Knowledge Graph-driven Tabular Data Discovery from Scientific Documents

Vijay S. Kumar[1], Varish Mulwad[2], Jenny Weisenberg Williams[1],

Tim Finin[3], Sharad Dixit[1], Anupam Joshi[3]

1: GE Research, Niskayuna, NY, USA
2: GE Research, Bengaluru, India
3: University of Maryland, Baltimore County, Baltimore, MD, USA

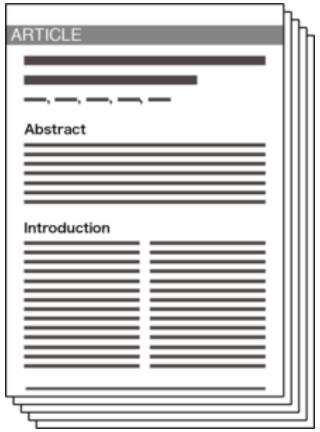
1st Tabular Data Analysis (TaDA) workshop,

VLDB, Sep 1, 2023

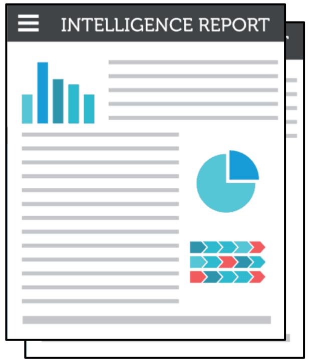
Documents and Tabular Data

Scientific/Technical documents

- Critical information embedded within structured elements (**tables**, charts, equations, …)
  - Supplement text with vital visual context
  - Structurally formatted for human consumption
- Increasing publication rates
  - Open-access, preprint servers, generative AI, …

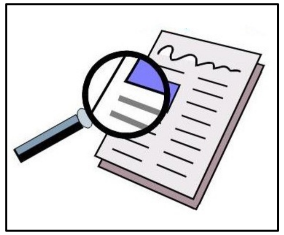


Scientific papers, preprint articles



Intelligence Reports

Patents



Maintenance manuals, legal agreements, etc.

Tables in Scientific Documents

- Significant volumes of tabular data locked away in these documents. Not easy to access & analyze
- Knowledge in tables critical to emerging applications
- Information discovery from documents focused on text and metadata; **Does not consider tabular data.**

Dataset	Document Type / Source	Domain	Corpus size	# tables
ChemTables	Patents / USPTO	Chemical	1,000	788
ArxivPapers	Preprints / arXiv	ML	104,723	277,996
ProCure (this work)	Papers & preprints / PubMed Central OA	Biomedical, clinical	62,777	120,417

We view scientific/technical documents as (also) a rich source of tabular data

Background

Extensive prior research on understanding information content of web tables, open data, and tables in enterprise data lakes

- annotate with semantic information → tables more discoverable
- address schema/data matching, data discovery and integration requirements

Recent advances in pre-trained / table representation learning models for **well-structured** tables

Some efforts specifically target tables in scientific/technical documents

Dataset	Downstream Task
PubTables-1M	Table detection, Table structure recognition
ChemTables	Table classification
ArxivPapers	Table extraction and segmentation
SciGen	Reasoning-aware table-to-text generation
TAT-QA	Question-answering over tables and text
S2abEL	Entity Linking for scientific tables

Discovery of relevant tabular data from (collections of) published documents is relatively under-explored

Motivating Use-cases

Information discovery for intelligence report generation and enhancement

Incorporate Effective Visual Presentations When Feasible

C-14. Analysts should present intelligence in a visual format to clarify an analytical conclusion and to complement or enhance the presentation of intelligence and analysis. In particular, visual presentations should be used when information or concepts, such as spatial or temporal relationships, can be conveyed better in graphic form, such as tables, flow charts, and images coupled with written text. Visual presentations

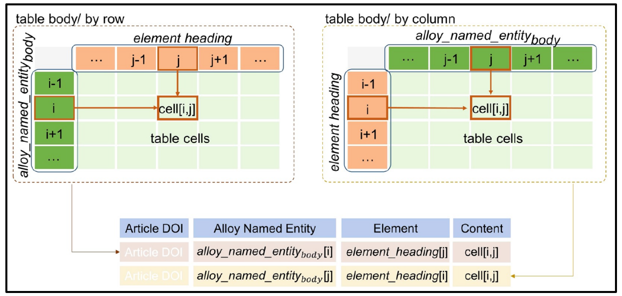
https://irp.fas.org/doddir/army/atp2-33-4.pdf

AI–assisted Scientific Research

1. Further augment understanding of and discovery from existing literature

- Allen AI's Semantic Reader, Elicit, SCISPACE, Explainpaper, SciSummary, HeyScience, AIrXiv

2. Help assemble training datasets (from documents) in low-data domains, e.g.,



Alloy design in materials science

Data Sets and Associated Data Creation/Preparation Tools (NSF APTO)

Data: e.g., aggregate historical data from lab notebooks and academic journals from 1730 to 2010 on telecommunication technologies' bandwidth, latency, and power requirements.

