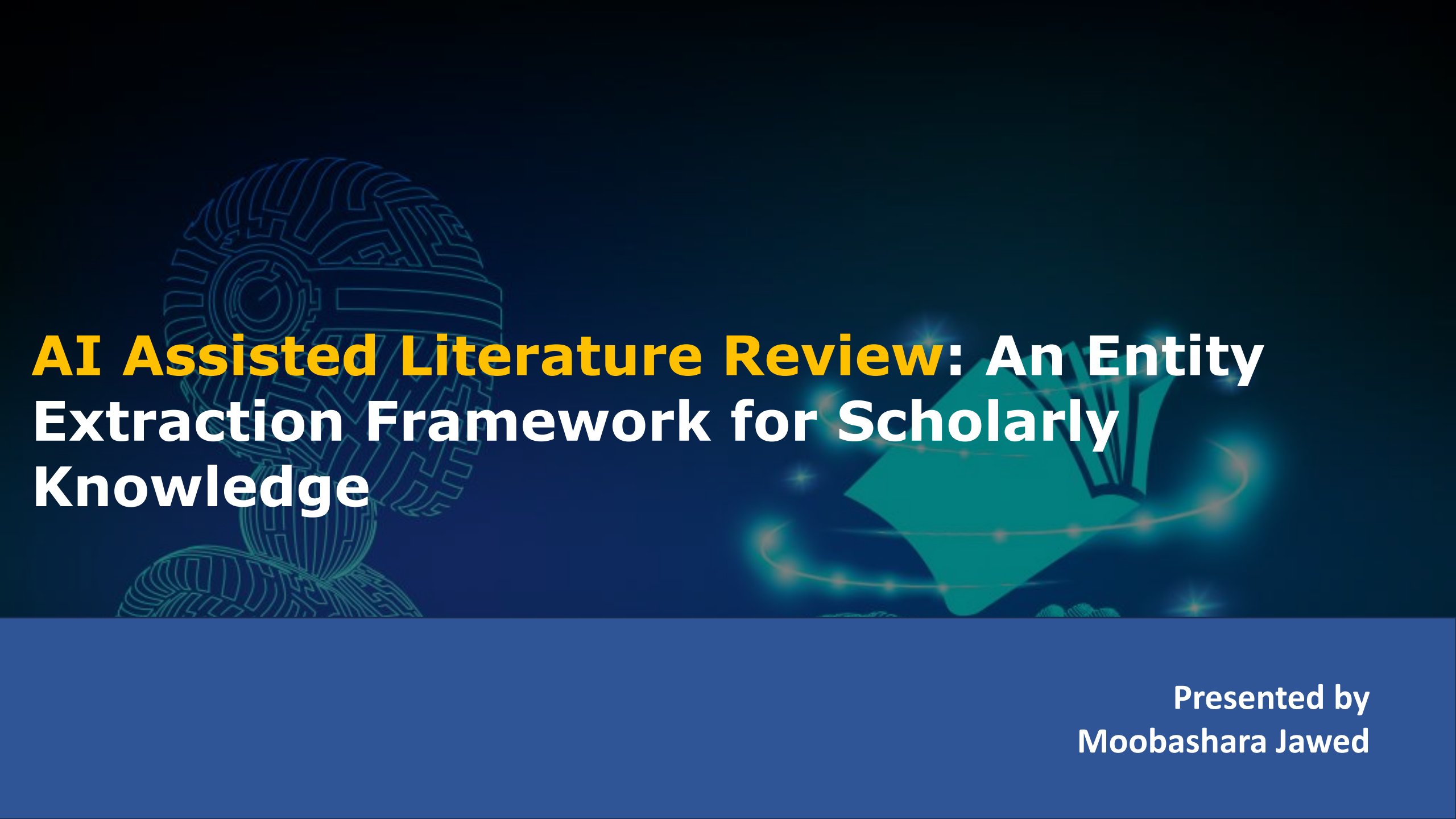


AI Assisted Literature Review: An Entity Extraction Framework for Scholarly Knowledge

