# **Project Plan: Full Text Screening - AI model**

Timeline in the method project:

Training sets 2 & 3 ready in December

Full abstract screening done around January?

Timeline of AI full text screening

What is the desired schedule, just so I know what to prioritize as I have another project as well.

The currently relevant topics are highlighted with orange.

## **Scope**

The goal is to assist screeners with inclusion/exclusion decisions based on full data.   
The project can be divided to two sub tasks:

1. Tagging the data
2. Use tagged data to exclude/include, automatically or with human in the loop.

Tagging can mean either word counts from pre-defined dictionary, or question answering/data answering with NLP. I will first focus on the possibilities with NLP.

**Input:**  Full pdf

**Input format for the model**: Pdf file with text. Image like pdfs are not considered yet.

How big proportion of the papers are image like pdfs where some off-shelf OCR should be applied? This can cause additional errors in the predictions. And if we take text like input, this means that Figures and Tables are not included.

**Desired output:**   
Tagging information with NLP  
PICO:

* Subject: [human, animal, in-vivo]
  + If animal, what species
  + If in-vivo, what cells
* Intervention / Exposure:
  + Radiation: [gamma, x, alpha,…]
  + Other interventions?
* Comparison
* Outcome

Decision Tree:

* Is literature review: y/n
* Includes ionizing radiation: y/n
* Only UV: y/n
* …

One possible approach is also to use some separate classifier to detect human/animal and in-vivo studies, and based on that apply the corresponding model for tagging.

**Output format:** JSON? .txt? .csv? and the structure of these outputs?

## **Data collection**

(abstract screening done manually)

**SD:**157 screened papers  
88 excluded papers  
Questions: (The numbers notate the number of papers that were annotated with this flag at least by one screener.)

* 'This article concerns '
  + Human (100)
  + Cells (36)
  + Animal (45)
  + More than one study type (5)
* 'Are you able to screen this article?'
  + Yes
  + No, there is something wrong with the pdf link
  + No, for another reason
  + No, the article is written in a language that I cannot read (5)
* 'Are you including or excluding this study?'
  + Include (92)
  + Exclude (88)
* 'Is the data reported separately by sex? '
  + FALSE
  + TRUE
* 'If excluding, which exclusion criteria applies to this study?'
  + '1. Study Design. Not a controlled study.' (43)
  + '5. Outcome. ' (8)
  + '4. Population. Age/Gender/Disease status etc. not matching our PICOS'(30)
  + '2. Radiation. No ionising radiation in intervention group.'(6)
  + '3. Co-intervention. Accompanying exposure (surgery accepted).' (12)
* 'Please give details'
  + 'Human and Animal', (This is sometimes annotated together with “Human” label)
  + 'cells and animal',
  + 'in vitro, in vivo and ex vivo studies',
  + 'Animal (frogs) and in vitro experiments both described. Only in vitro study is valid for data extraction.'
* 'Please provide details'
  + 'No PDF available'

**CNS:**Proposed question structure:

1. Does this article describe primary study? (reviews?)
2. 'This article concerns '
   * Human
   * Cells
   * Animal
   * More than one study type (can we make it multichoice?)
3. Does this article consider intervention or exposure with radiation?
4. Are the subjects not exposed to any other chemical … ? Are all the subjects exposed to other chemical, biological or physical agents then radiation? We do not consider surgeries as such agents.
5. Is there evidence for absence of a control group?
6. Is it only about neoplastic or cancer related effects?
7. 'This article concerns '
   * Human
   * Cells
   * Animal
   * More than one study type (can we make it multichoice?)

* Is the study literature review?
  + y/n
* Is there evidence that the study does not include ionising radiation health/biological effets?
  + y/n
* Is it only about UV radiation? / What type of radiation is considered?
  + ??
* Is the study about the effects on the CNS? / What effects does the study address? Is one of the outcomes investigated in the CNS?
  + y/n
* Is it only about neoplastic or cancer related effects?
  + y/n
* Is it only about anxiety levels/emotional distress?
  + y/n
* Is there evidence for absence of control group?
  + y/n
* Are all the subjects exposed to other chemical, biological or physical agents then radiation? We do not consider surgeries as such agents.
  + y/n

If human study

* How old are the study subjects?
* How many subjects are there?
* Is the study conducted on unborn/pregnant/deceased subjects?