Forecasting technology trajectories

As search (over documents) gets more driven by generative AI, need a way to verifiably synthesize tabular data

Scientific Tables

1. Domain-specific Entities

- typically more numerical cell content than text
- text, where present, usually in the form of Literals

Table 2

Developed serology tests for SARS-CoV-2 detection by different companies and researchers.

Developer	Platform	Target antigen	Target antibody	Other features	References
Abbott Laboratories	CMIA	Nucleocapsid	IgG	Return 100–200 test results in 1 h, specificity 99.6%, and sensitivity of 100%	Abbott Laboratories (2020b)
DiaSorin	CMIA	Spike	IgG	Fully automated, quantitative, 97.4% sensitivity, 98.5 specificity	DiaSorin (2020)
Pharmact AG	Lateral flow assay	–	IgG and IgM	POC, results in 20 min, can determine the phase of the disease, 99.8% agreement with PCR for non-affected cases	Pharmact (2020)
Hangzhou Biotest Biotech	Lateral flow assay	Spike	IgG and IgM	100% specificity for IgM and IgG, 100% sensitivity	(Hangzhou Biotest Biotech

Similar to web tables ... with domain-specific entities

Table 1

Sensitivity and specificity of the Elecsys® Anti–SARS-CoV-2 and LIAISON® SARS-CoV-2 S1/S2 IgG tests.

Test and result	COVID-19 NAAT test result		Sensitivity (%)	Specificity (%)	PPV (%) (COVID-19 prevalence 1/5/10%)	NPV (%) (COVID-19 prevalence 1/5/10%)
	Positive ($n = 40$)	Negative ($n = 161$)				
Elecsys® Anti–SARS-CoV-2						
Positive	37	2[b]	92.5 (CI: 79.6–98.4)	98.8 (CI: 95.6–99.9)	42.9/79.7/89.2	99.9/99.6/99.2
Negative	3[a]	159				
LIAISON® SARS-CoV-2 S1/S2 IgG						
Positive	35	4[b]	87.5 (CI: 73.2–95.8)	97.5 (CI: 93.8–99.3)	26.2/65.0/79.7	99.9/99.3/98.6
Negative	5	157				

Similar to open data … less text, more numbers … with ranges, multi-value cells; merged cells

Scientific Tables

2. **High structural heterogeneity**, more so than web tables

- optimized for human consumption; minimize information overload
- information compaction to ensure tables fit under space constraints

Table 2

Performance of serological assays in dependence of time after onset of symptoms.

		IgA						IgG				
		S1-assay		N-assay				S1-assay		N-assay		
	n	pos.	% ($CI_{95\%}$)	pos.	% ($CI_{95\%}$)	*p*	κ	pos.	% ($CI_{95\%}$)	pos.	% ($CI_{95\%}$)	*p*
$\text{Sensitivity}_{0\text{-}3\text{ d}}$	16	5	31.2 (12.1–58.5)	2	12.5 (2.2–39.6)	n.s.	0.470	2	12.5 (2.2–39.6)	2	12.5 (2.2–39.6)	n.s.
$\text{Sensitivity}_{4\text{-}7\text{ d}}$	23	12	52.2 (31.1–72.6)	7	30.4 (14.1–53.0)	n.s.		4	17.4 (5.7–39.5)	7	30.4 (14.1–53.0)	n.s.
$\text{Sensitivity}_{8\text{-}10\text{ d}}$	24	16	66.7 (44.7–83.7)	9	37.5 (19.6–59.2)	0.016		11	45.8 (26.2–66.8)	14	58.3 (36.9–77.2)	n.s.
$\text{Sensitivity}_{11\text{-}13\text{ d}}$	17	17	100 (0.77–100)	13	76.5 (49.8–92.2)	n.s.		13	76.5 (49.8–92.2)	15	88.2 (62.3–97.8)	n.s.
$\text{Sensitivity}_{\geq 14\text{ d}}$	25	24	96.0 (77.7–99.8)	16	64.0 (42.6–81.2)	0.008		22	88.0 (67.6–96.8)	24	96.0 (77.7–99.8)	n.s.
$\text{Sensitivity}_{\text{outpat.}}$	65	63	96.9 (88.4–99.5)	4	6.2 (1.9–15.5)	<0.001	0.004	64	98.5 (90.6–99.9)	56	86.2 (74.8–93.1)	0.021
Specificity	139	8	94.3 (88.6–97.3)	0	100 (96.7–100)	<0.001	nd	1	99.3 (95.5–99.9)	0	100 (96.7–100)	n.s.