Presented by
Moobashara Jawed

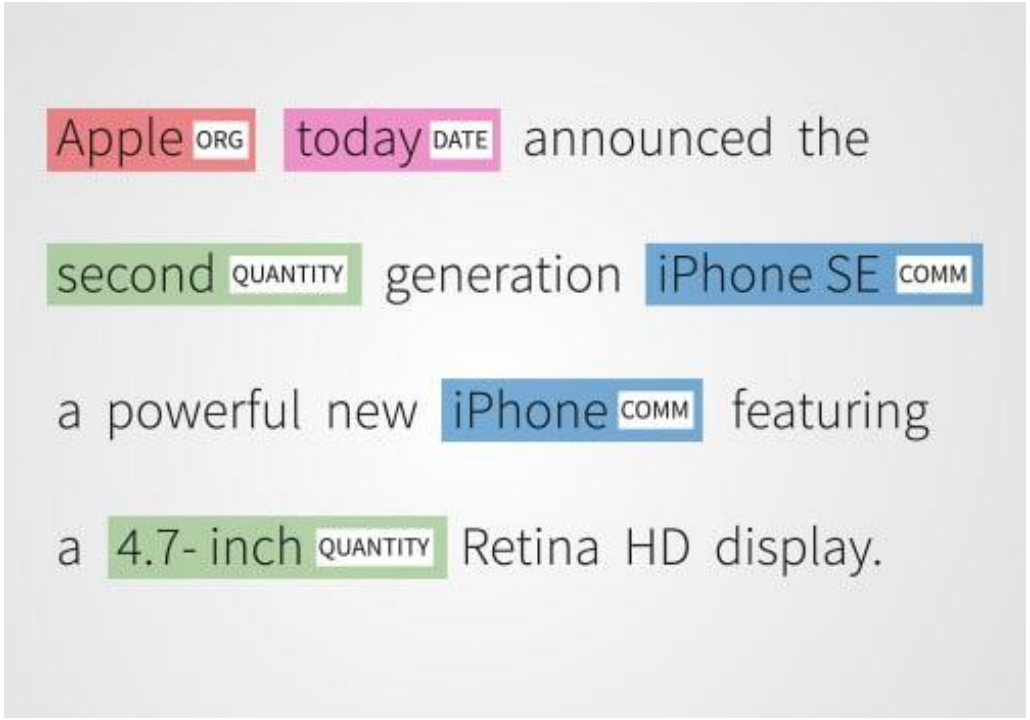
Named Entity Recognition (NER)

What is Entity Recognition?