If animal study

1. Which animals are considered?
2. Are the animals adults?
3. Are they healthy animals?
4. Is it comparing exposure to non exposure?
5. IS there a single group pre- vs. post-intervention?

If in-vitro

1. Which cells are considered?
2. Are the the cells from healhy subjects?
3. Are the cells from healthy animals?
4. Is it comparing exposure to non exposure?
5. IS there a single group pre- vs. post-intervention?

How were the pdfs extracted from Zotero? R Code provided by Edinburgh, run where and when?

## **DaTA Preprocessing**

#### Read paper

A close-up of a paper

Description automatically generated

If Image based pdf:

Use some OCR tool and proceed similarly.

A close-up of a paper

Description automatically generatedIf regular pdf with text:

1. Use pdfMinerSix to find bounding boxes of chapters   
   ([Welcome to pdfminer.six’s documentation! — pdfminer.six \_\_VERSION\_\_ documentation (pdfminersix.readthedocs.io)](https://pdfminersix.readthedocs.io/en/latest/))
2. Read the text using pyPDF2.
3. Find the most used formatting (font and size) in the document.
4. Only include the chapters where the most common formatting is used, and there is minimum of 5 letters (to exclude tables with only values)

* GROBIT tool could be used as well: [kermitt2/grobid: A machine learning software for extracting information from scholarly documents (github.com)](https://github.com/kermitt2/grobid)

#### Embed the paper

Embed all the included chapters and store the embeddings for later use.  
ChatGPT: $0.0001 / 1K tokens

#### Query

1. Select an appropriate prompt template
2. Find the most relevant chapters with K-nearest algorithm on the embedding space
3. Give these chapters as the context for the query. (Or make a summary based on the data)
4. Use e.g. ChatGPT API: $0.0030 / 1K tokens for the input

#### Sanity checking etc?

What to do if the model is unsure? Flag?

## **Possible methods**

Pre-Pico   
Abstract text 🡪 PICO labels  
 [Species, Intervention, Comparator, Strain, Induction, Outcome]  
([hemmlin/pre-pico-test (github.com)](https://github.com/hemmlin/pre-pico-test))  
From my initial exploration on the code it seems like data extraction from abstract  
Waiting for further Documentation  
BIRD based solutions might not be the best approach for long context: [2203.11258.pdf (arxiv.org)](https://arxiv.org/pdf/2203.11258.pdf)

Donut 🍩 Image 🡪 question answering, or data extraction to json[arxiv.org/pdf/2111.15664.pdf](https://arxiv.org/pdf/2111.15664.pdf)Nice open source pre trained model with easy api through HuggingFaceAlready working somehowCould be limited in the size of context?Similar approach to Microsoft DUBLIN [2305.14218.pdf (arxiv.org)](https://arxiv.org/pdf/2305.14218.pdf)

General purpose VDUse.g. LayoutLM, LayoutLMv2, BERT  
OCR 🡪 Transformer 🡪 Encoder

Spade  
[Spatial Dependency Parsing for Semi-Structured Document Information Extraction (aclanthology.org)](https://aclanthology.org/2021.findings-acl.28.pdf)similar to WYVERN  
graph based method that utilizes the OCR bounding boxes to obtain better performance  
Maybe not so useful with articles where the hierarchy of the text is quite simple.

PromptingText context 🡪 question answering.  
[lamini-ai/llm-classifier: Classify data instantly using an LLM (github.com)](https://github.com/lamini-ai/llm-classifier)  
- both zero- and few-shot methods available  
- requires docker 🡪 not possible to test in my work laptop.[2303.05352.pdf (arxiv.org)](https://arxiv.org/pdf/2303.05352.pdf)[2305.19835.pdf (arxiv.org)](https://arxiv.org/pdf/2305.19835.pdf)Would require paid version of ChatGPT API, but probably easy to implement and can be used for highlighting. Fully zero-shot method.Or using Llama or other open-source option.ChatGPT API pricing: [Pricing (openai.com)](https://openai.com/pricing)  
0.003$/1K tokens  
🡪 approx. 0.03$/paper + answer tokens.

