UNIVERSITÄT OSNABRÜCK

Matrix Data Extractor (MDE) Tool User Interface

Arnab Ghosh Chowdhury, Prof. Dr. Martin Atzmueller

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

Semantische Informationssysteme

Interreg North-West Europe Di-Plast

EUROPEAN UNION

European Regional Development Fund

For Di-Plast project specific use only

14th September, 2022

MDE Tool Functionality

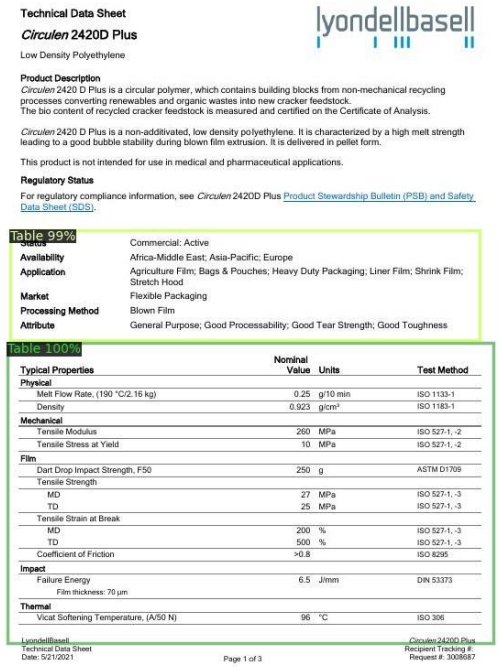
- Homepage
- Synchronize Datasheet
- Datasheet Information Extraction
- Tabular Data Extraction
- Detailed information mentioned in User Guide at Knowledge Hub wiki and in GitHub

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 LabelImg 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-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).

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/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	6.5	J/mm	DIN 53373
Film thickness: 70 µm			
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 Tracking #:
Request #: 3008687

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

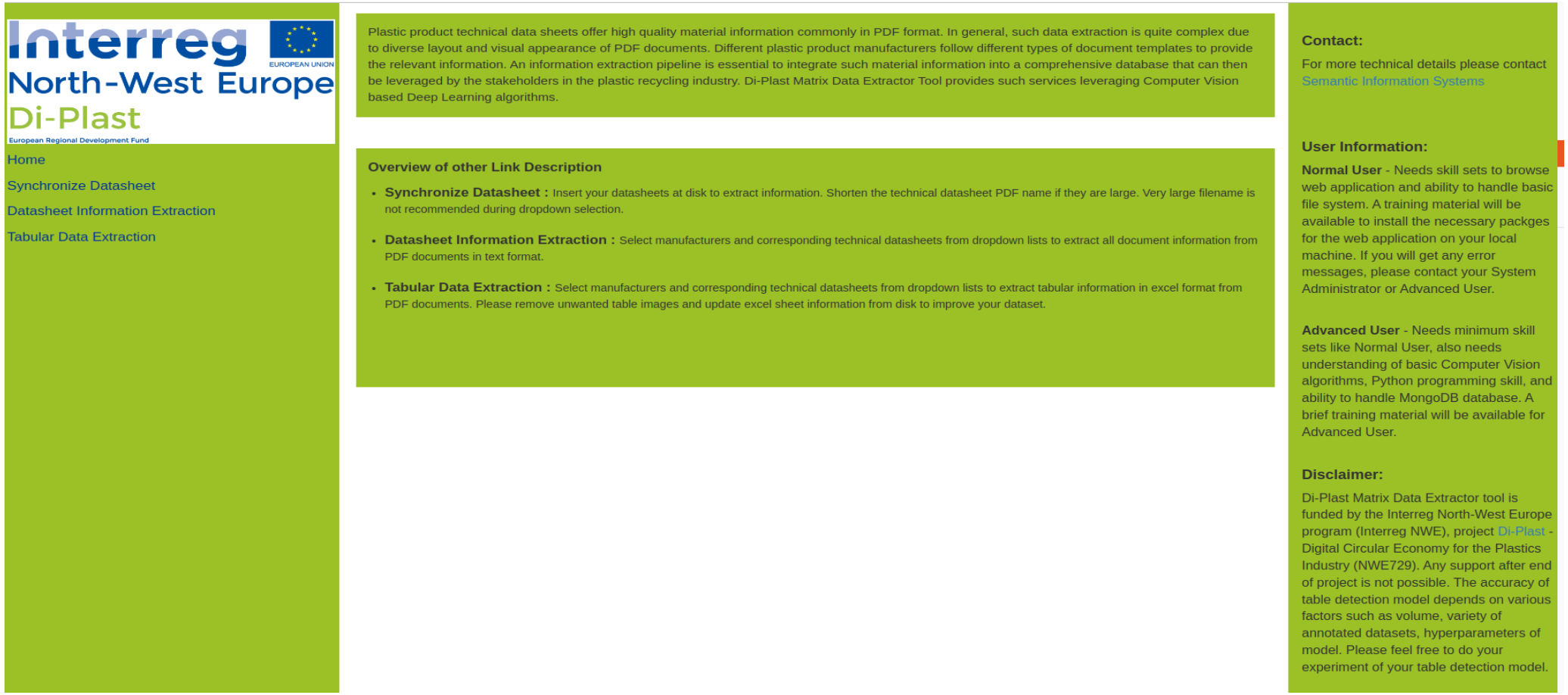
Typical Properties	Nominal Value	Units	Test Method
Physical			
Melt Flow Rate, (190 °C/2.16 kg)	0.25	g/10 min	ISO 1133-1
Density	0.923	g/cm³	ISO 1183-1
Mechanical			
Tensile Modulus	260	MPa	ISO 527-1, -2
Tensile Stress at Yield	10	MPa	ISO 527-1, -2
Film			
Dart Drop Impact Strength, F50	250	g	ASTM D1709
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	6.5	J/mm	DIN 53373
Film thickness: 70 µm			
Thermal			
Vicat Softening Temperature, (A/50 N)	96	°C	ISO 306

