

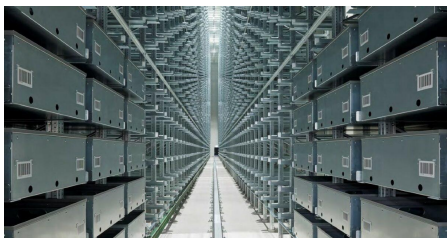
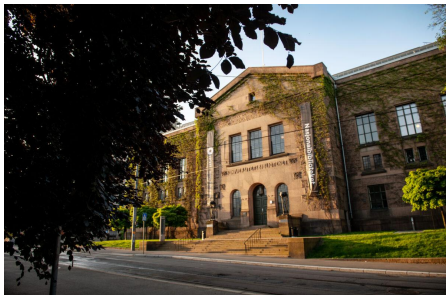
Automating Metadata Extraction and Cataloguing: Experiences from the National Libraries of Norway and Finland

Pierre Beauguitte, National Library of Norway
Osma Suominen, National Library of Finland

SWIB
26 November 2024



Legal deposit at the National Library of Norway



National Library of Norway

Digital legal deposit

Required fields are marked with *

Release information

Year of publication *

Place of publication

For example Oslo

Publication format *

The publication is mainly published/distributed in this format to subscribers/recipients

Print Digital

Does this publication have an ISBN? *

Yes No

ISBN *

Must consist of 13 digits. Hyphens are optional.

Example with hyphens: 978-82-450-0364-2

Example without hyphens: 9788245003642

Is this publication part of a series? *

Yes No

ISSN

Must consist of 8 characters. Hyphens are optional.

Example with hyphens: 0013-0818

Example without hyphens: 00130818

Series title

Nr. in series

[Back](#) [Next](#)

Last opp filer til Nasjonalbiblioteket

NB: Venligst vent på kvittering som bekrefter at filen har blitt lastet opp. Du kan bare laste opp en fil om gangen. Skal du laste opp flere filer kan du gjerne pakke dem sammen i et eksempel ZIP-formatet.



Drag and drop files
Browse your device

Before uploading, please be sure you trust this site, have the rights to the data, and want to share this content to the owner of this upload widget.

[View Box Terms of Service.](#)

[View Box Privacy Notice.](#)

box

Legal deposit at the National Library of Norway

1000s of PDF files delivered with no metadata...

Need to be catalogued manually

☰ DIMO Monografier > Ny avlevering

Ny avlevering

Type publikasjon *

Bok

Boktype (Valgfritt)

Skjønnlitteratur

Tittel*

Undertittel (valgfritt)

Originaltittel (valgfritt)

Utgivelsesår*

Ansvarsangivelser + Legg til

Du har ikke lagt til noen ansvarsangivelser i listen enda.

Utgivere * + Legg til

Du har ikke lagt til noen utgivere i listen enda.

Filer

LAST OPP FIL

Dra og slipp
eller

Velg fil

Du må laste opp en fil

Focus on grey material: public reports



Kostnader for bruk med ulikt driftsomsfang

En analyse av kostnader per dyr/dekar for bruk med ulik størrelse

NIBIO RAPPORT | VOL. 9 | NR. 41 | 2023



Håvard Rønningen Graarud og Anders Halland
Divisjon for kart og statistikk/ Avdeling for driftsøkonomisk analyse

TITTEL/TITLE Kostnader for bruk med ulikt driftsomsfang. En analyse av kostnader per dyr/dekar for bruk med ulik størrelse				
FORFATTER(E)/AUTHOR(S) Håvard Rønningen Graarud, Anders Halland				
DATO/DATE: 07.03.2023	RAPPORT NR./ REPORT NO.: 9/41/2023	TILGJENGELIGHET/AVAILABILITY: Åpen	PROSEKT NR./PROJECT NO.: 53183	SAKSNR./ARCHIVE NO.: 21/00836
ISSN: 978-82-17-03259-5		ISSN: 2464-1162	ANTALL SIDER/ NO. OF PAGES: 40	ANTALL VEDLEGG/ NO. OF APPENDICES: 1
OPPDRAUGGIVER/EMPLOYER: Landbruks og matdepartementet			KONTAKTPERSON/CONTACT PERSON: Siri Voll Dombu	
STIKKORD/KEYWORDS: Kostnader, stordriftsfordeler, strukturkostnader, skalafordeler, landbruksøkonomi, regnskapsdata, kornproduksjon, melkeproduksjon, sauedrift, frukt- og bær Economies of scale, agricultural economics, accounting data, cost structure, cereal production, dairy production			FAGOMRÅDE/FIELD OF WORK: Landbruksøkonomi Agricultural economics	
SAMMENDRAG/SUMMARY: Landbruks- og matdepartementet har bedt NIBIO belyse hvordan kostnader varierer mellom bruk med ulikt driftsomsfang og mellom bruk med samme driftsomsfang. Driftsgranskinger i jord- og skogbruk er et representativt utvalg på omtrent 920 bruk og er datagrunnlag for analysen i utredningen. NIBIO legger fram resultatene i form av figurer og tabeller, samt en kortere drøfting av disse. Metoden som er brukt, er punkt-diagrammer som viser kostnader per dekar/dyr og tabeller med gjennomsnittstall. Fullstendig sammendrag i rapporten.				
LAND/COUNTRY: Norge/ Norway				
FYLKE/COUNTY:				
KOMMUNE/MUNICIPALITY:				
STED/LOKALITET:				

