AI and Automation in Metadata: Experiments and Future Directions

17 Sep 2025

Abigail Huang (Principal Librarian/Resource Discovery)

Jeremy Goh (Librarian/Resource Discovery)

AGENDA

Case Study 1: The future of WAS Cataloguing

Background
Title and Abstract Generation
Subject Assignment
Quality assessment
Summary

Case Study 2: SGCAT – A cataloguing assistant prototype

Overview
What is SGCAT?
Benefits
What’s Next

Discussion

Case Study 1: The future of WAS Cataloguing

Using AI to create metadata for archived websites in the Web Archives Singapore

Read the published paper here:

https://go.gov.sg/was-paper-2025

Background

- Archived collection of .sg websites documenting Singaporean life, culture, and history in the 21st century

- Covering individuals, associations, businesses, cultural organizations, events, news

Inclusion Criteria

- **.sg websites** and selected **non .sg websites** with Singaporean ownership or institutions

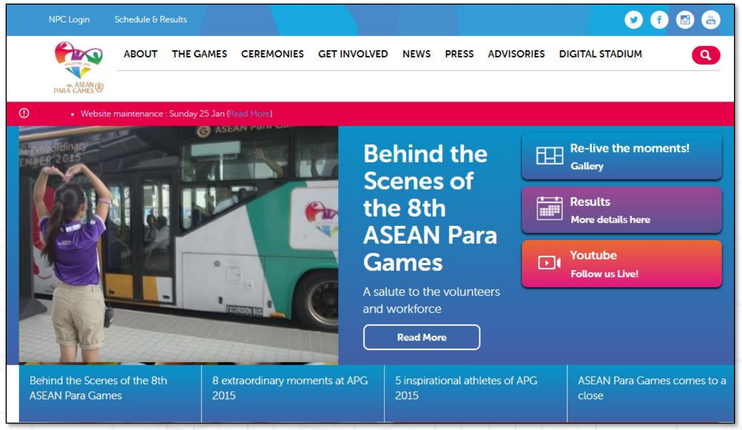
WAS WEB ARCHIVE SINGAPORE

320,543
Website versions

98,708
Unique websites

Highlights of websites archived

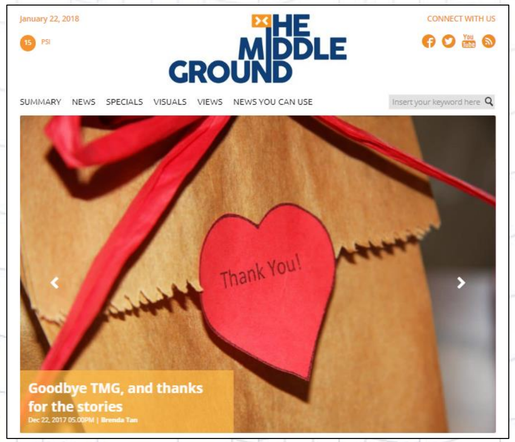
National Day Parade 2015



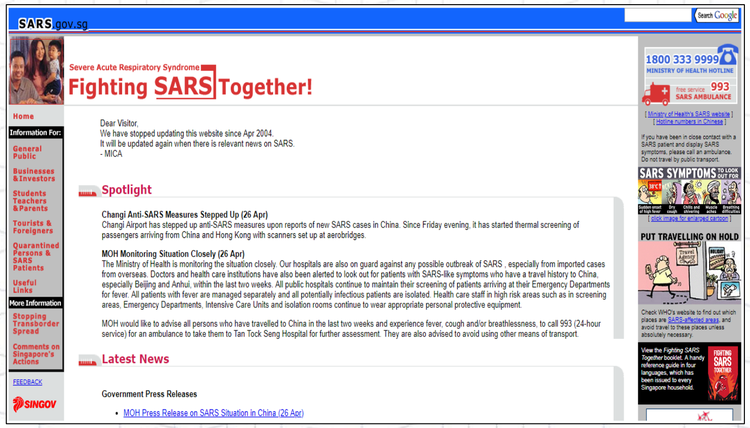
2015 ASEAN Para Games

Mothership article on Trump-Kim Summit

First Toa Payoh Secondary School



The Middle Ground



Fighting SARS Together

Why is the Web Archive Singapore Collection a Good Candidate for AI/ML Exploration?

- **It addresses a need:**
  - Number of WAS records increased significantly from FY20, due to annual whole domain crawl, while resources remain limited
- **Advantages:**
  - Content is born digital – no need to spend money digitising
  - Substantial amount of past data available (90K websites)
  - Many new websites similar to previous examples (eg condominium launch websites, employment agencies)
  - Automated QC removes low-quality websites from crawl
- **But...**
  - Extremely large size of website files
  - Data still located in an internal government network
  - Limited text in some websites (many graphics)
  - Variety of languages, though >90% are in English

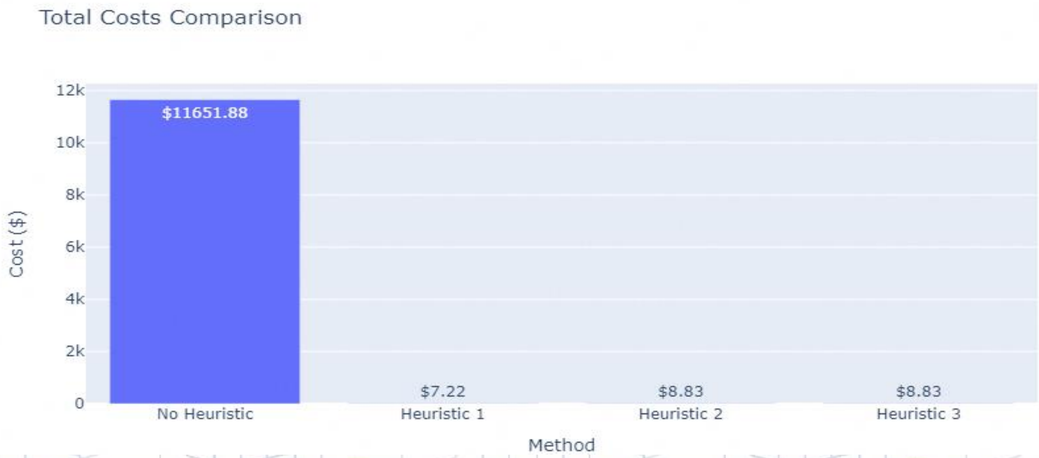
Title and Abstract Generation

Process WARC files & Reduce Data Size

- **Parsed 112 WARC (Web ARChive) files to extract web content**
  - Processes HTML responses, extracting titles and text content
  - Implements quality assurance checks to filter out low-quality content
  - Deduplicates records based on URL normalization
  - Uses multithreading for efficient processing of multiple files
  - Outputs results as a pandas DataFrame for further analysis

Filename	Title	Text
toa_payoh_sec.warc	First Toa Payoh Secondary School	Thank you Toa Payoh ...