LyondellBasell *Circulen* 2420D Plus

Fig. Document Table Image

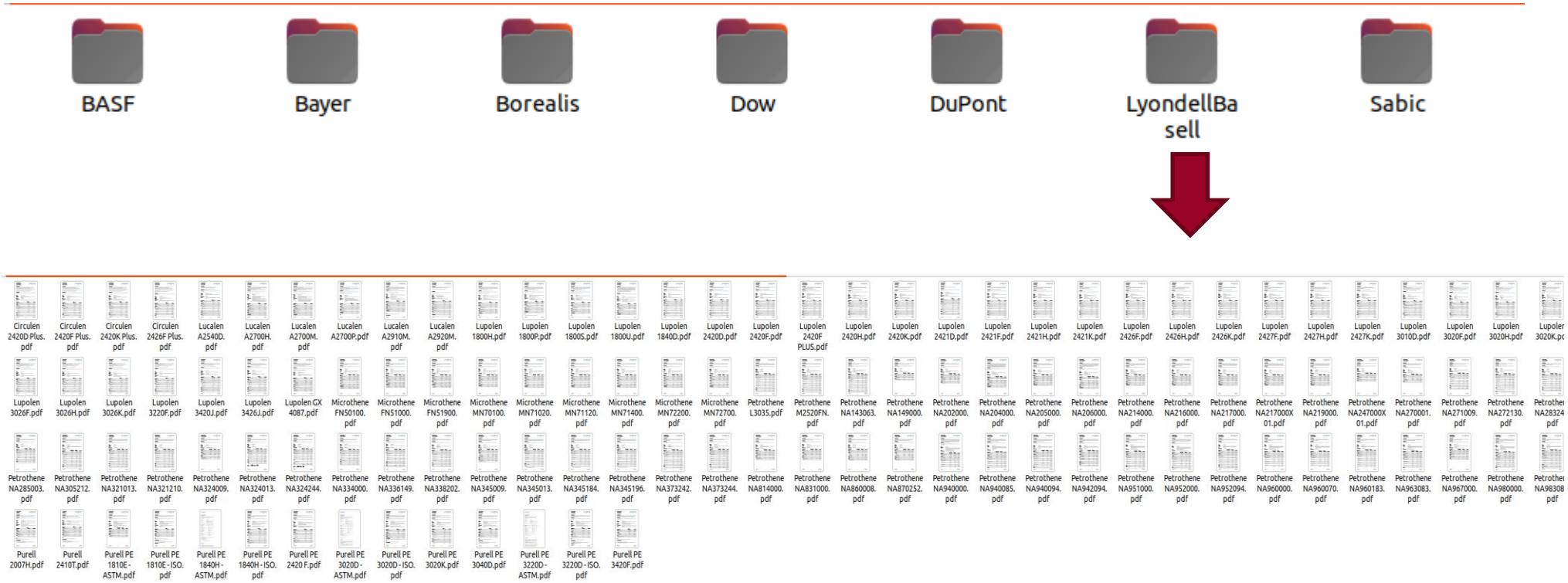
Homepage

- Describe information about basic functionalities of MDE tool



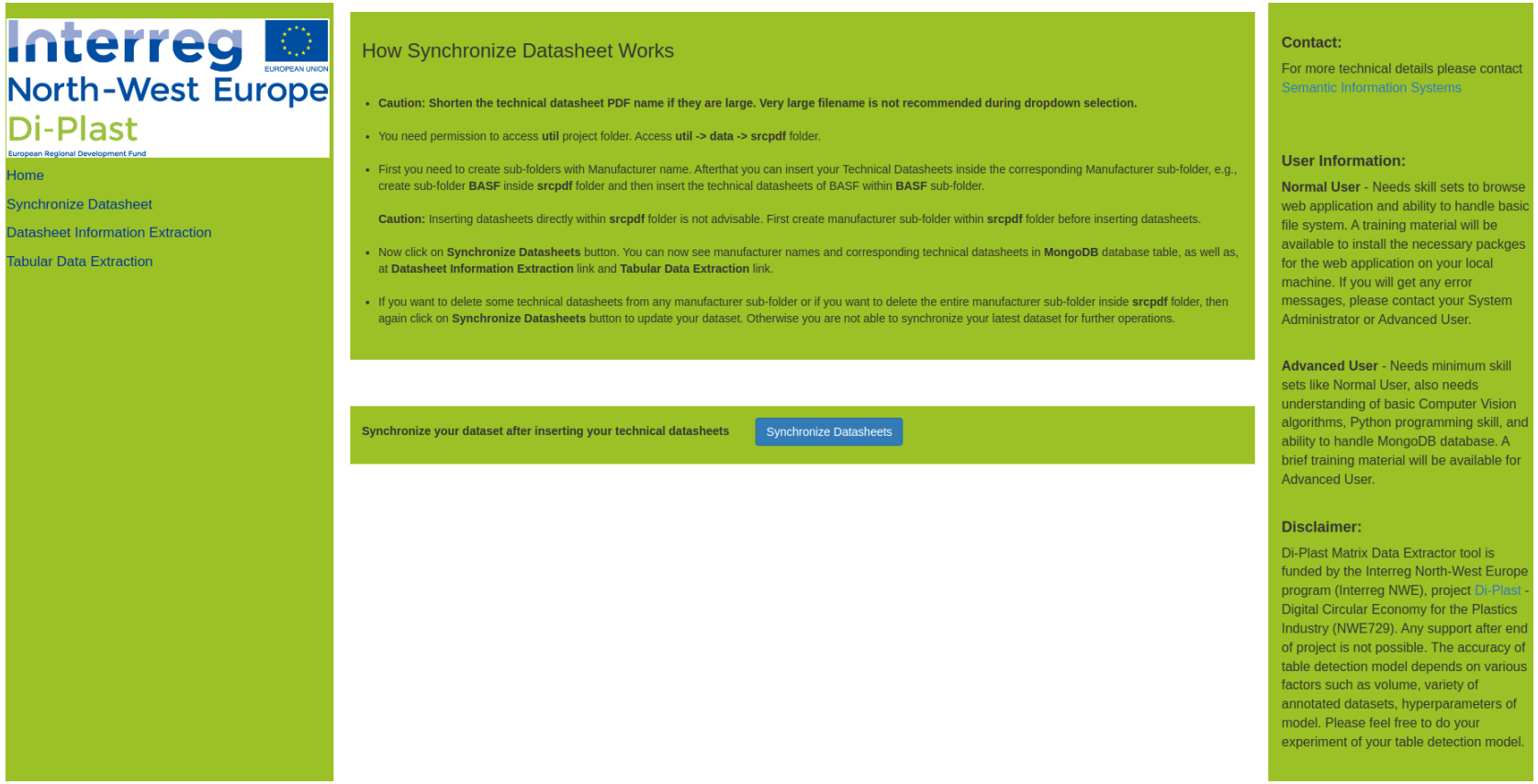