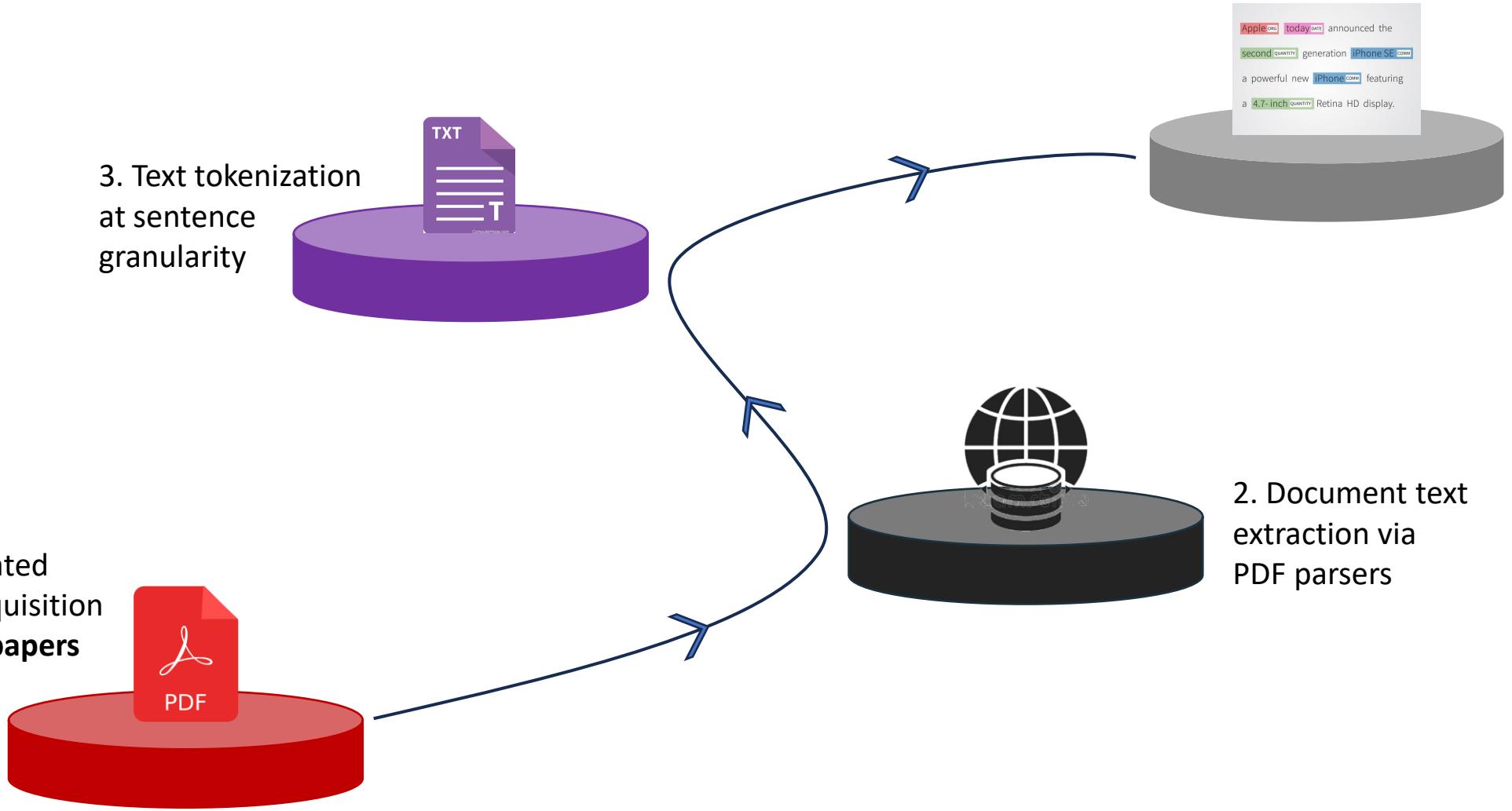
Automated identification of key terms (entities) in text into categories like diseases, species, or locations.

Why is it Needed?

- Manual reading and curation are time-consuming and inaccessible for many learners.
- Enables faster knowledge discovery, integration, and reasoning.
- Automated extraction democratizes knowledge access—from school students to expert researchers.



Workflow



1. Automated article acquisition via **Pygetpapers**

2. Document text extraction via PDF parsers

3. Text tokenization at sentence granularity

4. Named Entity Recognition (NER)

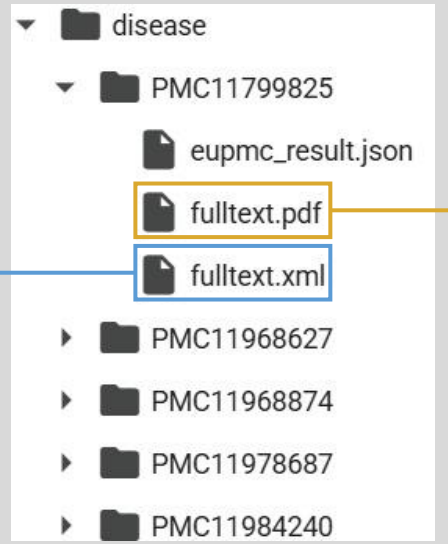
Step 1: Automated article acquisition via Pygetpapers

