NATIONAL LIBRARY OF FINLAND

Extracting metadata from grey literature using large language models

Osma Suominen

Tekoälykahvit 1.11.2023

Perustuu SWIB23-konferenssin salamaesitykseen 12.9.2023

Grey literature?

reports

working papers

government documents

white papers

preprints

theses

...

semi-formal non-commercial
PDFs published on the web
– lots of them!

Image made with DreamStudio
(based on Stable Diffusion)
Prompt: "A big pile of papers, reports, documents, PDF files, word documents, powerpoint slides, posters, articles and books forming a wave in the style of Hokusai"

Why extract metadata from grey literature PDFs?

1. help users of **digital repositories** who need to enter metadata when uploading PDFs

2. in **web archiving**, get more information about harvested PDF files

3. ease **cataloguing** of e.g. government reports and doctoral theses

Why is it **hard** to extract metadata from PDFs?

1. You can't rely on embedded PDF metadata (pdfinfo)

```
Title: Microsoft Word - Kurkela&RantanenMostFinal3.docx
Title: A9R122vgnv_1n6ctkf_ezs.tmp.pdf
```

- so you (also) have to look at document text

2. Diverse templates and creative text **LAY***OUTS*

3. PDF text extraction is messyand unreli-\nable

Early **experiments**

- 2019-2020: We tried GROBID. Didn't work with grey literature (intended for scholarly papers but not other document types)

- 2022: Lingsoft Inc. made a prototype for us (PDF text extraction and traditional machine learning)

- 2023: Experiment using OpenAI GPT-3 fine-tuning API https://github.com/osma/llm-metadata-extraction/

Predict metadata based on document text

Amir Farughian

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

ACTA WASAENSIA 500

Vaasan yliopisto
UNIVERSITY OF VAASA

ISBN	978-952-395-054-2 (print) 978-952-395-055-9 (online)
ISSN	0355-2667 (Acta Wasaensia 500, print) 2323-9123 (Acta Wasaensia 500, online)
URN	https://urn.fi/URN:ISBN:978-952-395-055-9

Hansaprint Oy, Turenki, 2022.

