SPECIAL PROJECT PROGRESS REPORT

All the following mandatory information needs to be provided. The length should *reflect the complexity and duration* of the project.

Reporting year	2023/2024……………………………………………………
Project Title:	ROPEWALK (Rescuing old data with people's efforts: Weather and climate archives from logbook records)
Computer Project Account:	mas@dmi.dk……………………………………………………
Principal Investigator(s):	Martin Stendel
Affiliation:	Danish Meteorological Institute
Name of ECMWF scientist(s) collaborating to the project (if applicable)	…………………………………………………………… ……………………………………………………………
Start date of the project:	2024…………………………………………………………
Expected end date:	2026…………………………………………………………

Computer resources allocated/used for the current year and the previous one
(if applicable)
Please answer for all project resources

		Previous year		Current year	
		Allocated	Used	Allocated	Used
EWCloud	(EWCU)	-	-	7,680.126	361.734
Data storage capacity	(Gbytes)	-	-	1000	?

Note that we have not been able to determine the used data storage.

Summary of project objectives (10 lines max)

On the ROPEWALK project, we aim to digitize weather observations on board of Danish ships from over the period 1680 to 1960. The original ship journals and logbooks are stored in the Danish National Archive. They fill more than 750 shelf meters with millions of observations. It would take several years to transcribe the observations by means of a crowd rescucing approach. Therefore we decided to use a machine-learning approach. To our knowledge this is the first attempt on a dataset of this size extending over so long a period. Consequently, we had to develop a scalable method to process the data.

To transcribe the millions of weather observations and make it computer-legible, a data model needs to be developed (finished recently) and tests need to be conducted (initiated in June 2024). Once these challenges have been addressed, a production run will be initiated.

Summary of problems encountered (10 lines max)

Note that we had the required access to the 40 GB GPU only from February 2024, which explains the relatively low use of resources. Over the past four months, the project has encountered a number of challenges. Developing methods and creating an image processing pipeline has been particularly difficult due to issues with determining variable table layout structures over time and recognizing handwritten text in ship logs. Image noise and the wide variation in handwriting have initially led to frequent errors, complicating the training of a robust model. Achieving a low Character Error Rate (CER) is crucial for accurate readings, which are essential to get useful data for climate models.

Additionally, using the European Weather Cloud has presented further challenges: (1) The requested GPU was not initially set up for access, and (2) training the machine learning models with sufficiently large batch sizes has been problematic due to the high GPU VRAM requirements.

Summary of plans for the continuation of the project (10 lines max)

To advance the project, additional training data will be generated, and parameter tuning on the machine learning models will be conducted. This will include retraining several models to enhance accuracy and reduce the Character Error Rate (CER) for both table layout identification and handwriting text recognition. New models will also be developed to extract text from various layout variations in the ship logs.

Tests for the image processing pipeline will be implemented to transition to a phase where weather data tables can be read robustly, efficiently, and accurately from the ship logs. Achieving the final improvements for each model and performance is anticipated to be the most challenging aspect.

List of publications/reports from the project with complete references

No publication or reports have been produced yet due to the initial phase of the project.