Synchronize Datasheet - I

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



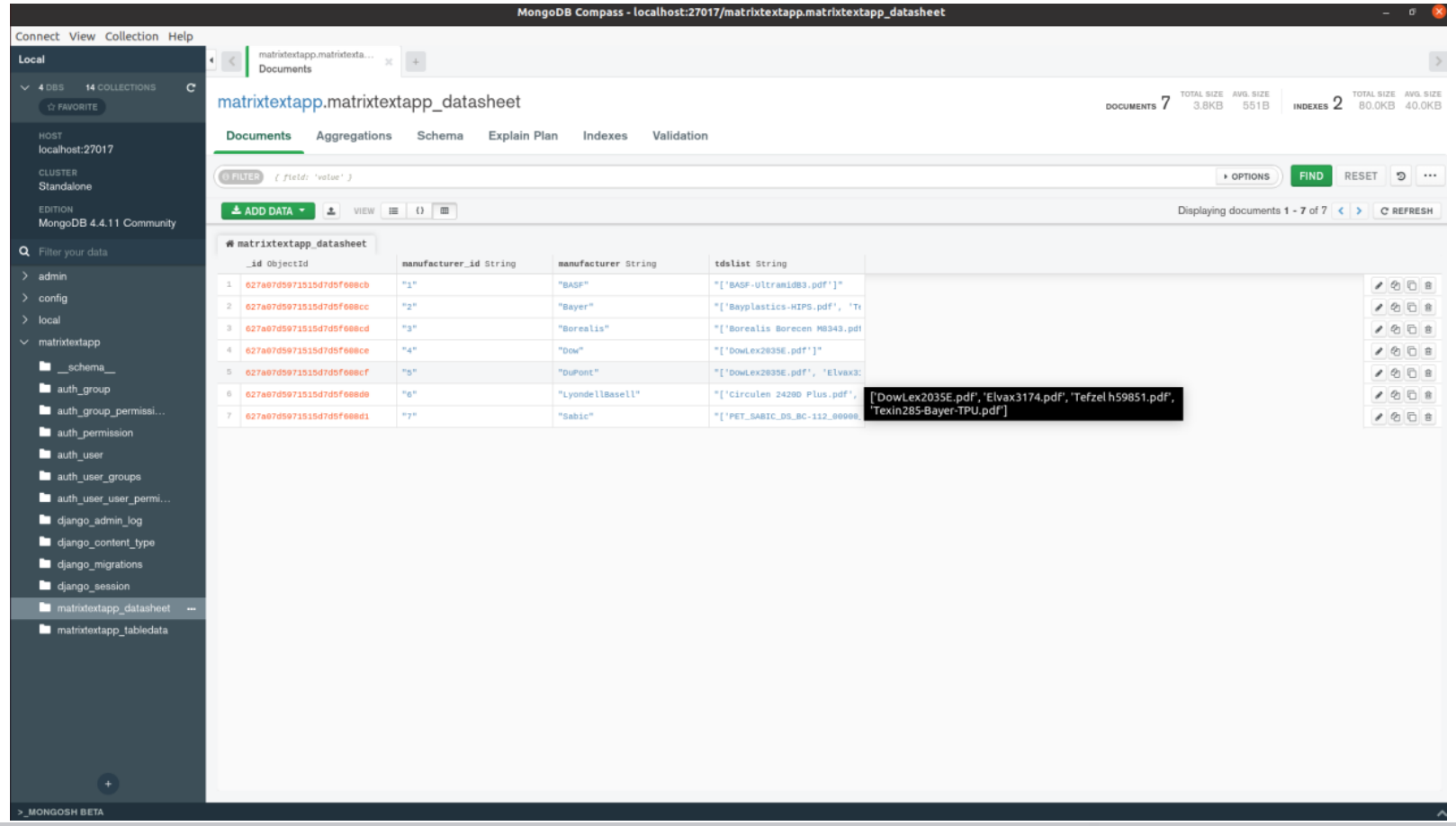
Synchronize Datasheet - II

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