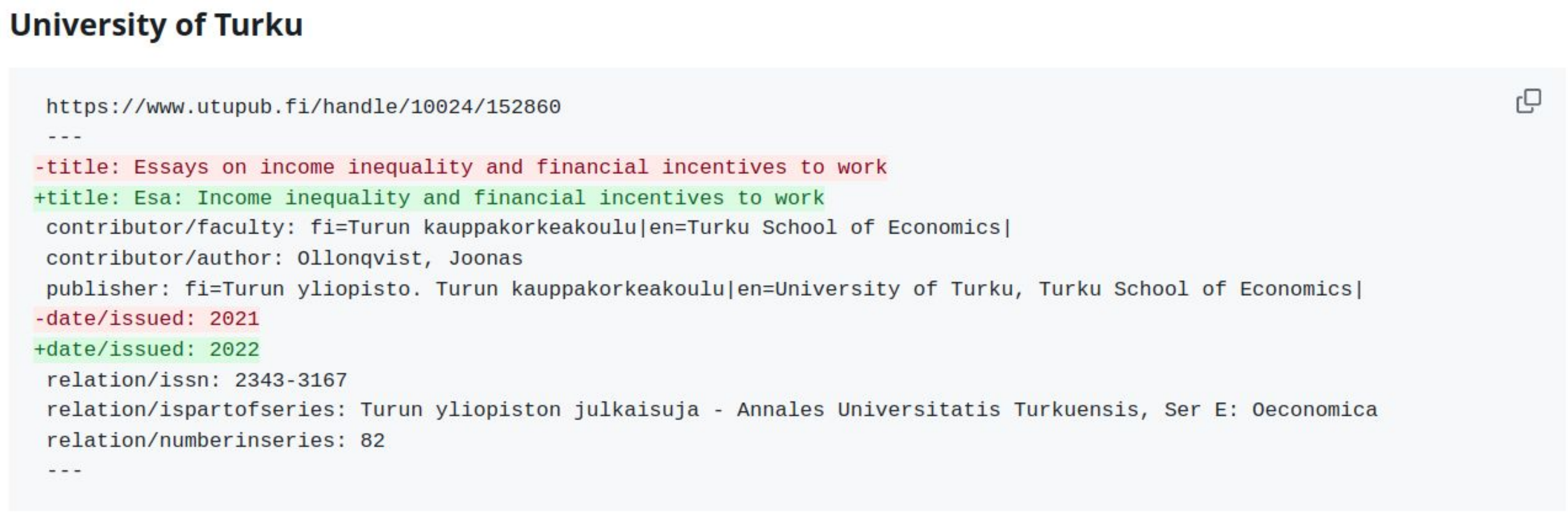
First 5 pages of text from PDF

Fine-tuned GPT language model

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

contributor/faculty: fi=School of
Technology and nnovations|en=School
of Technology and Innovations|

contributor/author: Farughian, Amir

contributor/organization: fi=Vaasan
yliopisto|en=University of Vaasa|

publisher: Vaasan yliopisto

date/issued: 2022

relation/issn: 2323-9123

relation/issn: 0355-2667

relation/isbn: 978-952-395-054-2

relation/ispartofseries: Acta
Wasaensia

relation/numberinseries: 500
```

Example of LLM extracted metadata

Diff view: human vs. LLM generated



University of Turku