Focus on grey material: public reports

Hypothesis:

using a PDF text layer (content, position, font and font size),
and a set of rules (regular expressions, keywords, heuristics),

we can automatically extract enough metadata for simple cataloguing

METEOR

<https://github.com/NationalLibraryOfNorway/meteor>

Open source - Apache License

Developed by the Text team at NLN (January 2023 -)

Written in Python with PyMuPDF

Usable as a module as well as REST API (FastAPI)


METEOR

Browse... No file selected.
or copy URL to a report: Submit
or drop a PDF file

1 of 56 70%

Kjøttfe på utmarksbeite
Beiteressursar i soner for arealltskot

NIBIO RAPPORT 1 VOL. 6 | NR 56 | 2020



YNGVE REKDAL OG MICHAEL ANGELOFF
Divisjon for kart og statistikk

Filename: NIBIO_RAPPORT_2020_6_56.pdf

Field	Value	Origin	Registry
year	2020	{'type': 'PDFINFO', 'pageNumber': 1}	
language	nno	{'type': 'LANGUAGE_MODEL'}	
title	Kjøttfe på utmarksbeite	{'type': 'FRONT_PAGE'}	
publisher	NIBIO	{'type': 'RAPPORT_PREFIX', 'pageNumber': 1}	Norsk institutt for bioøkonomi
publicationType			
author	Firstname: Yngve Lastname: Rekdal	{'type': 'INFO_PAGE', 'pageNumber': 2}	
author	Firstname: Michael Lastname: Angeloff	{'type': 'INFO_PAGE', 'pageNumber': 2}	
isbn	9788217025597	{'type': 'PAGE', 'pageNumber': 2}	
issn	2464-1162	{'type': 'PAGE', 'pageNumber': 2}	

Integration to existing pipeline

After development and test phase, *METEOR* was integrated to cataloguing tool as a suggestion provider.

- very fast response time
- *explainability*: suggestions are highlighted, and a user can see their origin.

Suggestions are saved for evaluation.

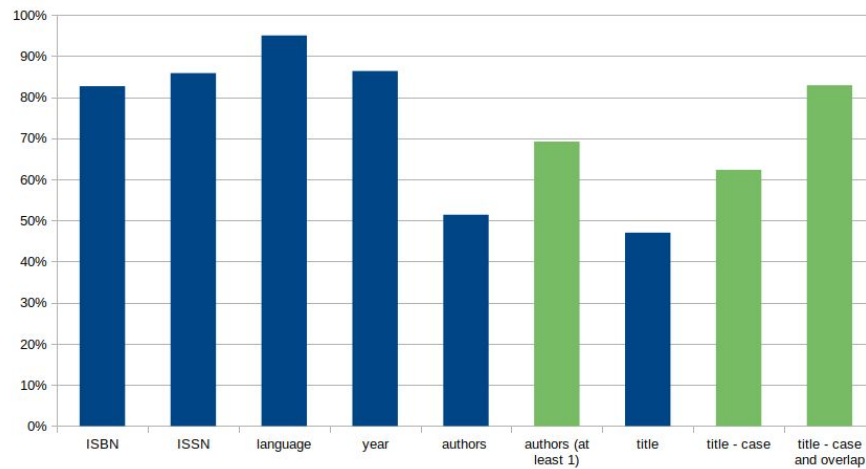
The screenshot displays a web interface for managing metadata. It features several sections:

- Ansvarsangivelser**: A table with columns for 'Rolle', 'Etternavn', and 'Fornavn'. It lists two entries: 'Forfatter Graarud Håvard Rønningen' and 'Forfatter Halland Anders'. A tooltip indicates 'Metadata autofyllt av Meteor™ fra kolofonside på side 2'. Action icons (edit, delete, star) are visible for each row.
- Utgivere ***: A field for 'Navn' containing 'Norsk institutt for bioøkonomi' with similar action icons.
- Organisasjoner**: A message box stating 'Du har ikke lagt til noen organisasjoner i listen enda.' with a '+ Legg til' link.
- Angi format for publisering ***: Radio buttons for 'Trykt' and 'Digital' (selected).
- Språk ***: A dropdown menu set to 'Norsk (bokmål)' with a star icon.

METEOR evaluation

n = 1839 reports

- ISBN: 82.7 %
- ISSN: 85.8 %
- Publication year: 86.4 %
- Language: 95 % (models from GiellaLT, The Arctic University of Norway, Tromsø)
- Authors (list of firstname, lastname): 51.4 %
 - at least one correct name in list: 69.2 %
- Title: 47 %
 - allowing casing difference: 62.3 %
 - allowing casing difference and prefix/suffix: 82.9 %



METEOR: limitations and future plans

PDFs can contain:

- text in images
- noise in text layer



- Threshold for full automation? or “What is good enough?”
- Integration with catalogues and authority registries (ex: obtain a publisher’s name from an ISSN). Higher accuracy, but loss of generality.

Metadata extraction at the National Library of Finland

- we provide repository services (DSpace installations) to many organisations
- many material types:
 - reports (by government organisations, research institutes, universities...)
 - theses (doctoral, master, bachelor, other...)
 - articles (preprints, postprints...)
 - books and book chapters
- we tested tools including GROBID and Meteor, but results were not very good
- LLM based approach seems to work much better! But needs curated data...

The image displays a collage of four different digital repository interfaces:

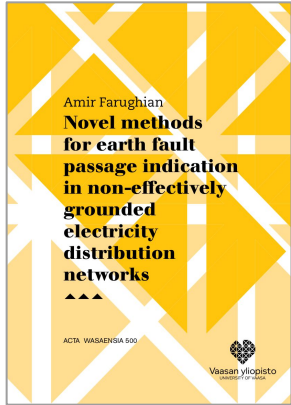
- Doria:** A dark-themed interface with a search bar and a list of collections including 'ELY-keskus [1994]' and 'Kansalliskirjasto [69735]'. It features the Doria logo and navigation icons.
- LUTPub:** A white interface with a search bar and a list of collections including 'Ammattikorkeakoulut'. It features the LUT University logo and navigation icons.
- Julkari:** A white interface with a search bar and a list of collections including 'Julkari etusivu'. It features the Julkari logo and navigation icons.
- Lauda:** A white interface with a search bar and a list of collections including 'Lauda'. It features the Lauda logo and navigation icons.

FinGreyLit data set

- PDFs and curated metadata from nine different DSpace repositories
 - academic libraries and government institutions
- ca. 800 documents (75% train, 25% test subsets)
- Dublin Core style metadata, ca. 10 fields actively curated + 20 more
- Finnish, Swedish, English language documents + some Northern Sámi

- CC0 license, available on GitHub, still work in progress
<https://github.com/NatLibFi/FinGreyLit>

Predict metadata based on document text layer



First few and last pages of text from PDF

ISBN 978-952-395-055-9 (paperback)
ISBN 978-952-395-054-2 (hardcover)
ISBN 978-952-395-055-9 (e-book)
ISBN 978-952-395-054-2 (e-book)
https://www.uva.fi/ISBN/ISBN978-952-395-055-9
Helsingin yliopisto, Helsinki, Suomi

2022

2022

2022

2022

2022

2022

Fine-tuned
language model

title: Novel methods for earth fault passage indication in non-effectively grounded electricity distribution networks