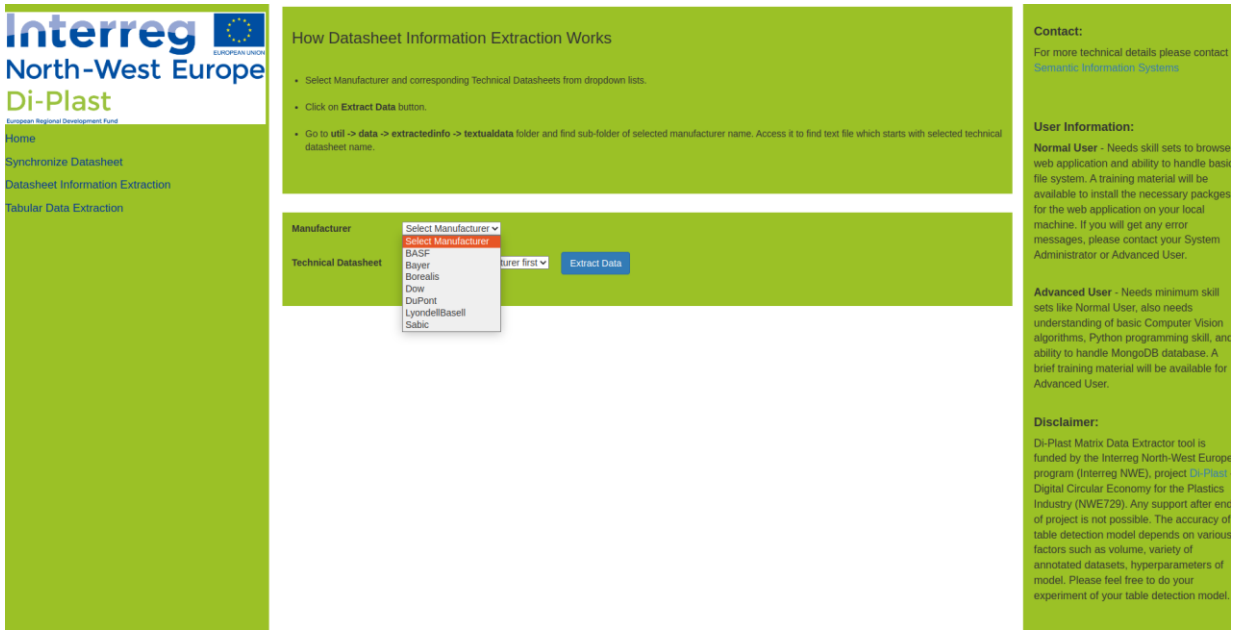
Synchronize Datasheet - III

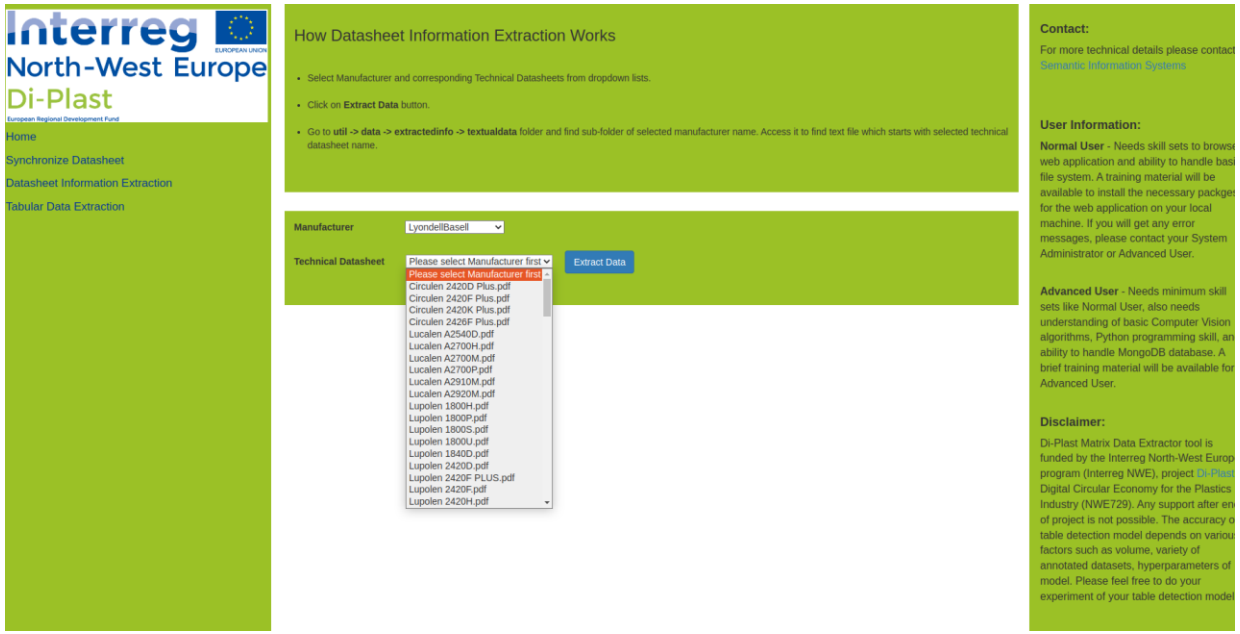
- Manufacturer names and relevant PDF filenames shown in MongoDB database UI (MongoDB Compass)