- **Cost of sending WARC files is not cost effective (>$11K for the tokens required)**
  - As cataloguers primarily use “main” and “about” pages for cataloguing, develop similar data optimisation heuristics (H)
    1. “About“ page or the shortest URL by length + website's URL.
    2. Shortest URL by length + website's URL
    3. Above + Regular expression rules
  - Identifies most relevant content and reduces token count (<$9)



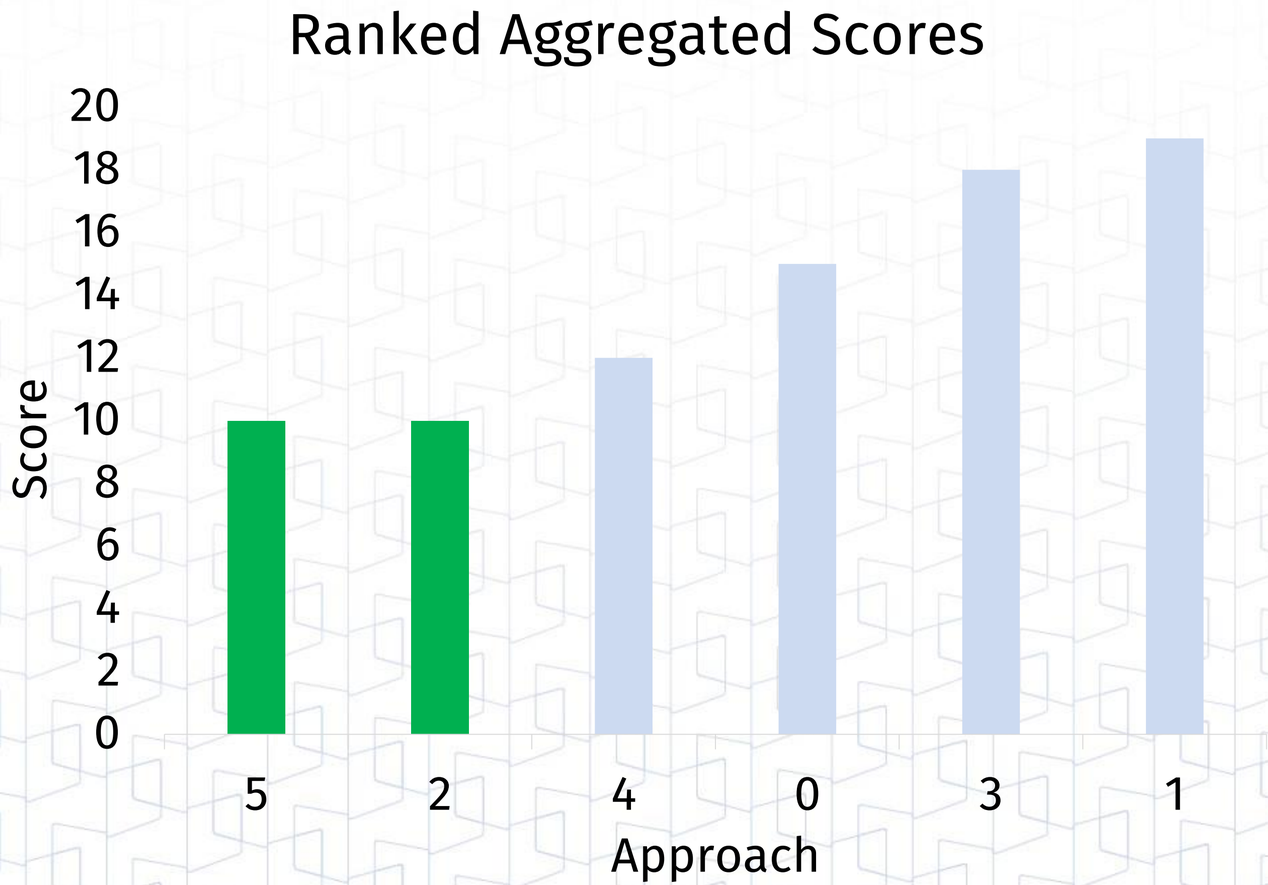
Approaches for the title and abstract scored automatically

- Generated results using 6 combinations of heuristics and prompts (with and without cataloguer rules)

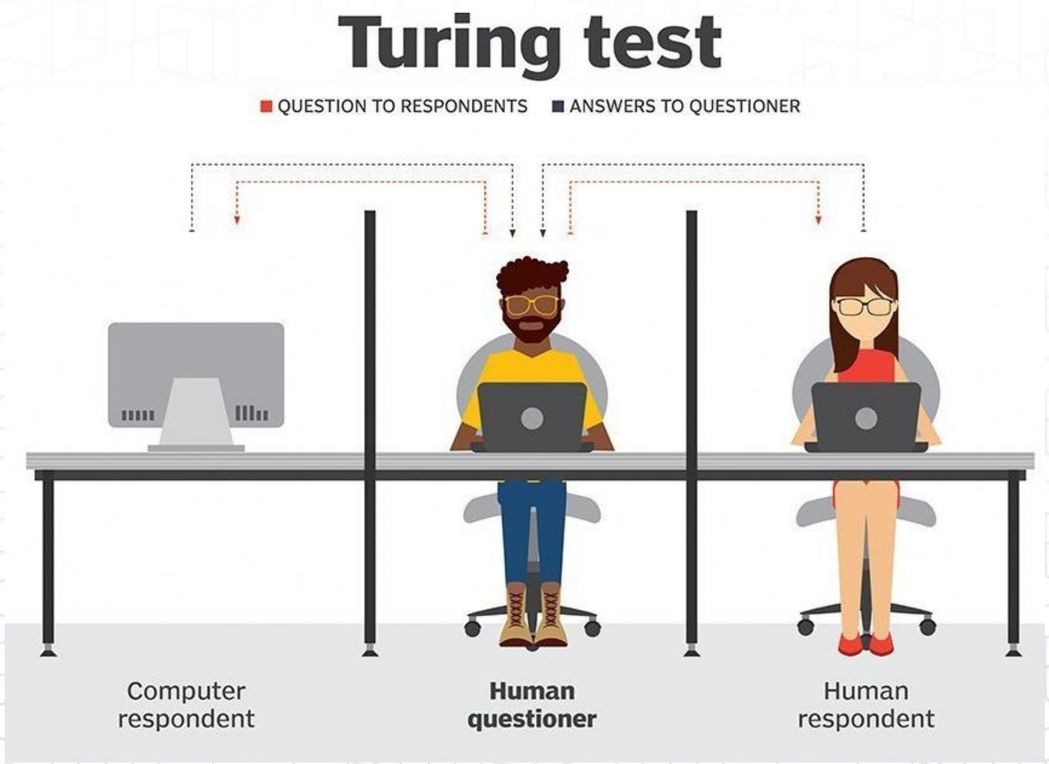
Selection criteria

- Maximum median of BERTScore
  - Semantic similarity between generated and manually curated abstracts
- Minimum median of Levenshtein
  - Character dissimilarity between generated and manually curated titles
- Minimum standard deviation