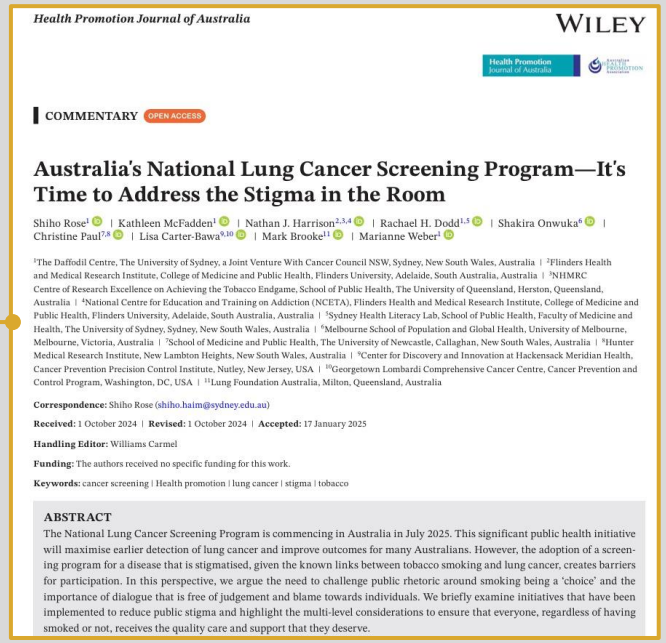
```
!pygetpapers --query '"lung cancer"' --xml --pdf --limit 5 --output disease --save_query
```

```
INFO: Total Hits are 578319
5it [00:00, 57456.22it/s]
INFO: Saving XML files to /content/disease/*/fulltext.xml
  0% 0/5 [00:00<?, ?it/s]INFO: Wrote the pdf file for PMC11968627
 20% 1/5 [00:03<00:14,  3.73s/it]INFO: Wrote the pdf file for PMC11984240
 40% 2/5 [00:06<00:09,  3.22s/it]INFO: Wrote the pdf file for PMC11968874
 60% 3/5 [00:10<00:06,  3.45s/it]INFO: Wrote the pdf file for PMC11978687
 80% 4/5 [00:15<00:04,  4.19s/it]INFO: Wrote the pdf file for PMC11799825
100% 5/5 [00:17<00:00,  3.48s/it]
```

This XML file does not appear to have any style information associated with it. The documen

```
<article xmlns:mml="http://www.w3.org/1998/Math/MathML" xmlns:xlink="http://www.w3.
  <?DTDIdentifier.IdentifierValue -//NLM//DTD JATS (Z39.96) Journal Archiving and I
  <?DTDIdentifier.IdentifierType public?>
  <?SourceDTD.DTDName JATS-archive-oasis-article1-mathml3.dtd?>
  <?SourceDTD.Version 1.1?>
  <?ConverterInfo.XSLTName jats-oasis2jats3.xsl?>
  <?ConverterInfo.Version 1?>
  <?properties open_access?>
  <processing-meta base-tagset="archiving" mathml-version="3.0" table-model="xhtml"
    <restricted-by>pmc</restricted-by>
  </processing-meta>
  <front>
    <journal-meta>
      <journal-id journal-id-type="nlm-ta">Health Promot J Austr</journal-id>
      <journal-id journal-id-type="iso-abbrev">Health Promot J Austr</journal-id>
      <journal-id journal-id-type="doi">10.1002/(ISSN)2201-1617</journal-id>
      <journal-id journal-id-type="publisher-id">HPJA</journal-id>
      <journal-title-group>
        <journal-title>Health Promotion Journal of Australia</journal-title>
      </journal-title-group>
      <issn pub-type="ppub">1036-1073</issn>
      <issn pub-type="epub">2201-1617</issn>
      <publisher>
        <publisher-name>John Wiley and Sons Inc.</publisher-name>
        <publisher-loc>Hoboken</publisher-loc>
      </publisher>
    </journal-meta>
    <article-meta>
      <article-id pub-id-type="pmcid">11799825</article-id>
      <article-id pub-id-type="pmid">39910978</article-id>
      <article-id pub-id-type="doi">10.1002/hpja.70011</article-id>
      <article-id pub-id-type="publisher-id">HPJA70011</article-id>
      <article-id pub-id-type="other">5524741</article-id>
      <article-categories>
```





Health Promotion Journal of Australia WILEY

COMMENTARY OPEN ACCESS

Australia's National Lung Cancer Screening Program—It's Time to Address the Stigma in the Room

Shiho Rose[1] | Kathleen McFadden[1] | Nathan J. Harrison[2,3,4] | Rachael H. Dodd[1,5] | Shakira Onwuka[6] | Christine Paul[7,8] | Lisa Carter-Bawa[9,10] | Mark Brooke[11] | Marianne Weber[1]