Datasheet Information Extraction - I

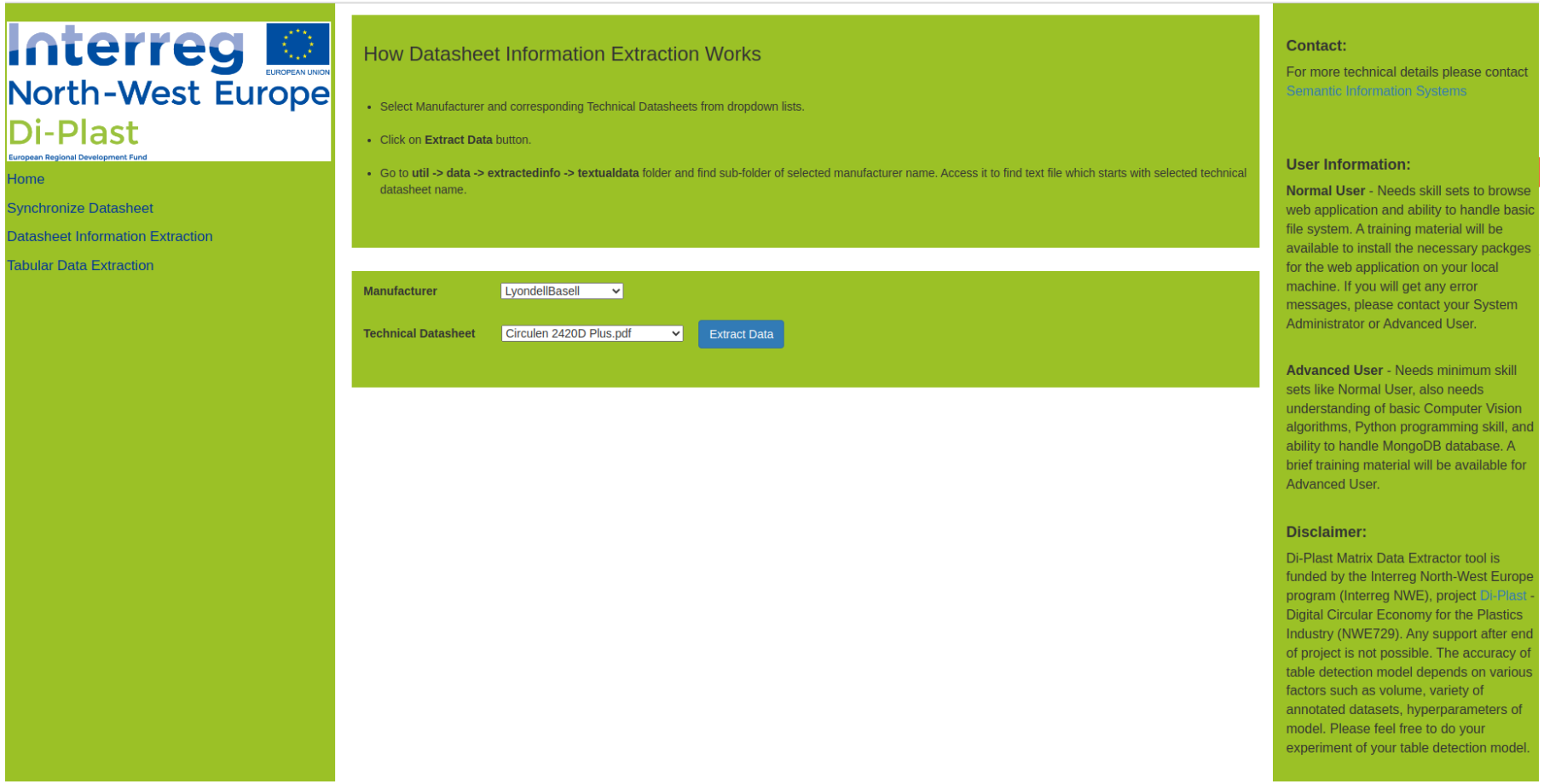
- Extract textual information (unstructured data) from PDF files into text files
- Can be useful for Big Data Analysis and Natural Language Processing (NLP)