- Selected Option **5 (H3 Without Rules)** and **2 (H2 With Rules)**



The Turing Test: Output evaluated by cataloguers



Objective

- Trained cataloguers manually evaluated whether 3 versions of title and abstract were valid (criteria listed below):
  - Version 1: Generated by gpt-4o, H2 with rules
  - Version 2: Generated by gpt-4o, H3 without rules
  - Version 3: Written by human cataloguer
- Evaluation without knowledge of the source

Criteria

Content must:

- accurately represent the website
- be relevant to the topic
- not include subjective interpretations, biased language, exaggerations, or misleading claims

Test Results

Null Hypothesis (H_0) :

- There are no significant differences between human and AI-generated titles and abstracts.

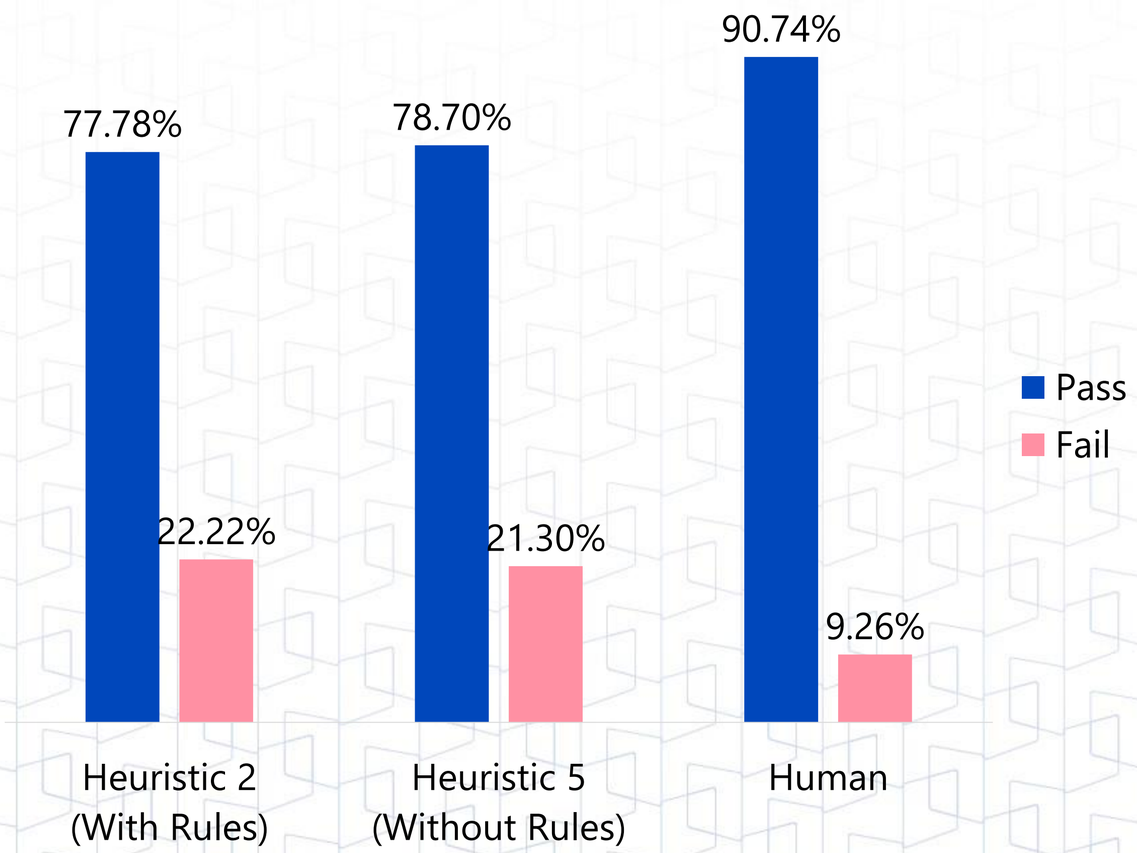
Alternative Hypothesis (H_a) :

- At least one condition is different.

Result:

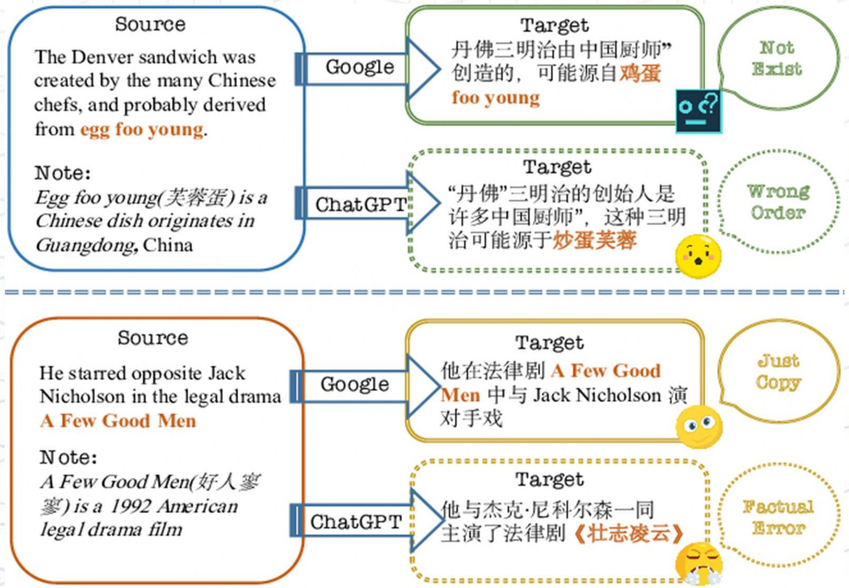
- p value (0.02) < 0.05
- Reject the Null Hypothesis

The titles and abstracts generated by GPT-4o do not have the same quality as those created by humans.

