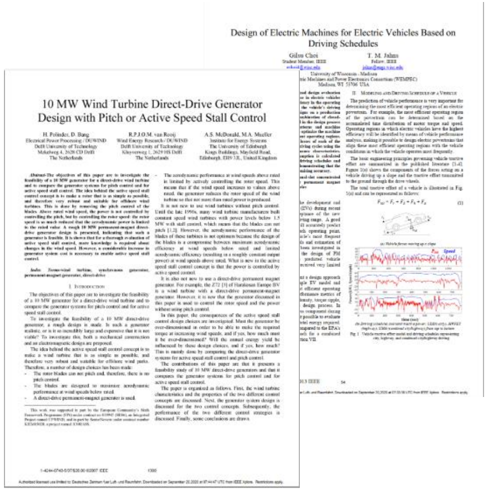
Information Extraction from PDFs

DLR Deutsches Zentrum für Luft- und Raumfahrt

Sarah Böning
Researcher
German Aerospace Center (DLR) – Institute of Data Science, Jena

Which data do I use …

- PDF files
- Currently two use cases:
  - Scientific papers describing motors (~1k PDFs)
  - Datasheets of satellite parts (~ 600 PDFs)
- Challenges with PDFs:
  - Digital but not machine-understandable
  - Combination of structured (tables) & unstructured (text) data
  - Very different layouts, even in just 1 document class



Design of Electric Machines for Electric Vehicles Based on Driving Schedules

Gilsu Choi, Student Member, IEEE

T. M. Jahns, Fellow, IEEE

University of Wisconsin - Madison

10 MW Wind Turbine Direct-Drive Generator Design with Pitch or Active Speed Stall Control

H. Polinder, D. Bang, Electrical Power Processing / DUWIND, Delft University of Technology, The Netherlands

R.P.J.O.M. van Rooij, Wind Energy Research / DUWIND, Delft University of Technology, The Netherlands

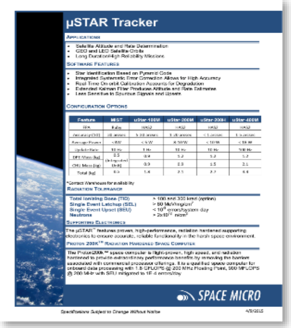
A.S. McDonald, M.A. Mueller, Institute for Energy Systems, The University of Edinburgh, United Kingdom

OVERVIEW

SPECIFICATIONS

HERITAGE

SYSTEM COMPONENTS



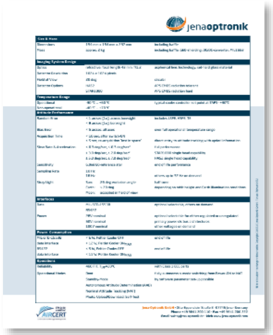
µSTAR Tracker

Applications

Software Features

Configuration Options

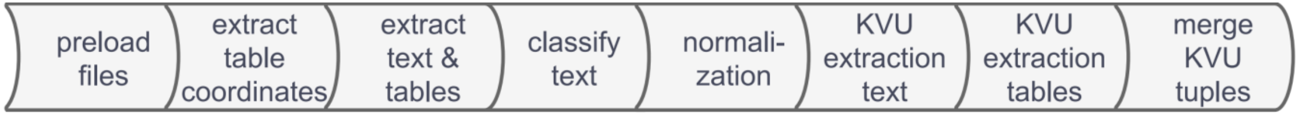
SPACE MICRO

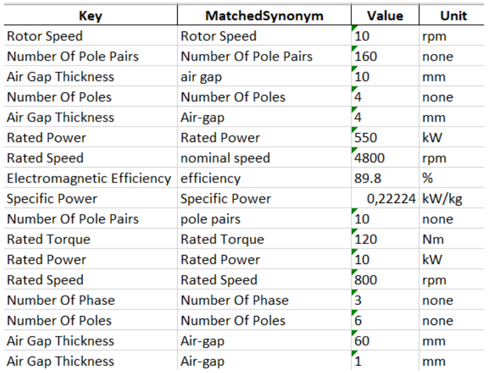


jenaoptronik

Which methods do I apply …

- Python 3[1] (incl. nltk[2], OpenCV[3], Camelot[4], pytesseract[5])
- Configurable, universial pipeline, easy to adapt
- Simple CV-based table detection + OCR for image-based tables
- Separate processing of unstructured (text) & structured (tables) data
- Extracted data post-processed to remove artifacts
- Normalization of text to simplify extraction with NLP methods
- Domain knowledge-based information extraction (e.g. via ontology)
  - 1st use case: (key, value, unit)-tuples



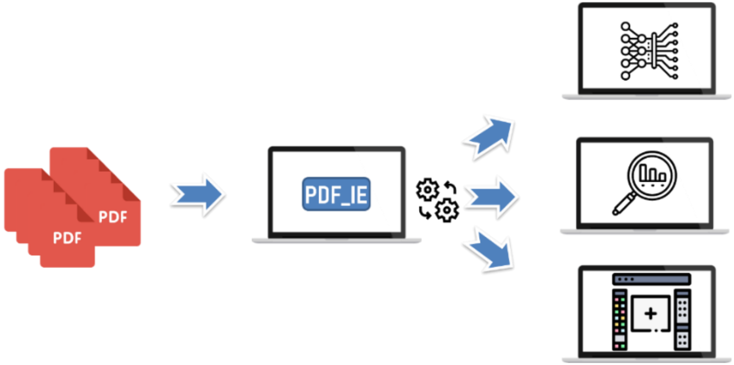


Key	MatchedSynonym	Value	Unit
Rotor Speed	Rotor Speed	10	rpm
Number Of Pole Pairs	Number Of Pole Pairs	160	none
Air Gap Thickness	air gap	10	mm
Number Of Poles	Number Of Poles	4	none
Air Gap Thickness	Air-gap	4	mm
Rated Power	Rated Power	550	kW
Rated Speed	nominal speed	4800	rpm
Electromagnetic Efficiency	efficiency	89.8	%
Specific Power	Specific Power	0,22224	kW/kg
Number Of Pole Pairs	pole pairs	10	none
Rated Torque	Rated Torque	120	Nm
Rated Power	Rated Power	10	kW
Rated Speed	Rated Speed	800	rpm
Number Of Phase	Number Of Phase	3	none
Number Of Poles	Number Of Poles	6	none
Air Gap Thickness	Air-gap	60	mm
Air Gap Thickness	Air-gap	1	mm

[1] https://www.python.org/
[2] https://www.nltk.org/
[3] https://pypi.org/project/opencv-python/
[4] https://pypi.org/project/camelot-py/
[5] https://pypi.org/project/pytesseract/
All accessed 14.10.2021

What is the goal of my work …

- Digitalization of documents
- Develop a universal method to extract any information from PDFs that can be easily adapted & specified for any domain
- Make PDFs machine-understandable & –processable
- Also comprehensible for humans
- Make information within PDFs accessible, findable, & easy-to-use for further automatic processing


PDF
PDF
PDF_IE