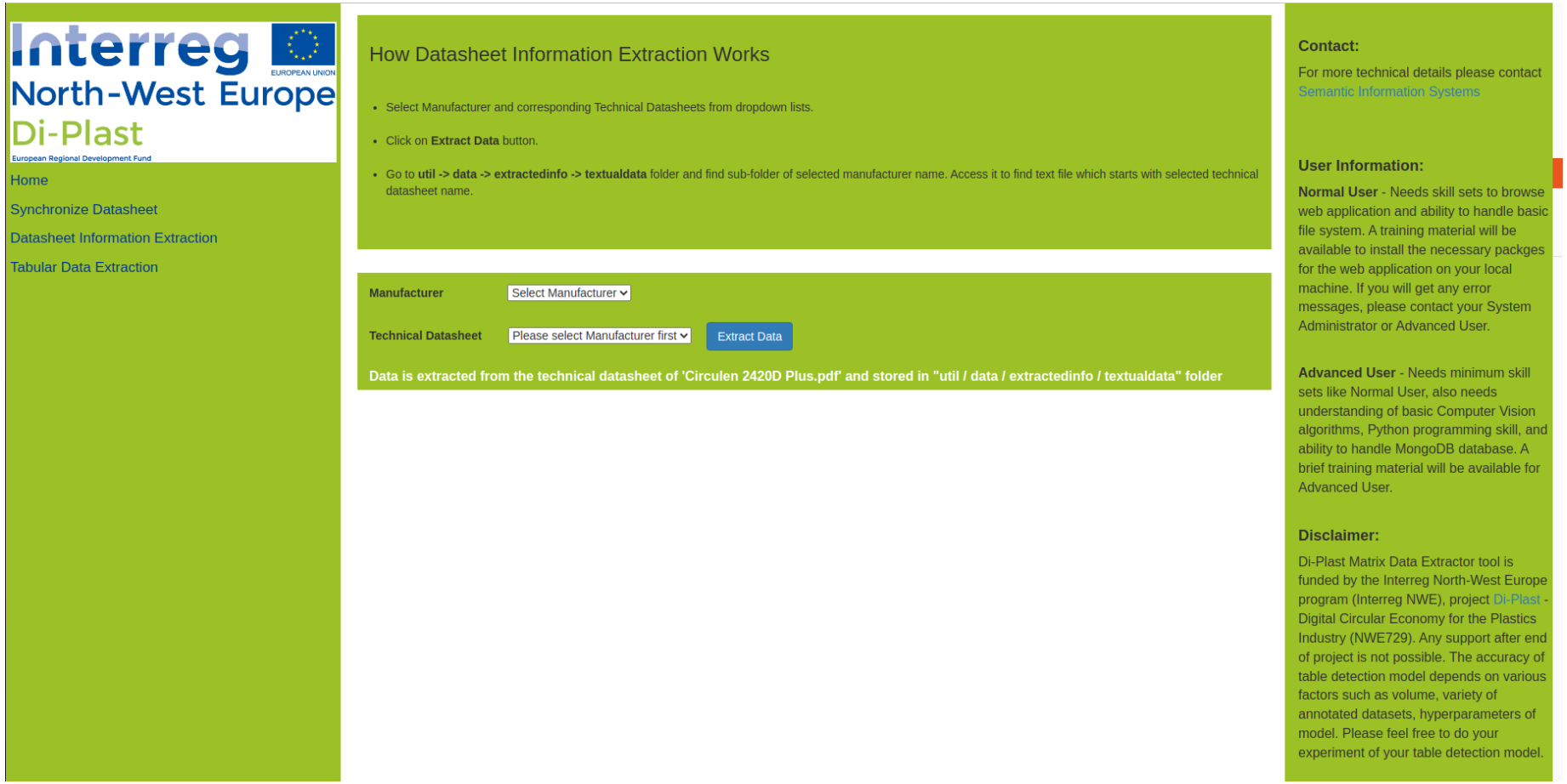
Datasheet Information Extraction - II

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



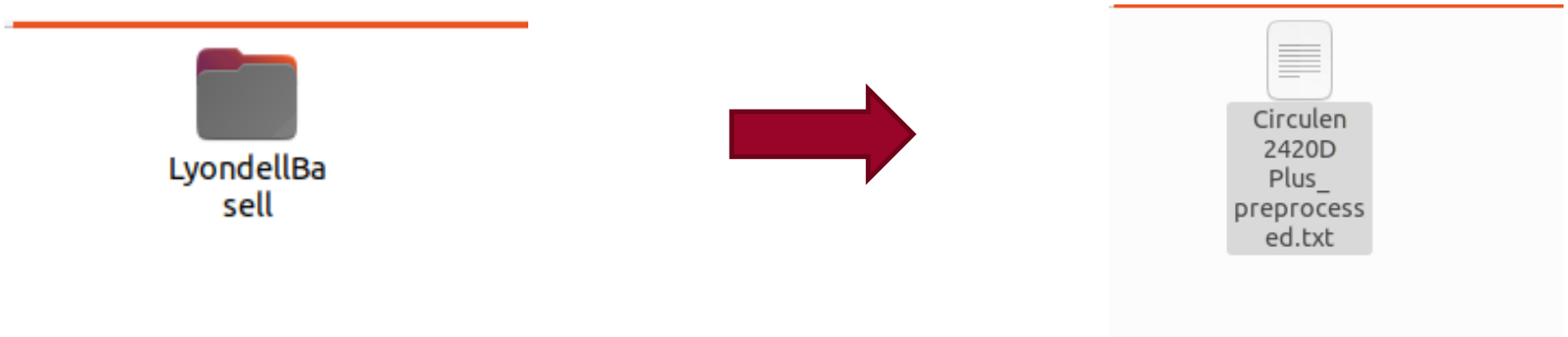
Datasheet Information Extraction - III

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



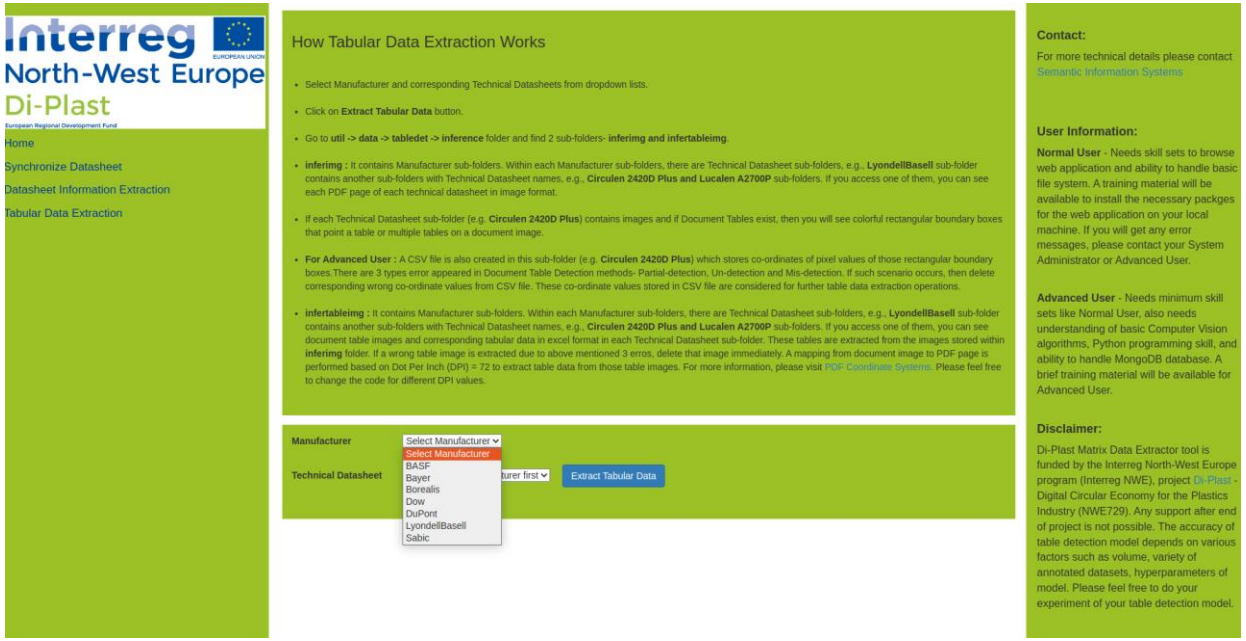
Datasheet Information Extraction - IV