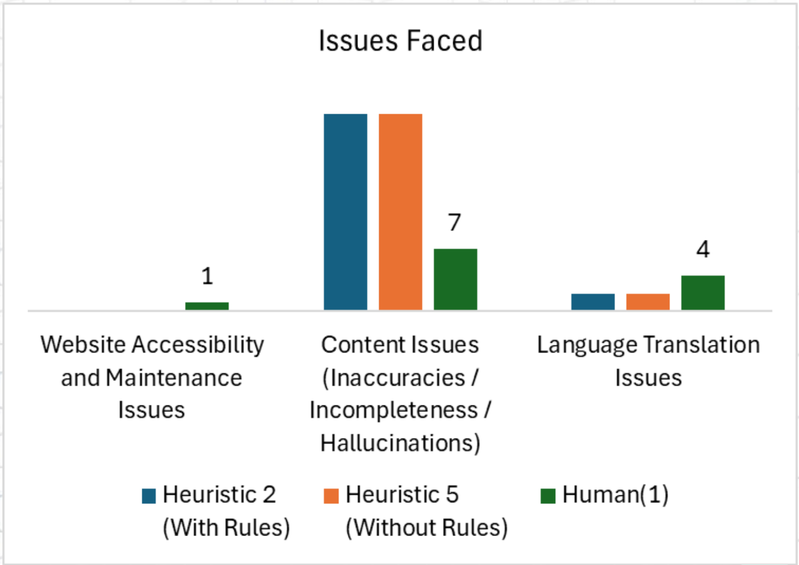


Issues in Generation and Evaluation

- **Title/Abstract Generation Issues**
  - Multi-lingual websites
  - gpt-4o may hallucinate and give results in multiple languages
  - Prompting for results in English does not guarantee English results
  - Accurate machine translation is not easy to obtain



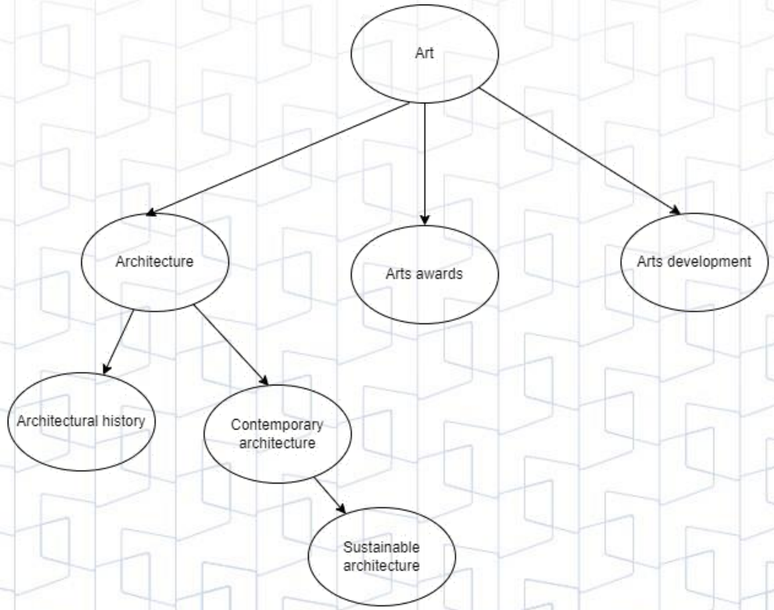
- **Evaluation Issues**
  - Can't access the website
  - Subjective grading based on cataloguer judgement
  - Biased language (Promotional Content)
  - Hallucinations
  - Different results on a different day