creator: Farughian, Amir

publisher: University of Vaasa

date: 2022

e-issn: 2323-9123

p-issn: 0355-2667

e-isbn: 978-952-395-055-9

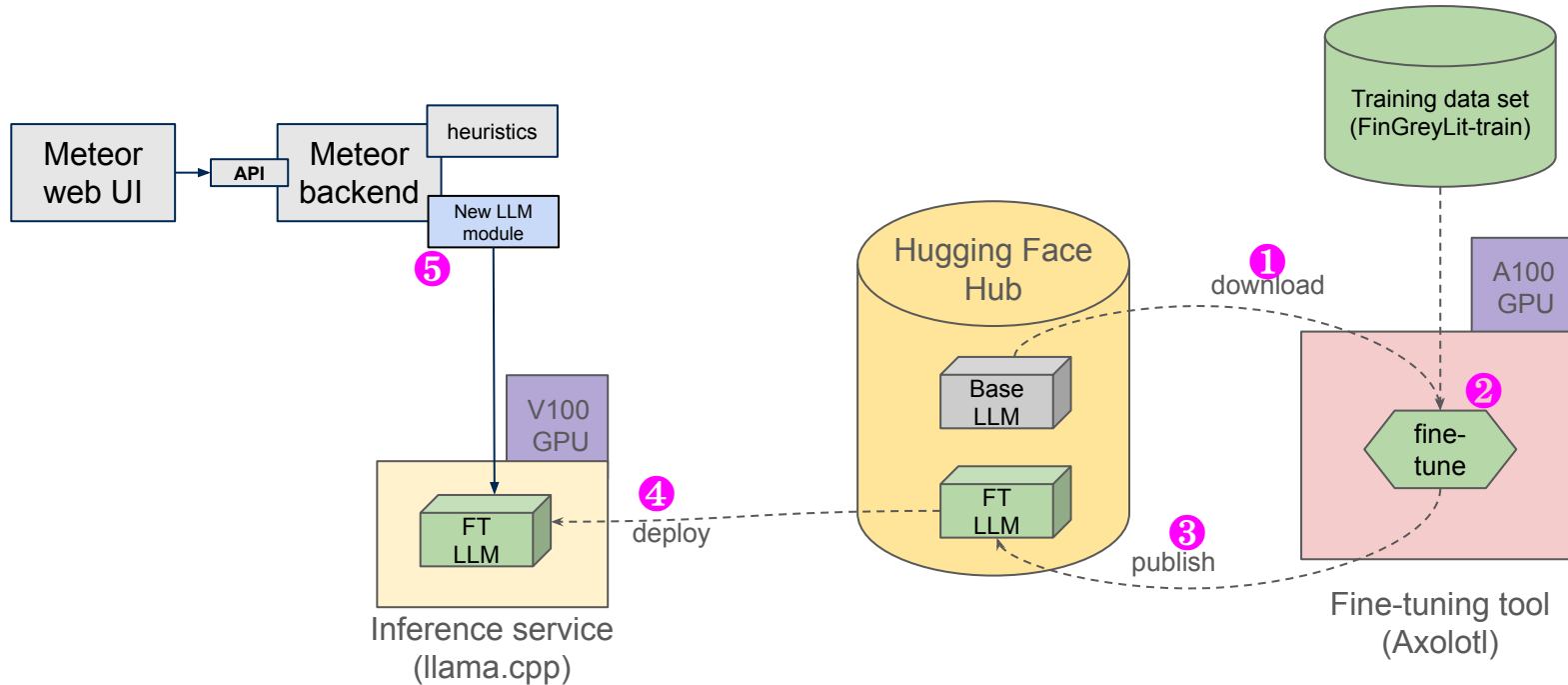
p-isbn: 978-952-395-054-2

Actually
JSON...

Fine-tuned LLMs

- testing small language models/families, e.g. **Mistral**, Llama, Gemma, StableLM, **Qwen2**...
- fine-tuned and published selected models to Hugging Face Hub:
 - [NatLibFi/Nous-Hermes-2-Mistral-7B-DPO-FinGreyLit](#)
for "production" usage
 - [NatLibFi/Qwen2-0.5B-Instruct-FinGreyLit-GGUF](#)
for local CPU-only testing and development

Fine-tuning & Meteor+LLM integration



NatLibFi is using the [University of Helsinki High Performance Computing \(HPC\) environment](#) (computing cluster with modern GPUs for UH researchers)

Meteor+LLM demo!