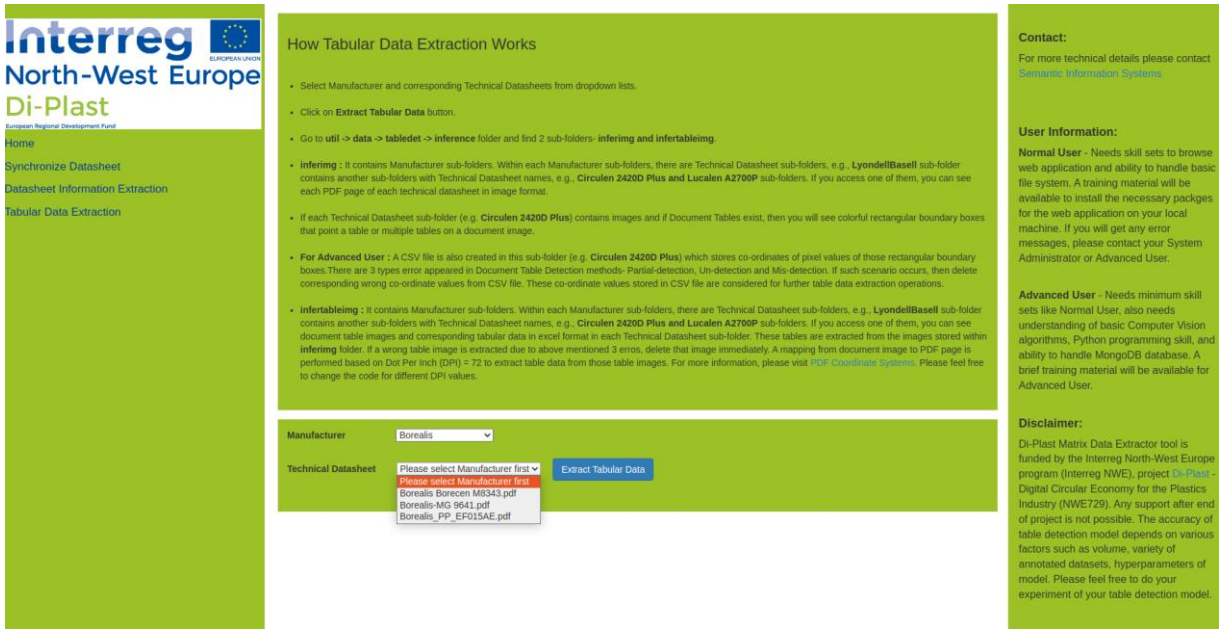
- Confirmation of textual information extraction
- For example, LyondellBasell sub-folder and Circulen 2420D Plus_preprocessed.txt file created to store textual data



Tabular Data Extraction - I

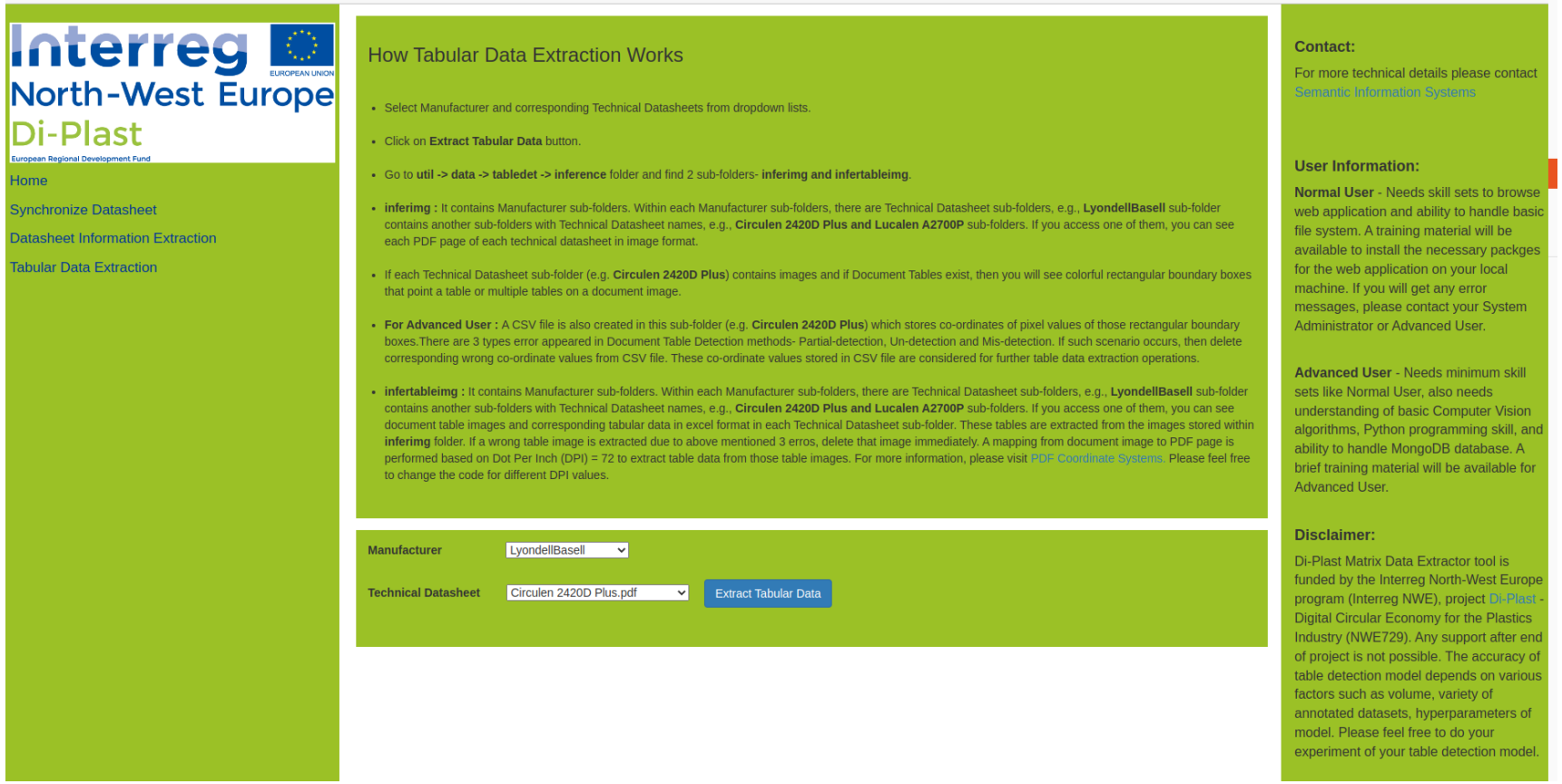
- Extract tabular data from PDF files into excel files
- Useful for Di-Plast Matrix application