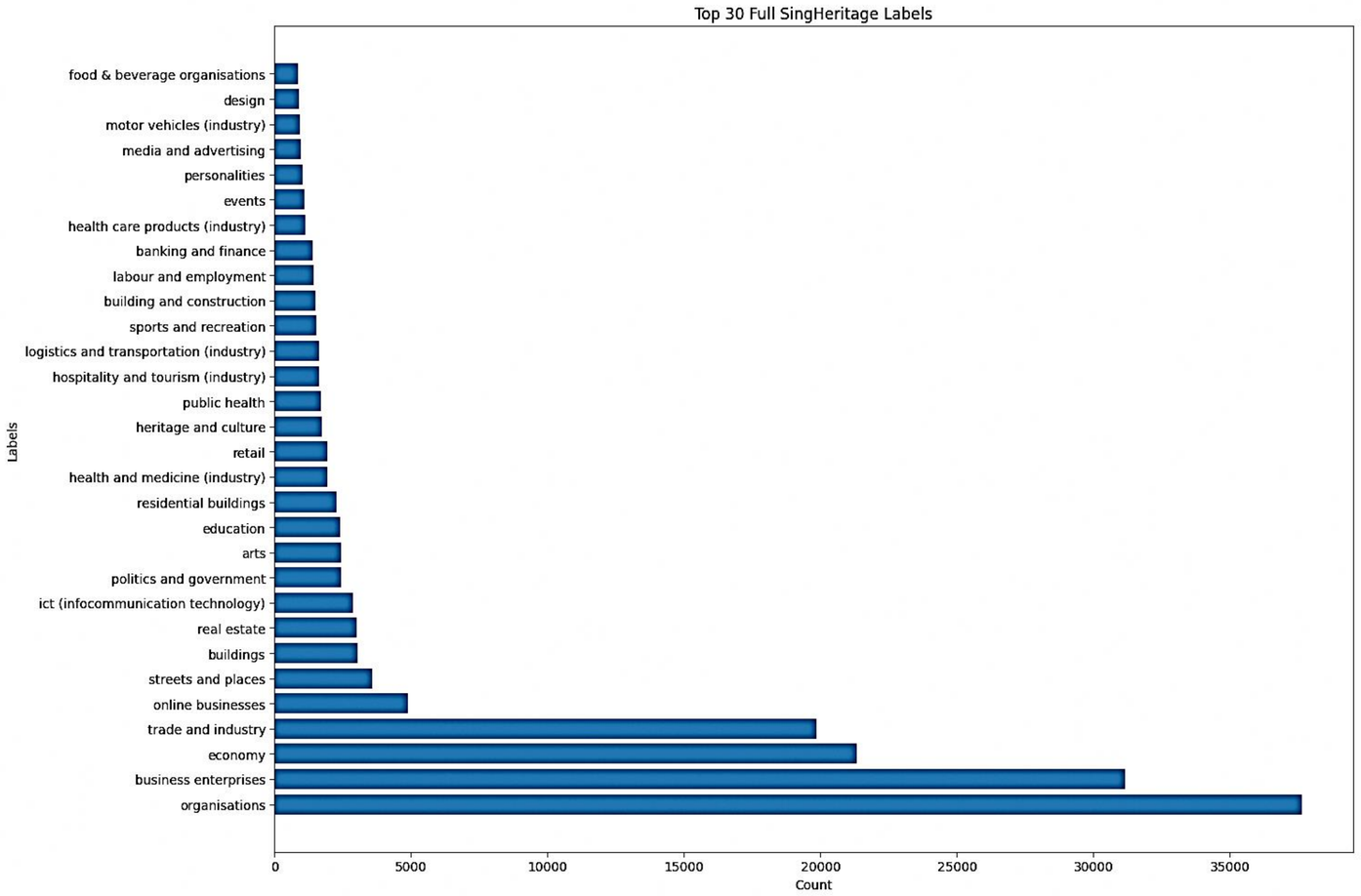


Subject Assignment

Singheritage Vocabulary

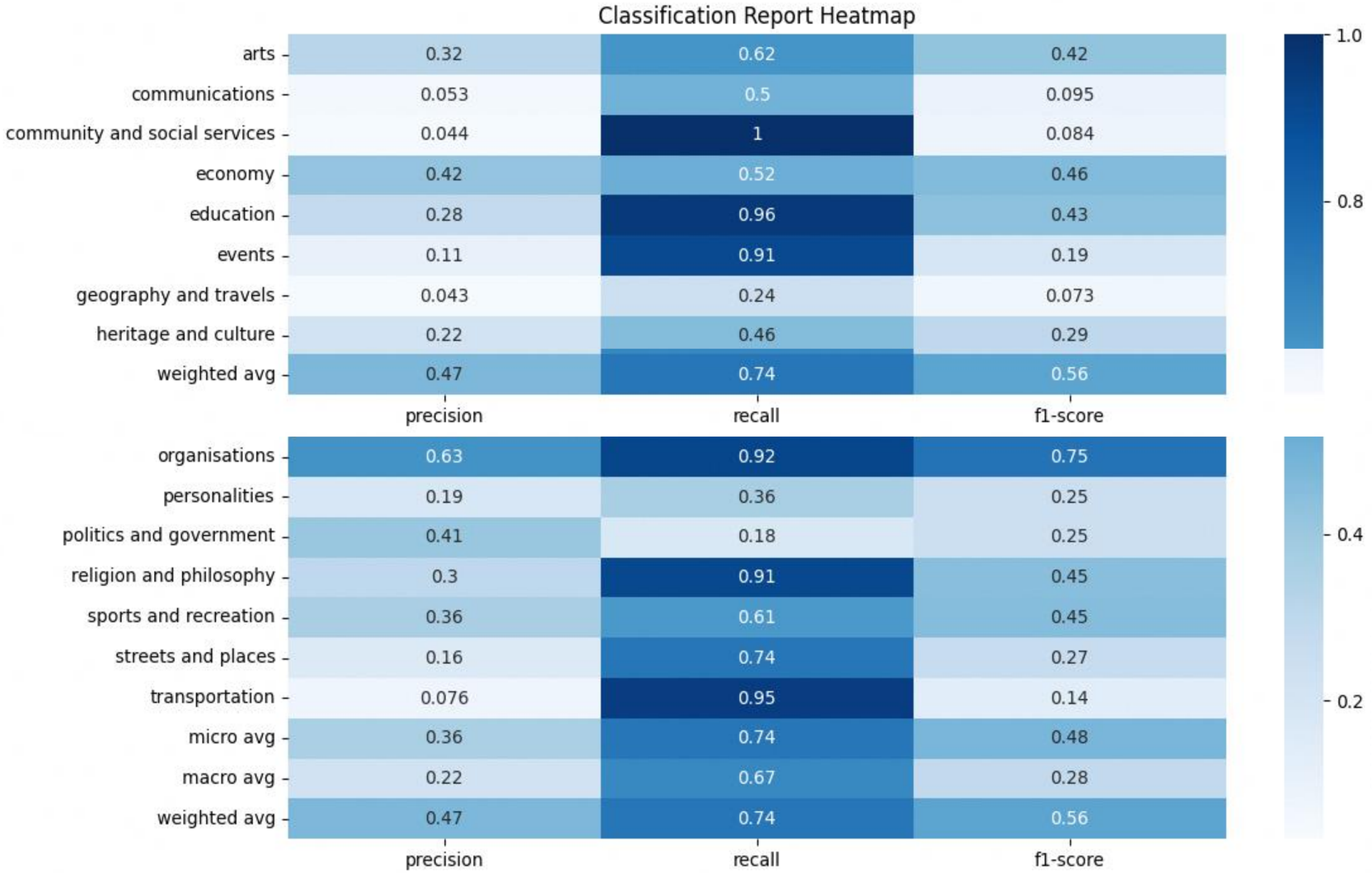
- Local taxonomy used for navigational purposes
- 1148 labels
- 8 levels
- Imbalanced data





Singheritage Assignment using GPT-4o

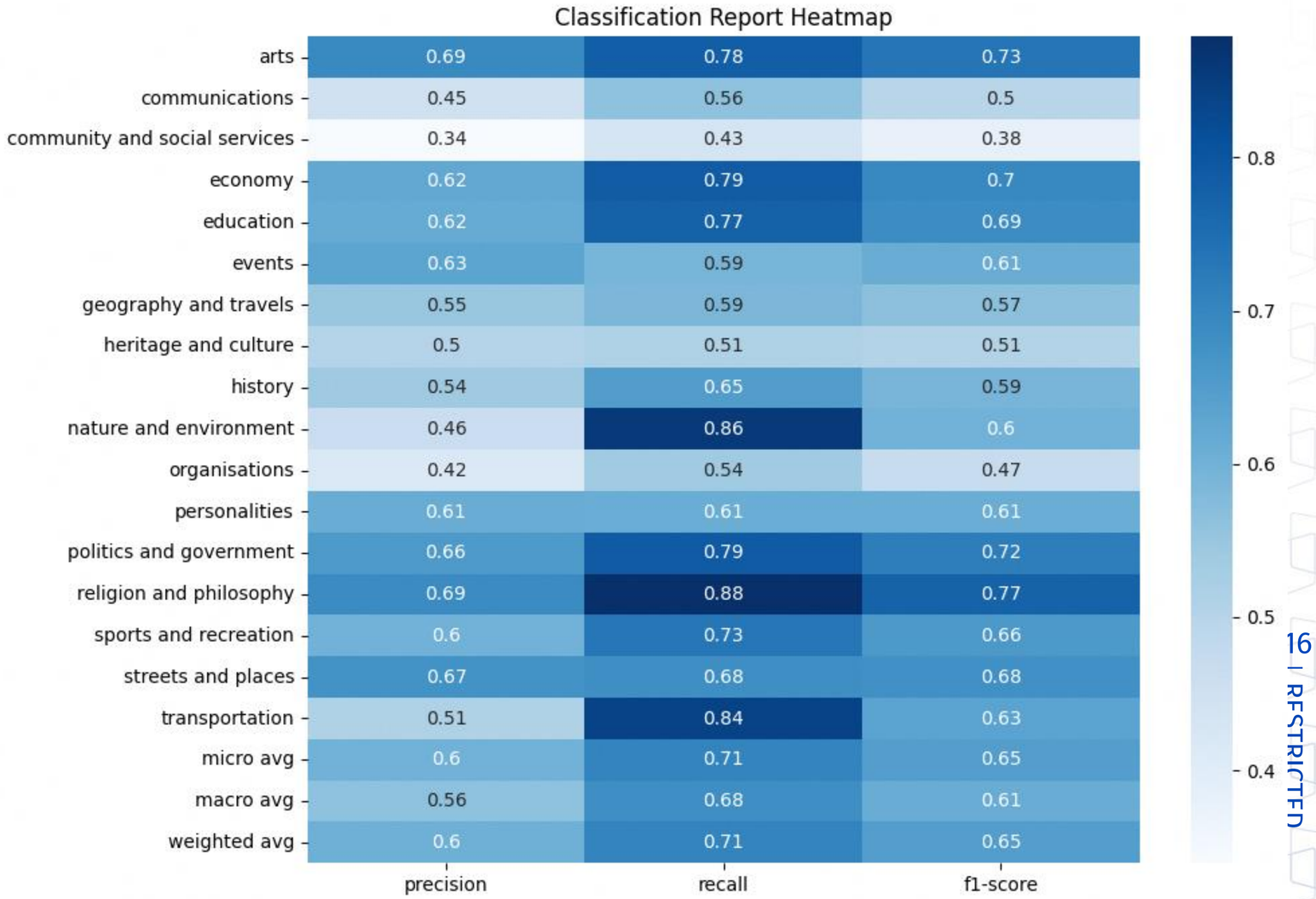
- 1st level
- 17 subjects
- 6271 abstracts
- gpt-4o selects labels for each abstract
- Compared against the manually curated dataset
- Weighted Average F1 Score ≡ 56%