Classifiers for long context  
[applsci-12-04554-v2.pdf](file:///C:\Users\linda.hemmann\Downloads\applsci-12-04554-v2.pdf)  
[2203.11258.pdf (arxiv.org)](https://arxiv.org/pdf/2203.11258.pdf)  
- bench mark  
- truncated BERT is quite good 🡪 is it wise to do this with AI?  
- should we just reuse the abstract classifier with random or ranked sentences from the paper?  
[Building a classification service with Llama2 in Python (mikulskibartosz.name)](https://mikulskibartosz.name/building-classification-service-with-llama2-in-python)  
[sh0416/llama-classification: Text classification with Foundation Language Model LLaMA (github.com)](https://github.com/sh0416/llama-classification/tree/master)  
-Llama2 based classifiers seem to be quite popular  
- unfortunately, the maximum context size is **2048 tokens**

### Other papers:

[Screening for in vitro systematic reviews: a comparison of screening methods and training of a machine learning classifier | Clinical Science | Portland Press](https://portlandpress.com/clinsci/article/137/2/181/232436/Screening-for-in-vitro-systematic-reviews-a)

**LLM for interpreting tables:**[2305.14336.pdf (arxiv.org)](https://arxiv.org/pdf/2305.14336.pdf)  
[2001.01469.pdf (arxiv.org)](https://arxiv.org/pdf/2001.01469.pdf)

**OCR libraries:**Pytesseract: [pytesseract · PyPI](https://pypi.org/project/pytesseract/)

easyOCR: [JaidedAI/EasyOCR: Ready-to-use OCR with 80+ supported languages and all popular writing scripts including Latin, Chinese, Arabic, Devanagari, Cyrillic and etc. (github.com)](https://github.com/JaidedAI/EasyOCR)

PyPDF: [pypdf · PyPI](https://pypi.org/project/pypdf/)

LangChain: [Introduction | 🦜️🔗 Langchain](https://python.langchain.com/docs/get_started/introduction.html)  
- works both with image and text pdfs

trOCR: [TrOCR (huggingface.co)](https://huggingface.co/docs/transformers/model_doc/trocr)

docTR: [docTR documentation (mindee.github.io)](https://mindee.github.io/doctr/)  
- Gives the spatial information and the text, can be used to recreate the document for highlighting etc.

PaddleOCR: [paddleocr · PyPI](https://pypi.org/project/paddleocr/)  
- gives bounding box, for highlighting  
- English and Chinese

Clova OCR   
- commercial

## **Results**

Clearly define the desired evaluation statistics and benchmarking metrics based on the project's objectives. What will be considered a successful outcome for this project?

Without any tuning the human/animal/cell classification performs as:

Confusion Matrix:

Animal Cells Human Unsure

Animal 10 1 0 2

Cells 2 2 4 1

Human 2 1 14 8

Unsure 0 0 0 0

Animal Cells Human Unsure

Animal 5 0 0 2

Cells 2 0 2 0

Human 2 1 9 6

Unsure 0 0 0 0

Animal Cells Human Unsure

Animal 18 3 1 4

Cells 3 11 2 0

Human 1 7 20 27

If the classification is separated to two phases, 11 miss annotated papers in the data (10%).

## **Documentation and Resources**

Github Project: [ESA-RadLab/FullTextTagging (github.com)](https://github.com/ESA-RadLab/FullTextTagging)  
Front end?

## **Paper outline**:

Format?

How to differentiate from paper-qa and other similar if taking more machine learning oriented approach? ([whitead/paper-qa: LLM Chain for answering questions from documents with citations (github.com)](https://github.com/whitead/paper-qa))

## **Glossary**

VDU – Visual Document Understanding  
OCR – Optical Character Recognition  
IE – Information Extraction  
DocVQA – Document Visual Question Answering

## **Other notes**

The performance of general-purpose VDUs relies on the performance of OCR (Donut)

A graph of different colored bars

Description automatically generated with medium confidence

# Questions

* What to do “papers” like *../data/test\_SD\_files//2017\_Abstracts.pdf* and what does the screening annotations for such summaries even mean?
* How to handle studies with several classes? Can you choose several options in SyRF?
* Why are not all the “relativePfdPaths” not found in the Zotero?
* Where is the database of all the pdf files?