[1]The Daffodil Centre, The University of Sydney, a Joint Venture With Cancer Council NSW, Sydney, New South Wales, Australia | [2]Flinders Health and Medical Research Institute, College of Medicine and Public Health, Flinders University, Adelaide, South Australia, Australia | [3]NHMRC Centre of Research Excellence on Achieving the Tobacco Endgame, School of Public Health, The University of Queensland, Herston, Queensland, Australia | [4]National Centre for Education and Training on Addiction (NCETA), Flinders Health and Medical Research Institute, College of Medicine and Public Health, Flinders University, Adelaide, South Australia, Australia | [5]Sydney Health Literacy Lab, School of Public Health, Faculty of Medicine and Health, The University of Sydney, Sydney, New South Wales, Australia | [6]Melbourne School of Population and Global Health, University of Melbourne, Melbourne, Victoria, Australia | [7]School of Medicine and Public Health, The University of Newcastle, Callaghan, New South Wales, Australia | [8]Hunter Medical Research Institute, New Lambton Heights, New South Wales, Australia | [9]Center for Discovery and Innovation at Hackensack Meridian Health, Cancer Prevention Precision Control Institute, Nutley, New Jersey, USA | [10]Georgetown Lombardi Comprehensive Cancer Centre, Cancer Prevention and Control Program, Washington, DC, USA | [11]Lung Foundation Australia, Milton, Queensland, Australia

Correspondence: Shiho Rose (shiho.haim@sydney.edu.au)

Received: 1 October 2024 | **Revised:** 1 October 2024 | **Accepted:** 17 January 2025

Handling Editor: Williams Carmel

Funding: The authors received no specific funding for this work.

Keywords: cancer screening | Health promotion | lung cancer | stigma | tobacco

ABSTRACT

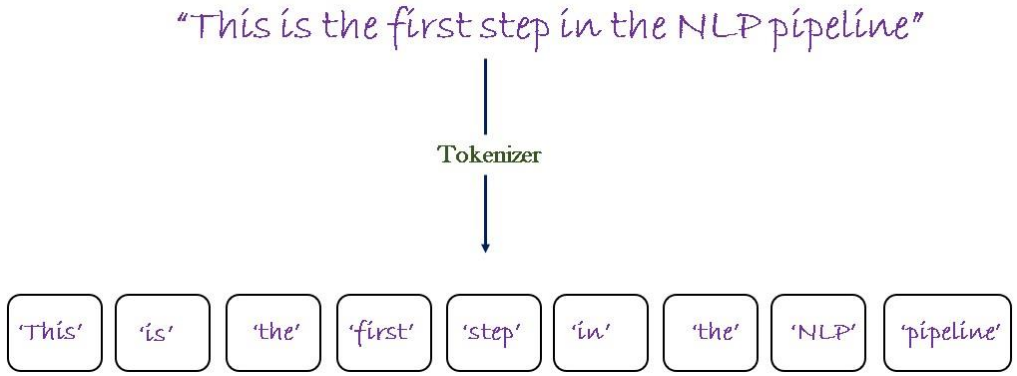
The National Lung Cancer Screening Program is commencing in Australia in July 2025. This significant public health initiative will maximise earlier detection of lung cancer and improve outcomes for many Australians. However, the adoption of a screening program for a disease that is stigmatised, given the known links between tobacco smoking and lung cancer, creates barriers for participation. In this perspective, we argue the need to challenge public rhetoric around smoking being a 'choice' and the importance of dialogue that is free of judgement and blame towards individuals. We briefly examine initiatives that have been implemented to reduce public stigma and highlight the multi-level considerations to ensure that everyone, regardless of having smoked or not, receives the quality care and support that they deserve.