SingHeritage Assignment using Graph Neural Networks

SingHeritage Hierarchy Level	Weighted F1 Score
1	65%
2	62%
3	56%

- More computational resources (GPUs) required to classify deeper levels
- Higher accuracy is not guaranteed



Content Quality Assessment

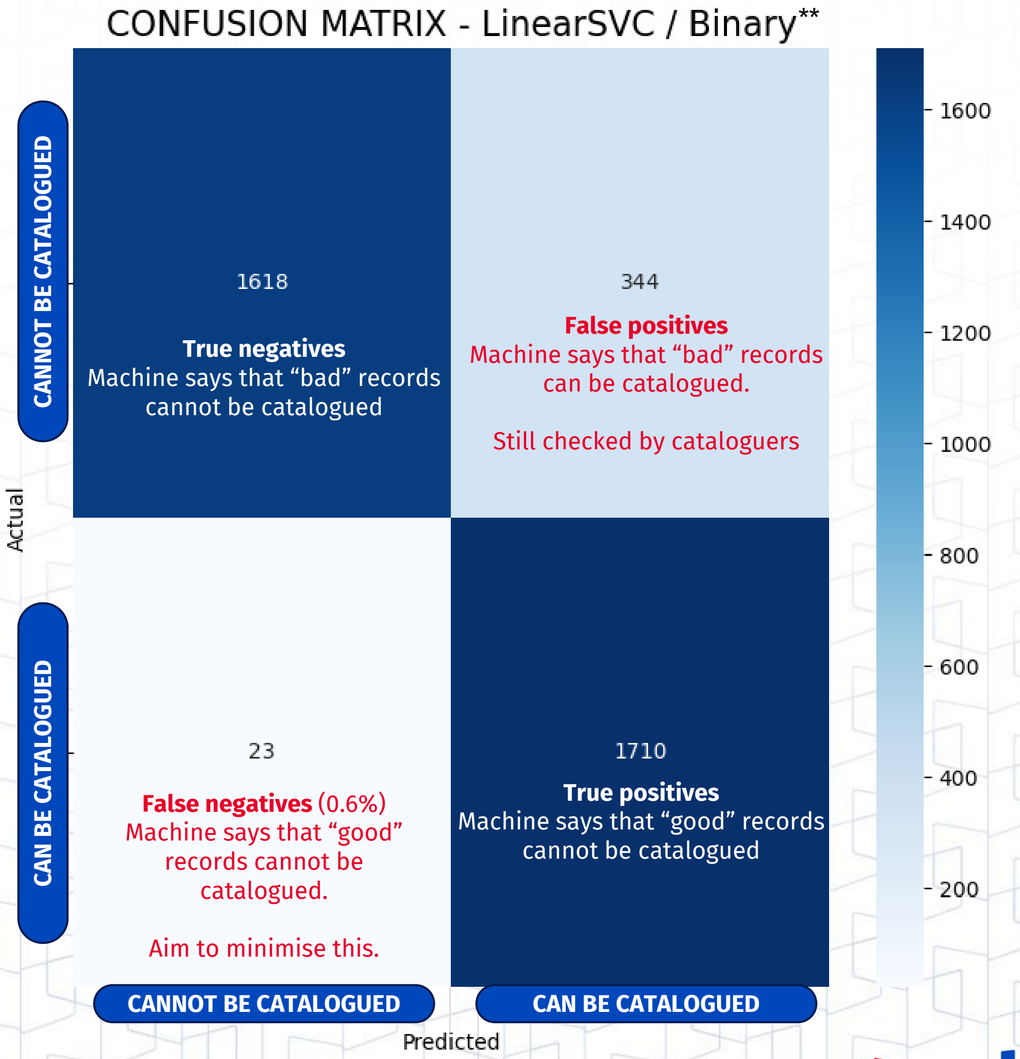
Upfront Quality Assessment: Removing unusable websites.

- **Not all websites can be catalogued.** >50% of archived websites cannot be opened or contain insufficient content (filler info, error messages etc.)
- **GI-GO.** AI is dependent on the quality of input data. Poor inputs are more likely to elicit hallucinations or incorrect responses.
- Pre-filtering can save cost and improve output quality