Method (Finder heuristics or LLM) can be chosen

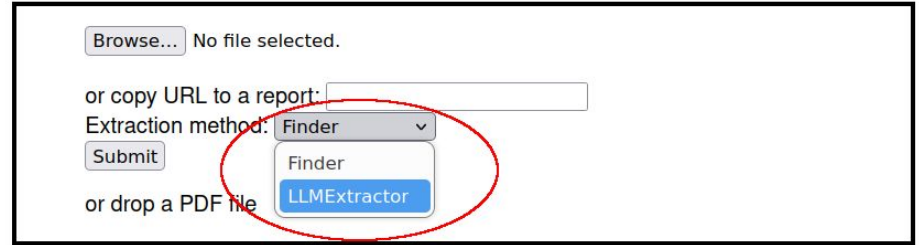
Prototype running on University of Helsinki/NLF container platform

Mistral-7B based LLM on llama.cpp, V100 GPU

Example document:

<https://www.fafo.no/images/pub/20905.pdf>

METEOR



Browse... No file selected.

or copy URL to a report:

Extraction method: **Finder** ▼

Submit

or drop a PDF file

Finder

LLMExtractor

[API-dokumentasjon](#)

[Swagger API-dokumentasjon](#)

Filename: https://www.doria.fi/bitstream/handle/10024/189909/wang_qingbo.pdf?sequence=1&isAllowed=y

Field	Value	Origin	Registry
year	2024	{'type': 'LLM'}	
language	eng	{'type': 'LLM'}	
title	Wood-Derived Functional Biomaterials towards 3D Bioprinting	{'type': 'LLM'}	
publisher	Åbo Akademi University	{'type': 'LLM'}	
publicationType			
author	Firstname: Qingbo Lastname: Wang	{'type': 'LLM'}	
isbn	9789521244032	{'type': 'LLM'}	
issn			

LLMs can follow (some) cataloguing conventions

Tiina Kähkönen

On the title page:

EMPLOYEE TRUST REPAIR IN THE CONTEXT OF
ORGANIZATIONAL CHANGE – IDENTIFICATION AND
MEASUREMENT OF ACTIVE TRUST REPAIR PRACTICES

Extracted as:

author

Firstname: Tiina
Lastname: Kähkönen

LLM can separate
first and last names

title

Employee trust repair in the context of organizational
change : identification and measurement of active trust
repair practices

LLM can reformat
titles & fix punctuation

Evaluation: How well does metadata extraction work?

Train set	-	-	FGL-train	FGL-train	FGL-train + NR
Eval set	NR	FGL-test	FGL-test	FGL-test	FGL-test
	Meteor	Meteor	Qwen2 0.5B	Mistral Nemo 12B	Mistral Nemo 12B
language	98%	97%	98%	99%	100%
title	42%	40%	73%	90%	91%
alt_title	-	-	73%	80%	83%
creator	67%	65%	72%	89%	91%
publisher	43%	8%	69%	85%	86%
year	85%	72%	87%	91%	90%
e-isbn	90%	83%	90%	94%	95%
e-issn	92%	86%	95%	93%	95%
p-isbn	-	-	89%	92%	92%
p-issn	-	-	92%	95%	95%
doi	-	-	97%	98%	98%
type_coar	-	-	85%	88%	90%
	76%	63%	85%	91%	92%

NR: NorRep data set (130 reports from Norwegian public institutions)

FGL: FinGreyLit data set (train: 619 docs; test: 179 docs)

Meteor: original Finder heuristics of Meteor evaluated on either **NorRep** or **FinGreyLit**
→ results for NorRep much better than for FGL

Qwen2 0.5B: small LM, 500 million parameters, CPU
→ good for testing and development on laptops

Mistral Nemo 12B: larger LM, 12B parameters, GPU
→ current best **FGL** results
→ adding **NR to train set** increases quality a little!

Lessons learned, so far

1. Multilingual models (Mistral, StableLM...) are surprisingly good even for Finnish and Swedish languages they don't officially support.
2. The fine-tuning approach allows data-driven expansion: extracting one more piece of information (e.g. DOI) just requires collecting and curating training data, but very little extra code.
3. It's possible to extract metadata using small-ish LMs (~10B) with reasonable speed (3-5 seconds / document) on locally hosted infrastructure.
4. Ecosystem of LLMs, fine-tunes, tools etc. moving extremely fast; today's state of the art may be obsolete next month!

Future work and collaboration

1. Expanding the data set used for training; pooling our data sets; fine-tuning and evaluating newer, more capable LLMs
2. Indicating the origin (e.g. page number) of found information in the LLM approach requires more detailed ground truth / training data
3. Working towards a stable enough LLM to be used in production settings
4. Refining evaluation metrics: what is a "correct enough" answer for real use cases?
5. Better extraction of information from PDFs, including text in images
6. Resolving names (person, organization...) to identifiers
 - a. Meteor supports a simple SQL authority database with name to ID mapping
 - b. Using the W3C Reconciliation API to find matching identifiers?

Thank you!

pierre.beauguitte@nb.no

osma.suominen@helsinki.fi