

Information Extraction from PDFs



Deutsches Zentrum
für Luft- und Raumfahrt

Sarah Böning

Researcher

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

The logo for the HIDA Annual Conference is a white square frame with a thick border. The text 'HIDA ANNUAL CONFERENCE' is written in a bold, white, sans-serif font inside the frame. The background of the slide is a vibrant blue with a complex, swirling pattern of concentric lines that create a sense of motion and depth.

HIDA
ANNUAL
CONFERENCE

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

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

Abstract: The objective of this paper is to develop the design of a 10 MW generator for a direct-drive wind turbine. The design is based on the use of a synchronous generator with a permanent magnet excitation system. The generator is designed to operate at a constant speed of 1000 rpm. The design is based on the use of a synchronous generator with a permanent magnet excitation system. The generator is designed to operate at a constant speed of 1000 rpm.

Fig. 1: A diagram showing the mechanical structure of the generator, including the stator and rotor.

Fig. 2: A graph showing the performance characteristics of the generator, including torque and power output.

OVERVIEW: This component is designed for high-performance applications. It features a robust construction and is capable of operating in a wide range of environments.

SPECIFICATIONS:

Parameter	Value
Power	100 W
Voltage	5 V
Current	20 A

SYSTEM COMPONENTS: The system consists of several key components, including a power supply, a control unit, and a sensor array.

µSTAR Tracker

Performance Metrics:

Parameter	Value
Power	100 W
Voltage	5 V
Current	20 A

jenaoptronic

Technical Specifications:

Parameter	Value
Power	100 W
Voltage	5 V
Current	20 A

Which methods do I apply ...

- Python 3¹ (incl. nltk², OpenCV³, Camelot⁴, pytesseract⁵)
- Configurable, universal 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

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



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

- 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