Step 2 & 3: PDF parsing and Text tokenization

2. Parsing is the process of extracting and structuring data from documents, such as PDFs.

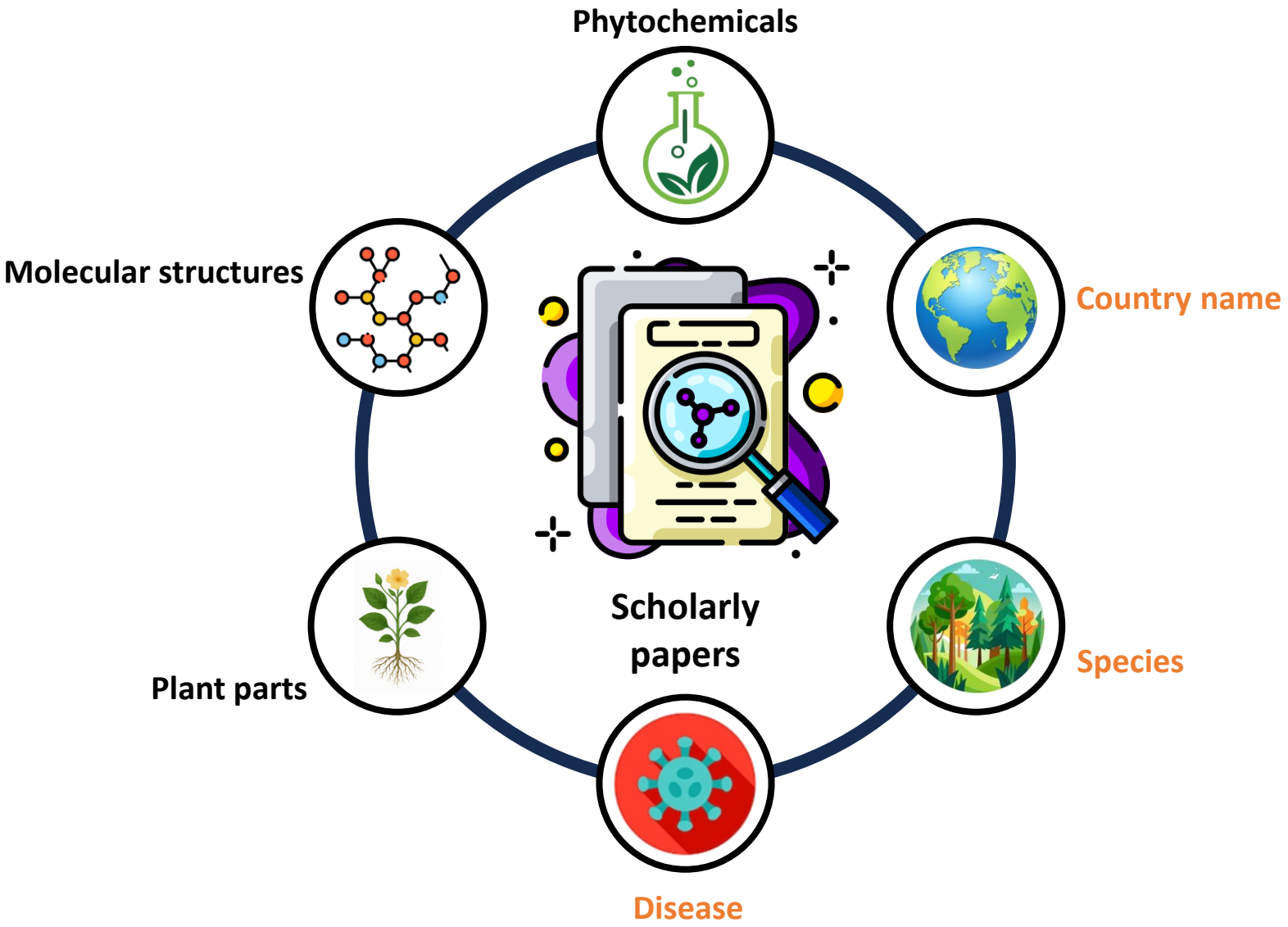
- PDF Parsing Packages: **fitz** (**from PyMuPDF**)
- **Fitz**: Extracts raw texts from pdfs

3. Tokenization is the process of splitting text into smaller units, such as sentences or words (tokens).



- Tokenization: **nltk.tokenize.sent_tokenize**
- Splits text into sentences

Step 4: Named Entity Recognition (NER)



Country Extraction

- NER Model: **spaCy en_core_web_lg**
- Validation: **Babel**'s official country list

Frequency table

	Country	Frequency	PMC11984240	PMC11799825	PMC11978687	PMC11968627	PMC11968874
1	China	22	7	0	6	4	5
2	Australia	16	0	15	0	0	1
3	Taiwan	1	1	0	0	0	0
4	Germany	1	1	0	0	0	0
5	Switzerland	1	0	0	0	0	1

Word cloud

Pearson Correlation between Countries

	Country1	Country2	PearsonCorrelation
0	Germany	Taiwan	0.993999
3	Germany	Switzerland	0.935414
1	Switzerland	Taiwan	0.902454

Species Extraction

- NER Model: scispaCy en_core_sci_md (scientific text)
- Validation:
  - Regex for binomial nomenclature (Genus species)
  - GBIF API verification (EXACT species matches)