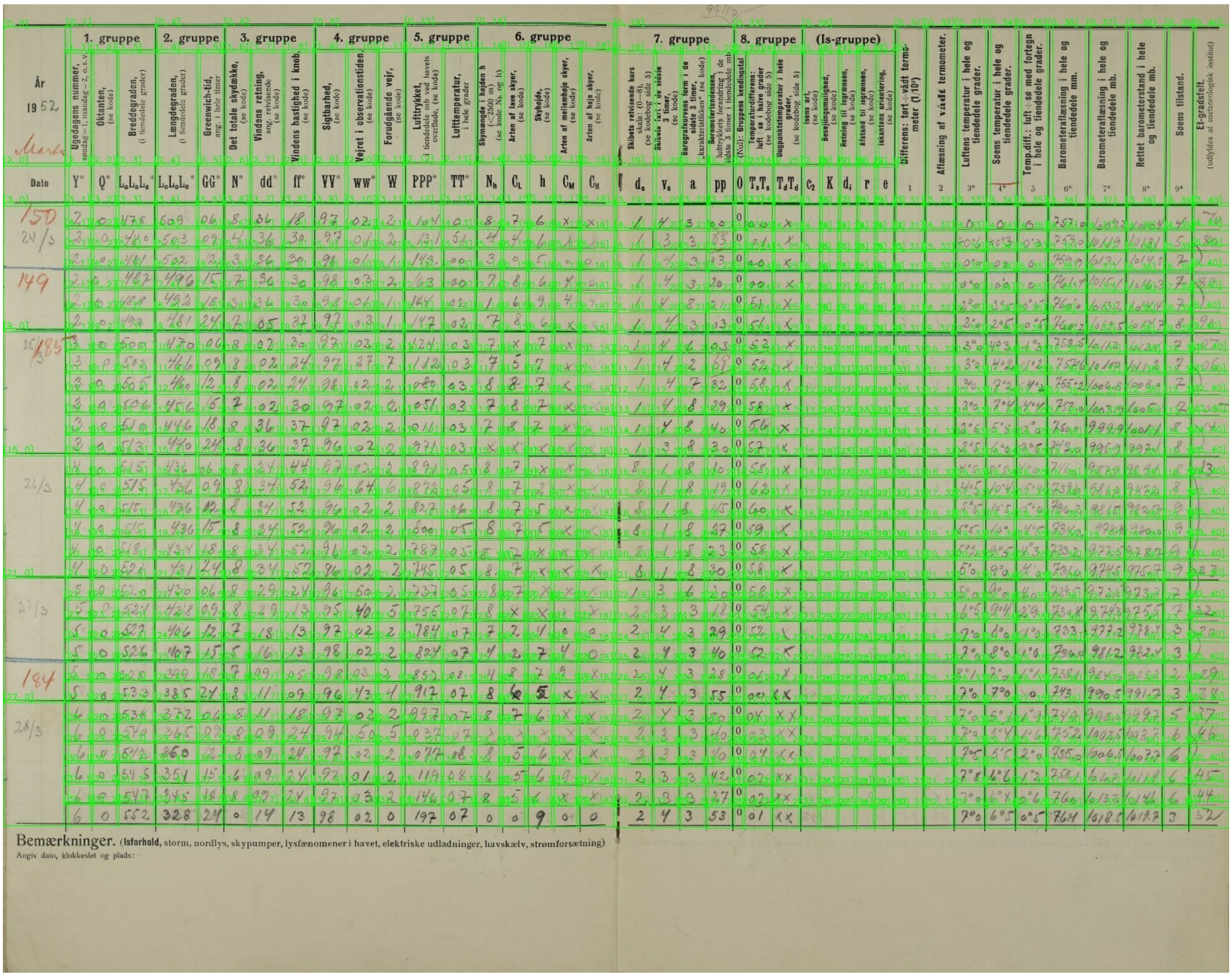
Summary of results

If submitted **during the first project year**, please summarise the results achieved during the period from the project start to June of the current year. A few paragraphs might be sufficient. If submitted **during the second project year**, this summary should be more detailed and cover the period from the project start. The length, at most 8 pages, should reflect the complexity of the project. Alternatively, it could be replaced by a short summary plus an existing scientific report on the project attached to this document. If submitted **during the third project year**, please summarise the results achieved during the period from July of the previous year to June of the current year. A few paragraphs might be sufficient.

During the first period, substantial progress was made in developing methods for the core components of the image processing pipeline. Models were trained to identify table layout structures and train handwritten text recognition to read relevant table fields. Although these methods are still being refined, continuous improvement in the machine learning models has been observed as more training data is annotated.

A brief overview of a result of identifying a table layout structure is presented below:

The example of a complete ship log image ("page") demonstrates that we are able to identify the table layout structure. The result is here highlighted with green boxes identifying table fields, along with green numbers indicating the row and column numbers:



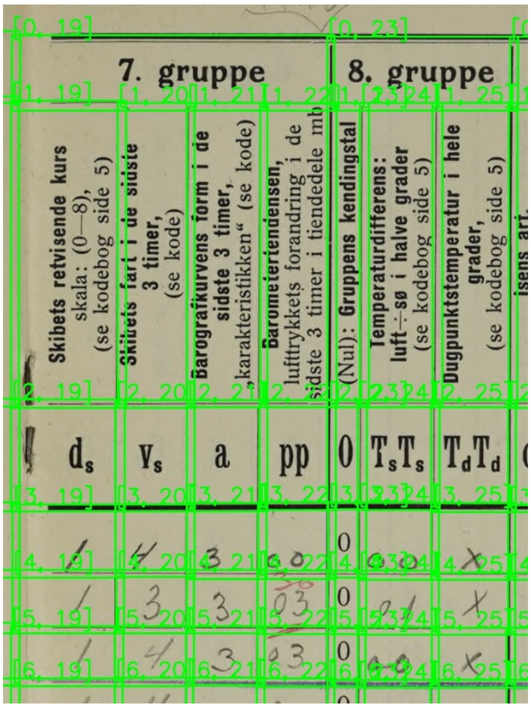
År 1952

1. gruppe 2. gruppe 3. gruppe 4. gruppe 5. gruppe 6. gruppe 7. gruppe 8. gruppe (Is-gruppe)