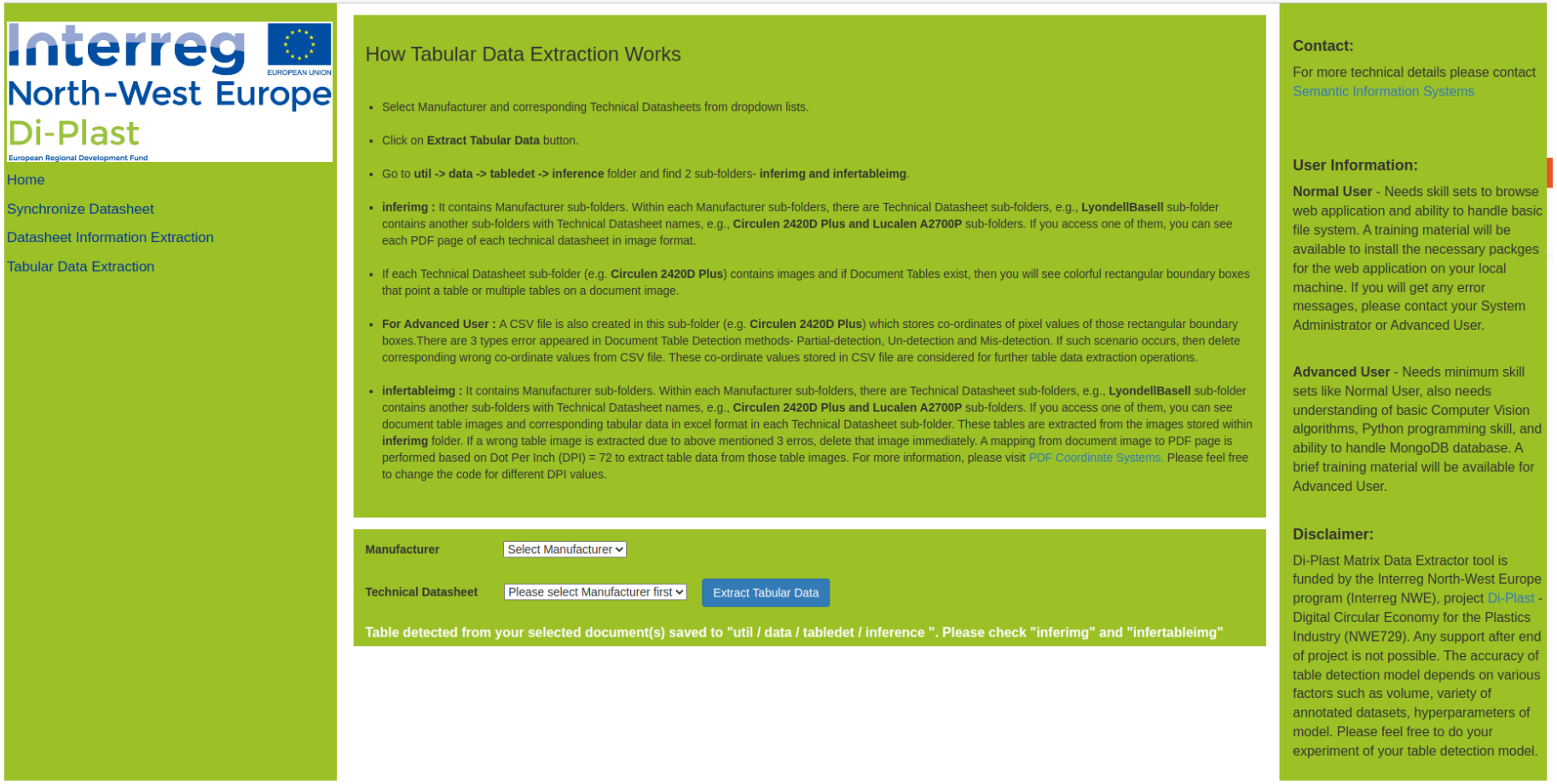
Tabular Data Extraction - II

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



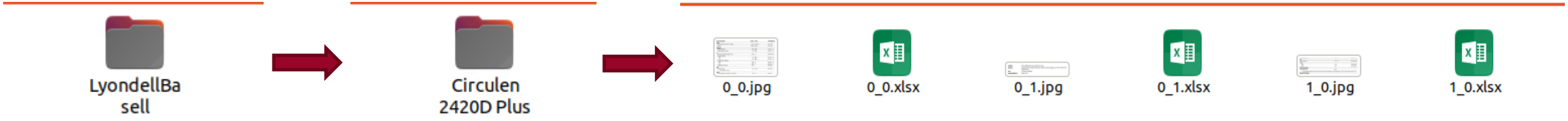
Tabular Data Extraction - III

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



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