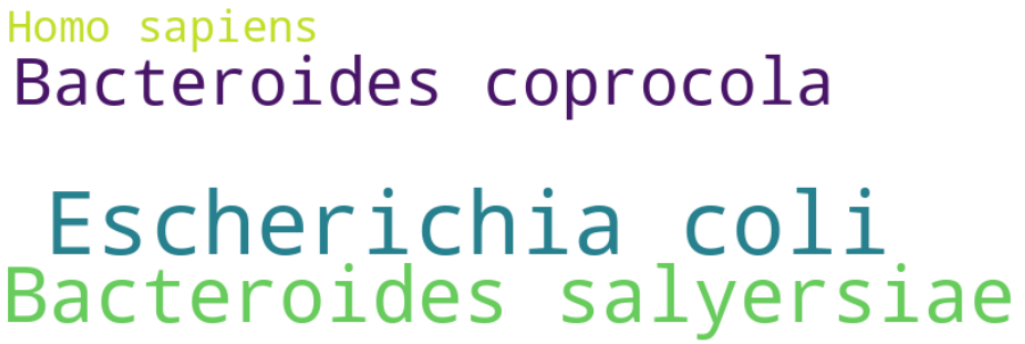
Frequency table

	Species	Frequency	PMC11984240	PMC11799825	PMC11978687	PMC11968627	PMC11968874
1	Escherichia coli	4	0	0	0	4	0
2	Bacteroides salyersiae	4	0	0	0	4	0
3	Bacteroides coprocola	3	0	0	0	3	0
4	Homo sapiens	1	0	0	0	0	1

Pearson Correlation between Specieses

	Species1	Species2	PearsonCorrelation
0	Bacteroides salyersiae	Escherichia coli	0.980154
3	Bacteroides coprocola	Bacteroides salyersiae	0.936586
1	Bacteroides coprocola	Escherichia coli	0.848528

Word cloud



Disease Extraction

- NER Model: scispaCy en_ner_bc5cdr_md (biomedical)
- Validation:
  - Disease Ontology (DOID) terms
  - Fuzzy matching (95% similarity threshold)

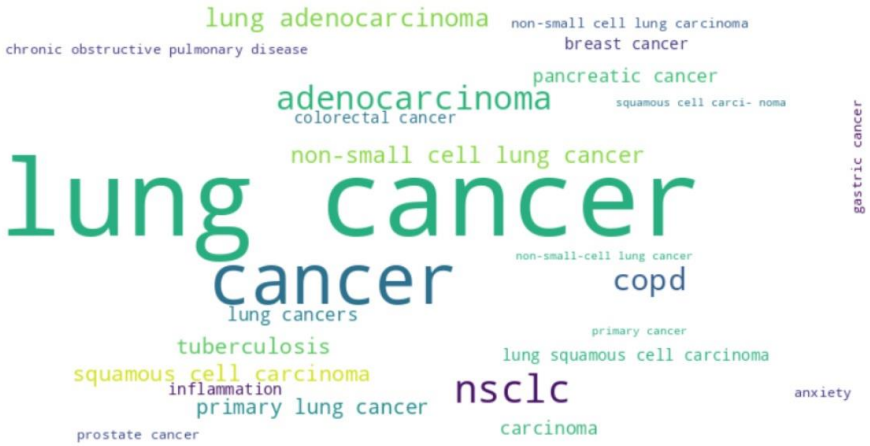
Frequency table

	Disease	Frequency	PMC11984240	PMC11799825	PMC11978687	PMC11968627	PMC11968874
1	lung cancer	383	76	38	66	80	123
2	cancer	104	16	9	9	10	60
3	adenocarcinoma	11	10	0	0	0	0
4	copd	10	9	0	0	0	0
5	nsclc	10	9	0	0	0	0

