technical Datasheets from dropdown lists.
- Click on **Extract Data** button.
- Go to url -> 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:

Technical Datasheet:

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 packages 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.

How Datasheet Information Extraction Works

- Select Manufacturer and corresponding Technical Datasheets from dropdown lists.
- Click on **Extract Data** button.
- Go to url -> 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:

Technical Datasheet:

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 packages 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.

Datasheet Information Extraction - II

- Extract Data button - Extracts textual information from PDF files into text files



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

Technical Datasheet Extract Data

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 packages 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.

Datasheet Information Extraction - III

- Notification for textual information extraction (in White color message)



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

Technical Datasheet

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 packages for the web application on your local machine. If you will get any error messages, please contact your System Administrator or Advanced User.

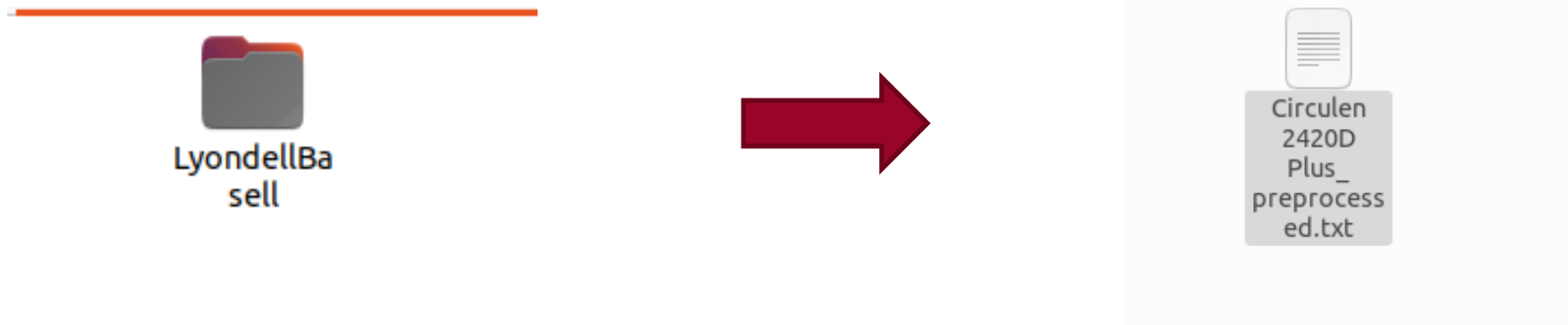
Advanced User - Needs minimum skill sets like Normal User, also needs understanding of basic Computer Vision algorithms, Python programming skill, and ability to handle MongoDB database. A brief training material will be available for Advanced User.

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.

Datasheet Information Extraction - IV

- Confirmation of textual information extraction
- For example, LyondellBasell sub-folder and Circulen 2420D Plus_preprocessed.txt file created to store textual data



Tabular Data Extraction - I

- Extract tabular data from PDF files into excel files
- Useful for Di-Plast Matrix application

Interreg North-West Europe Di-Plast

Home
Synchronize Datasheet
Datasheet Information Extraction
Tabular Data Extraction

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- `infering` and `infertableimg`
- **infering** : 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 `infering` folder. If a wrong table image is extracted due to above mentioned 3 errors, 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:
 Technical Datasheet: **Extract Tabular Data**

Manufacturer dropdown options: BASF, Bayer, Borealis, Dow, DuPont, LyondellBasell, Sabic

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 packages 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.

Interreg North-West Europe Di-Plast

Home
Synchronize Datasheet
Datasheet Information Extraction
Tabular Data Extraction

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- `infering` and `infertableimg`
- **infering** : 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 `infering` folder. If a wrong table image is extracted due to above mentioned 3 errors, 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:
 Technical Datasheet: **Extract Tabular Data**

Technical Datasheet dropdown options: Please select Manufacturer first, Borealis Borecan M8343.pdf, Borealis-MG 9641.pdf, Borealis_PP_EF03SAE.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 packages 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.

Tabular Data Extraction - II

- Extract Tabular Data button – Extract tabular data from PDF files into excel files using Camelot python package



Home

Synchronize Datasheet

Datasheet Information Extraction

Tabular Data Extraction

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- inferring and infertableimg.
- **inferring** : 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 **inferring** folder. If a wrong table image is extracted due to above mentioned 3 errors, 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

Technical Datasheet Extract Tabular Data

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 packages 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.

Tabular Data Extraction - III

- Notification for tabular data extraction (in White color message)



Home

[Synchronize Datasheet](#)

[Datasheet Information Extraction](#)

[Tabular Data Extraction](#)

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 errors, 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

Technical Datasheet

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 packages for the web application on your local machine. If you will get any error messages, please contact your System Administrator or Advanced User.

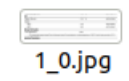
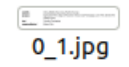
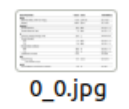
Advanced User - Needs minimum skill sets like Normal User, also needs understanding of basic Computer Vision algorithms, Python programming skill, and ability to handle MongoDB database. A brief training material will be available for Advanced User.

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.

Tabular Data Extraction - IV

- Confirmation of tabular data extraction
- For example, LyondellBasell and Circulen 2420D Plus sub-folders created to store tabular data
- Circulen 2420D Plus sub-folder contains table images and relevant excel files



Please visit our GitHub profile for more details:

<https://github.com/cslab-hub/MatrixDataExtractor>