```
 https://www.utupub.fi/handle/10024/152860
 ---
-title: Essays on income inequality and financial incentives to work
+title: Esa: Income inequality and financial incentives to work
 contributor/faculty: fi=Turun kauppakorkeakoulu|en=Turku School of Economics|
 contributor/author: Ollonqvist, Joonas
 publisher: fi=Turun yliopisto. Turun kauppakorkeakoulu|en=University of Turku, Turku School of Economics|
-date/issued: 2021
+date/issued: 2022
 relation/issn: 2343-3167
 relation/ispartofseries: Turun yliopiston julkaisuja - Annales Universitatis Turkuensis, Ser E: Oeconomica
 relation/numberinseries: 82
 ---
```

What we **found out**

This LLM extraction thing could actually work!

...but we need more consistent "ground truth" metadata.

– and we probably don't want to rely only on OpenAI for this

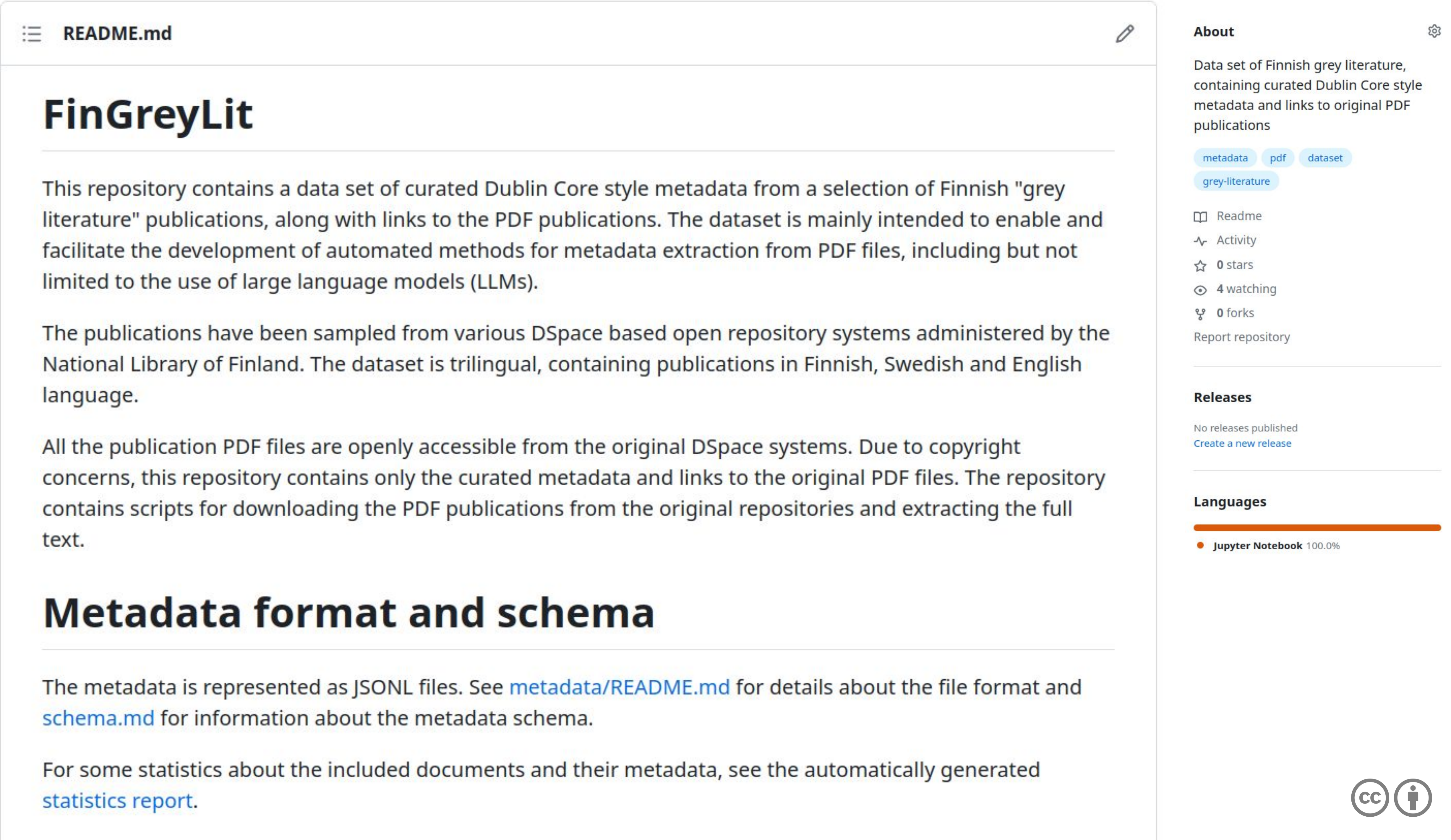
(proprietary LLM, vendor lock-in, copyright and privacy concerns, ethics...)

FinGreyLit data set

- PDFs and metadata collected from nine different DSpace repositories
  - academic libraries and government institutions

- ca. 700 documents curated by a summer intern & refined by us
- Dublin Core style metadata, ca. 30 fields in use
- Finnish, Swedish and English language documents
- JSON Lines file format for ease of use in experiments
- CC0 license

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



README.md

FinGreyLit

This repository contains a data set of curated Dublin Core style metadata from a selection of Finnish "grey literature" publications, along with links to the PDF publications. The dataset is mainly intended to enable and facilitate the development of automated methods for metadata extraction from PDF files, including but not limited to the use of large language models (LLMs).

The publications have been sampled from various DSpace based open repository systems administered by the National Library of Finland. The dataset is trilingual, containing publications in Finnish, Swedish and English language.

All the publication PDF files are openly accessible from the original DSpace systems. Due to copyright concerns, this repository contains only the curated metadata and links to the original PDF files. The repository contains scripts for downloading the PDF publications from the original repositories and extracting the full text.

Metadata format and schema

The metadata is represented as JSONL files. See metadata/README.md for details about the file format and schema.md for information about the metadata schema.

For some statistics about the included documents and their metadata, see the automatically generated statistics report.

About

Data set of Finnish grey literature, containing curated Dublin Core style metadata and links to original PDF publications

metadata pdf dataset grey-literature

Readme
Activity
0 stars
4 watching
0 forks
Report repository

Releases

No releases published
Create a new release

Languages

● Jupyter Notebook 100.0%

WIP: **New experiments** using FinGreyLit data set

- using the **Meteor** metadata extraction tool
    - just published as Open Source by the National Library of Norway
    - https://github.com/NationalLibraryOfNorway/meteor
    - extracts title, authors, language, publisher, year, ISBN, ISSN

- again using the **GPT-3 fine-tuning API**, but with better metadata

- fine-tuning a **LLaMA 2** language model (semi-open LLM by Meta) locally on a GPU using the resource-efficient QLoRA technique

- TODO: fine-tuning **Finnish GPT-3** models published by TurkuNLP group

Results

It's looking pretty good so far.

The **data set must still be improved**.

No results to publish yet, stay tuned!

1640
NATIONAL LIBRARY
OF FINLAND

Thank you!

Osma Suominen

osma.suominen@helsinki.fi
@osma@sigmoid.social