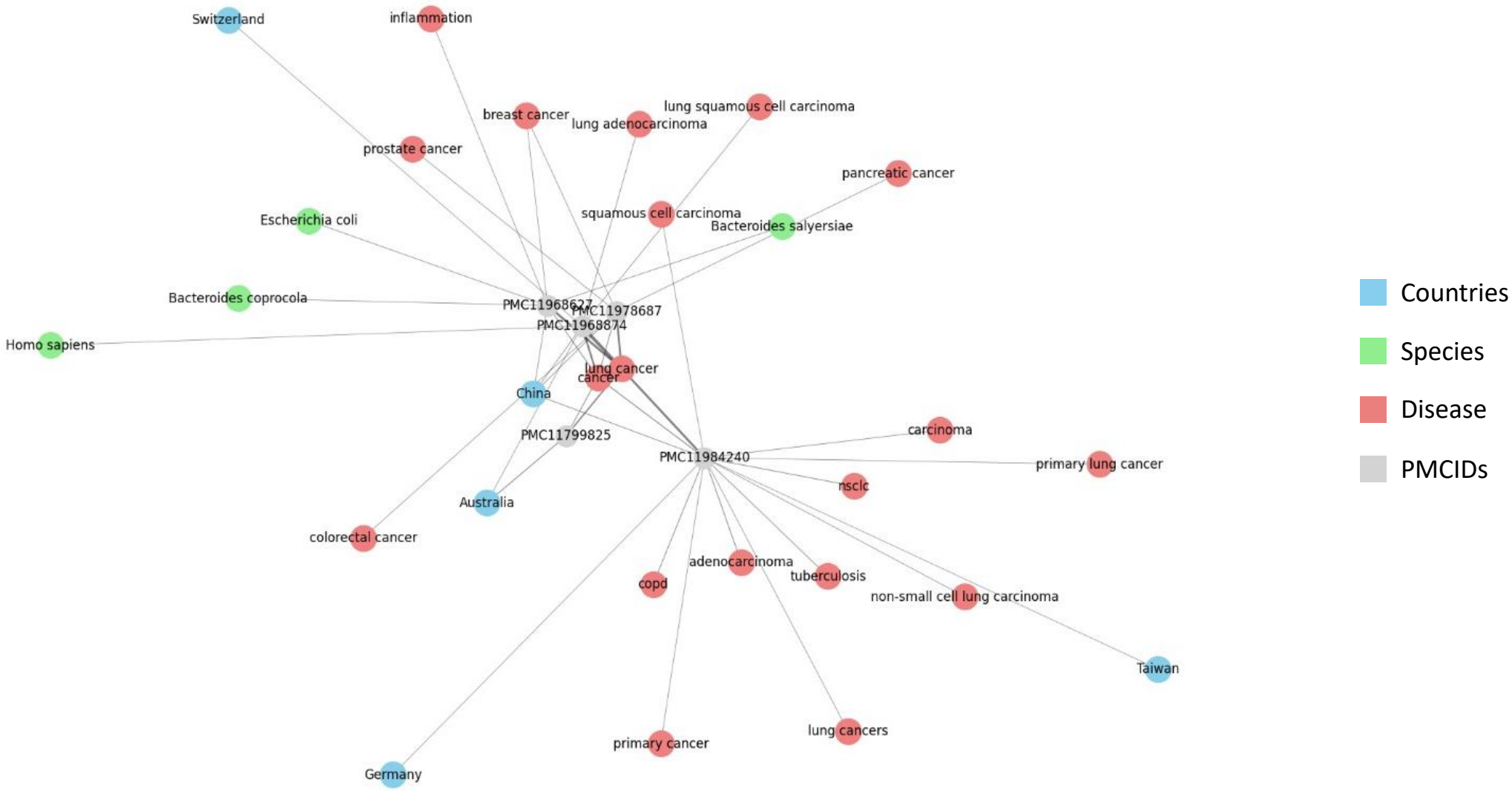
Pearson Correlation between Diseases

	Disease1	Disease2	PearsonCorrelation
370	non-small-cell lung cancer	squamous cell carci- noma	0.999993
389	non-small-cell lung cancer	primary cancer	0.996739
371	primary cancer	squamous cell carci- noma	0.996693
317	non-small cell lung carcinoma	primary cancer	0.995477
351	anxiety	squamous cell carci- noma	0.995002
352	anxiety	non-small-cell lung cancer	0.994845
333	chronic obstructive pulmonary disease	squamous cell carci- noma	0.994325
219	colorectal cancer	prostate cancer	0.994189
334	chronic obstructive pulmonary disease	non-small-cell lung cancer	0.994145
297	gastric cancer	squamous cell carci- noma	0.992542

Word cloud



Network of Countries, Species and Diseases



THANK YOU