PROCESS

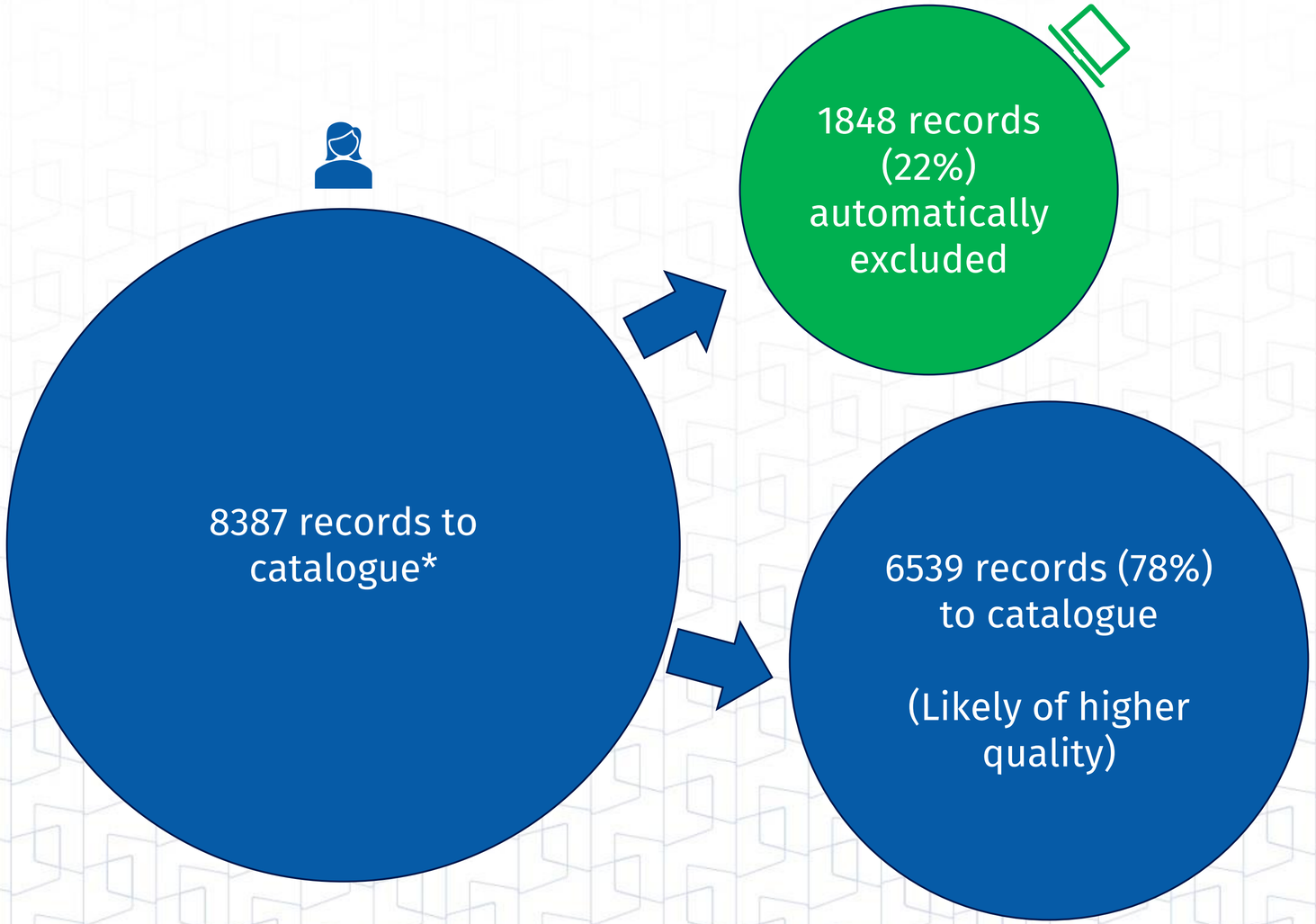
Archived websites are scraped for text → Machine learning model* trained & tuned on past cataloguer data → Trained model makes predictions on text from unseen websites

**Support vector classifier*



***Results from test set*

Upfront Quality Assessment: Removing unusable websites.



There is still room to improve performance!

**Actual implementation on a batch of records*

What's next?

- **The current results are promising, but not good enough yet**
  - 77-78% of generated Titles and Abstracts were valid, but a substantial number still suffer from hallucinations and biased language
  - Subject Assignment needs more investigation and synthetic data to train from.
- **Models are updated at a rapid pace, but there is a need for refinement**
  - Subsequent testing with another model, Google Gemini 1.5 Flash, without refined prompts resulted in only 50% valid Titles and Abstracts.
  - Choice of model and prompt engineering are necessary for better performance.
- **Alternatives include training a local model or using small models (need $$$ and technical expertise)**
  - Automating the workflow is just as important as using AI/ML, or there won't be savings

Without automation, time savings are smaller

Automated workflows enhance speed significantly

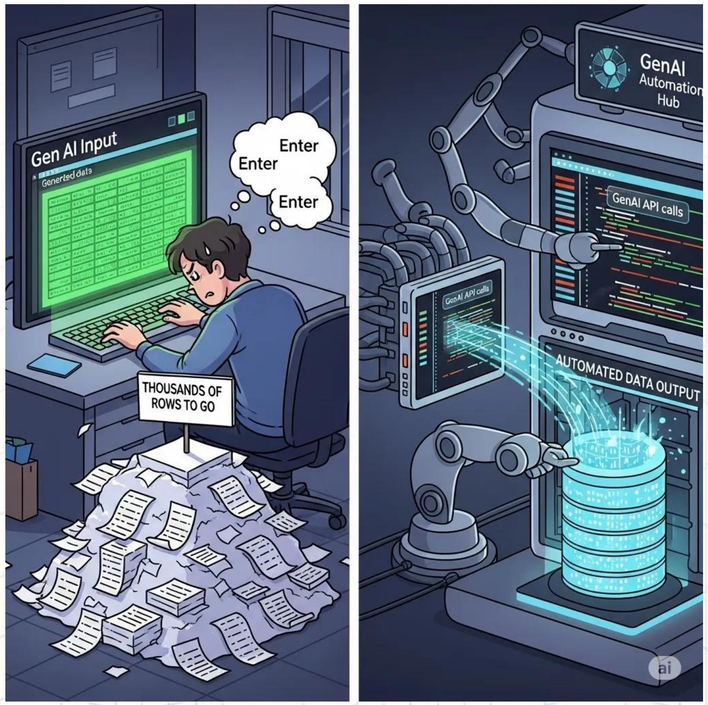


Image generated by AI for illustration

Case Study 2: SGCAT

SGCAT

A custom GPT prototype cataloguing assistant that can draft bibliographic records based on order & selection information and internal cataloguing practices to increase cataloguing efficiency

Developed by: Yogasvari, Mindy, Hui Ling, Munifah

OVERVIEW

Why SGCAT?

PROBLEM: Cataloguing is highly manual, effortful and time-consuming.

- Cataloguing is a highly manual activity that takes substantial focused manhours.
- Libraries need to catalogue books that are new to library databases.
- This involves going through stacks of new books individually to create valuable metadata according to comprehensive standards (e.g. MARC21, RDA) so that resources can be discovered by our users.

THUS, WE AIM TO:

- Automate routine aspects of cataloguing to enable cataloguers to focus on qualitative metadata enhancement
- Improve the speed and efficiency of cataloguing workflows

Joanna Maciejewska (MoSS is on preorder now!)
@AuthorJMac

You know what the biggest problem with pushing all-things-AI is? Wrong direction.
I want AI to do my laundry and dishes so that I can do art and writing, not for AI to do my art and writing so that I can do my laundry and dishes.

7:50 PM · Mar 29, 2024 · **3.2M** Views

What is cataloguing's "laundry"?