Dato Y* Q* LaLaLa* LoLoLo* GG* N* dd* ff* VV* ww* W PPP* TT* Nh CL h CM CH ds vs a pp 0 TsTs TdTd c2 K di r e 1 2 3* 4* 5 6* 7* 8* 9*

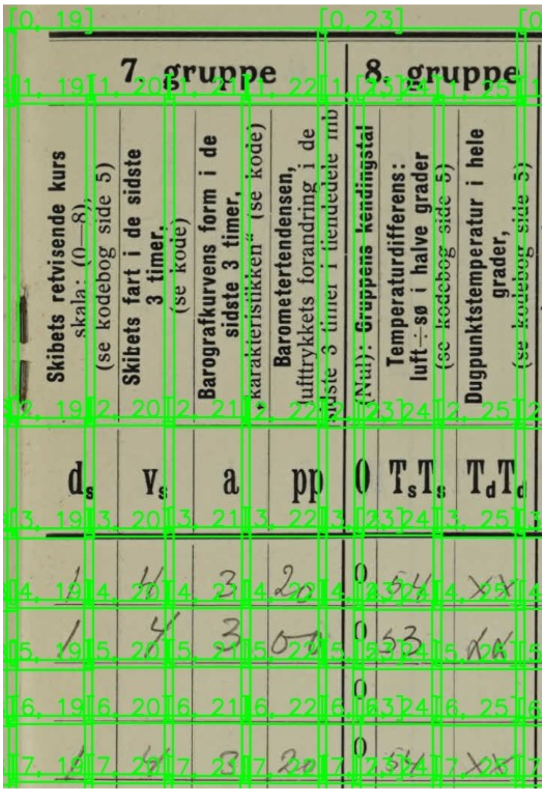
Bemærkninger. (Isforhold, storm, nordlys, skypumper, lysfænomener i havet, elektriske udladninger, havskælv, strømforsætning)
Angiv dato, klokkeslet og plads:

A closer look at the results reveals that the green boxes, with calculated overlap, align with the preprinted lines. This is important for the ROPEWALK data, since preprinted tables became available already in the early 18th century.



7. gruppe				8. gruppe			
Skibets retvisende kurs skala: (0—8), (se kodebog side 5)	Skibets fart i de sidste 3 timer, (se kode)	Barografkurvens form i de sidste 3 timer, „karakteristikken" (se kode)	Barometertendensen, lufttrykkets forandring i de sidste 3 timer i tiendedele mb	(Nul): Gruppens kendingstal	Temperaturdifferens: luft ÷ sø i halve grader (se kodebog side 5)	Dugpunktstemperatur i hele grader, (se kodebog side 5)	isens art.
d_s	v_s	a	pp	0	T_sT_s	T_dT_d	
1	4	3	00	0	00	x	
1	3	3	30 03	0	21	x	
1	4	3	03	0	00	x	

In contrast, another example illustrates a case where the table layout structure was not accurately identified, as again indicated by the green boxes. Note the shifted green lines, which can result in error-prone handwritten text recognition:



7. gruppe				8. gruppe		
Skibets retvisende kurs skala: (0—8), (se kodebog side 5)	Skibets fart i de sidste 3 timer, (se kode)	Barografkurvens form i de sidste 3 timer, „karakteristikken" (se kode)	Barometertendensen, lufttrykkets forandring i de sidste 3 timer i tiendedele mb	(Nul): Gruppens kendingstal	Temperaturdifferens: luft ÷ sø i halve grader (se kodebog side 5)	Dugpunktstemperatur i hele grader, (se kodebog side 5)
d_s	v_s	a	pp	0	T_sT_s	T_dT_d
1	4	3	20	0	54	xx
1	4	3	01	0	53	xx
				0		
1	4	3	20	0	54	xx