What is cataloguing?

Typically, 2 components to cataloguing:

DESCRIPTIVE CATALOGUING

About characteristics of the physical item:

- Who is the author?
- What is the title? Subtitle?
- How many pages?
- Is there an index?

Factual, not subjective:
"It is what it is"

Potential for automation with AI!

The "Laundry"

SUBJECT CATALOGUING

About the item's content

- What is the book about?
- Where would we shelve this item?

Based on cataloguer's judgment

SGCAT: A GPT-powered prototype quick cataloguing assistant

SGCAT semi-automates descriptive cataloguing

Broadclass
Information provider by selectors which may suggest associated headings or classification that we can get the GPT to process accordingly.

For example, TH would likely indicate that this is a thriller, for which the genre heading "thrillers (fiction)" would be applied).

Internal data sources

Order information
[ISBN, Title, Author, Publisher, Series, Edition, Broadclass, Remarks, etc]

NLB cataloguing guidelines

SGCAT

Custom GPT

Google Books API

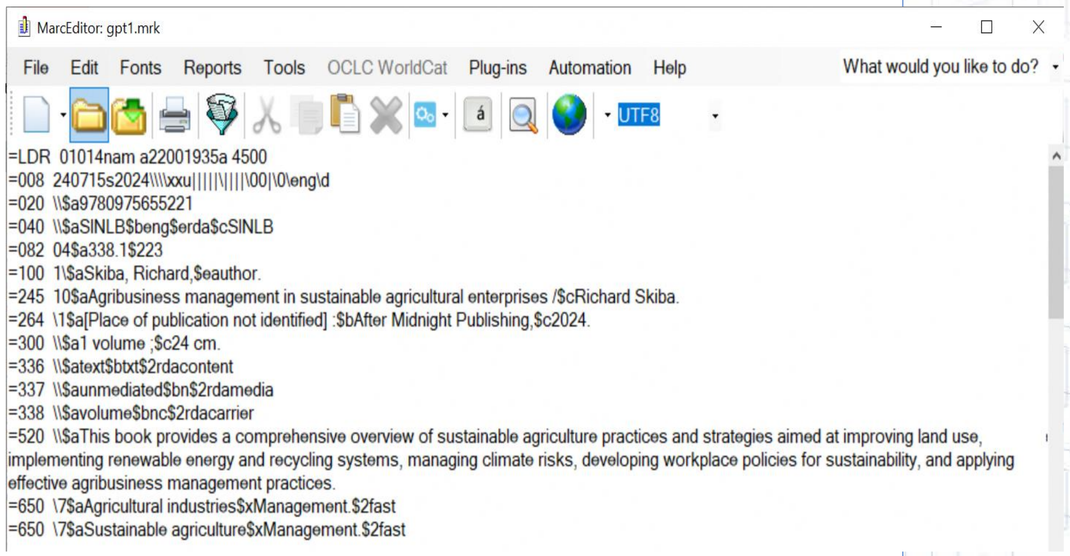
Open Library API

Enhance & cross-validate with external data sources

Seeking out in particular, information such as abstracts, which cataloguers can review; instead of crafting from scratch.

Scripted validation & editing in MarcEdit

MARC record



SGCAT: A GPT-powered prototype quick cataloguing assistant

Benefits: Improved efficiency

AI as a copilot to human cataloguers.

Shift in cataloguer's role

Creator → Reviewer

Streamline bibliographic metadata creation

Cuts cataloguing time by nearly 3x

Automate the mundane to focus on the **fun**

Cataloguers can focus more on qualitative enhancements to **enhance discovery**:

- Subject cataloguing
- Data disambiguation for accuracy
- Entity management
- Identify & capture entity-relationship information to represent the resource for better discovery

What's next for SGCAT?
Future Explorations

AI is not perfect. **Human participation is still required.**

Resolving issues in fixed fields

Prompt engineering to instruct GPT on fixed fields.

Explore other formats

Reviewing our knowledge base/documents to set up other formats like serials/AV.

Expand scope to non-English material

Explore romanisation of non-Latin scripts.

For Chinese language titles, the title, author, and publisher information in Chinese should be captured in separate MARC tag 880. In contrast, the Romanised title, author, and publisher information should be captured in the corresponding regular MARC tags. MARC Tag 880 should be associated with its corresponding regular MARC Tags by subfield 6 (linkage). An example is given below:

=100 1\$6880-01$aMurong Xue Cun,$eauthor.
=245 10$6880-02$aZhongguo, shao le yi wei yao /$cMurong Xue Cun.
=264 \1$6880-03$aBeijing :$bZhongguo he ping chu ban he,$c2010.
=880 1\$6100-01/1a慕容雪村,$eauthor.
=880 10$6245-02/$1$a中国, 少了一味药 /$c慕容雪村.
=880 \1$6264-03/$1$a北京 :$b中国和平出版社,$c2010.

Summary

Overall SWOT Analysis

	Good	Bad
Internal	**Strengths** • Faster, cheaper • Long term solution	**Weaknesses** • Hallucinations and mistakes • Manual intervention needed to identify/fix errors • Garbage in, Garbage out
External	**Opportunities** • Performance improvements with finer model tuning	**Threats** • Dependent on service providers • Privacy and copyright issues

Any questions?

Your turn: DISCUSSION

AI/automation and metadata

1. Small group discussions in breakout rooms
2. Regroup to share your learning points.

DISCUSSION QUESTIONS

Joanna Maciejewska (MoSS is on preorder now!)
@AuthorJMac

You know what the biggest problem with pushing all-things-AI is? Wrong direction.
I want AI to do my laundry and dishes so that I can do art and writing, not for AI to do my art and writing so that I can do my laundry and dishes.

7:50 PM · Mar 29, 2024 · **3.2M** Views

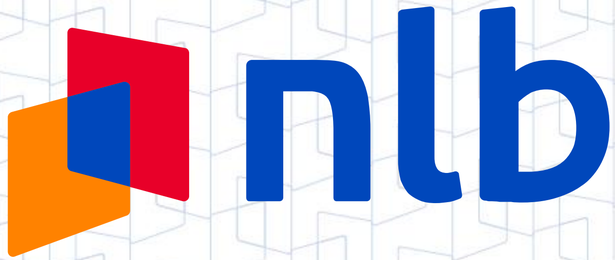
1. In the library world, are there equivalents of "art and writing" and "laundry and dishes"?

2. Have you started working on metadata projects with AI? How did you get started? Share some of the services or tools that you found was useful?

3. For those who haven't started AI projects at your work, what are some areas that are of interest to you? Share a theoretical AI project you think might be a good fit for your organisation.

4. What are some characteristics of the current state of AI/technology that limits their use in library work?

Please share your feedback with LAS



THANK YOU