Row and column headers … sub-columns … abridged header cells

Characterization	System Count	Precision	Recall
Tables with Header Rows	113,582	1.00	0.94
Tables with Header Columns	48,733	1.00	0.55
Tables with Concise Header Rows	36,182	0.84	0.94
Tables with Multi-level Header Rows	32,169	1.00	0.97
Tables with ONLY Numeric Data Cells	12,969	1.00	0.83
Tables with Concise Body	40,158	0.97	0.67
Horizontal Tables	21,863	0.95	0.50
Vertical Tables	7205	0.91	0.62

Our automated rule-based structural characterization of 120,000+ tables showing high variability amongst scientific tables

Scientific Tables

3. Diffuse context

- additional context needed to infer table (cell/column/row) semantics
- may be explicit but outside body of table, or implicit – based on other cells in row or column.

Table 2

Performance of serological assays in dependence of time after onset of symptoms.

		IgA						IgG				
		S1-assay		N-assay				S1-assay		N-assay		
	n	pos.	% ($CI_{95\%}$)	pos.	% ($CI_{95\%}$)	*p*	κ	pos.	% ($CI_{95\%}$)	pos.	% ($CI_{95\%}$)	*p*
$\text{Sensitivity}_{0\text{-}3\ d}$	16	5	31.2 (12.1–58.5)	2	12.5 (2.2–39.6)	n.s.	0.470	2	12.5 (2.2–39.6)	2	12.5 (2.2–39.6)	n.s.
$\text{Sensitivity}_{4\text{-}7\ d}$	23	12	52.2 (31.1–72.6)	7	30.4 (14.1–53.0)	n.s.		4	17.4 (5.7–39.5)	7	30.4 (14.1–53.0)	n.s.
$\text{Sensitivity}_{8\text{-}10\ d}$	24	16	66.7 (44.7–83.7)	9	37.5 (19.6–59.2)	0.016		11	45.8 (26.2–66.8)	14	58.3 (36.9–77.2)	n.s.
$\text{Sensitivity}_{11\text{-}13\ d}$	17	17	100 (0.77–100)	13	76.5 (49.8–92.2)	n.s.		13	76.5 (49.8–92.2)	15	88.2 (62.3–97.8)	n.s.
$\text{Sensitivity}_{\geq 14\ d}$	25	24	96.0 (77.7–99.8)	16	64.0 (42.6–81.2)	0.008		22	88.0 (67.6–96.8)	24	96.0 (77.7–99.8)	n.s.
$\text{Sensitivity}_{\text{outpat.}}$	65	63	96.9 (88.4–99.5)	4	6.2 (1.9–15.5)	<0.001	0.004	64	98.5 (90.6–99.9)	56	86.2 (74.8–93.1)	0.021
Specificity	139	8	94.3 (88.6–97.3)	0	100 (96.7–100)	<0.001	nd	1	99.3 (95.5–99.9)	0	100 (96.7–100)	n.s.

" … Seropositivity for IgA, IgG and IgM in 139 expected negative specimens and 170 specimens from 51 hospitalized and 65 outpatients with PCR-positive COVID-19 relative to days from onset of symptoms. Values for sensitivity and specificity are given as percentages with 95% Wilson-confidence intervals. McNemar's Test was used to compare diagnostic properties for two tests used on a single population and Fleiss' kappa was chosen as a measure of agreement. pos. = number of positive tested samples; n.d. = not determinable, n.s. = not significant. "

Scientific Tables

4. Lack of information reliability

- Not all tables can be treated the same. Some inherently more/less trustworthy

Description of LoM used to evaluate the efficacy of EIDD-2801 for SARS-CoV-2 pre-exposure prophylaxis and treatment.

EIDD-2801 pre-exposure prophylaxis	59	M	T33	13
	60	M	T33	13
	61	M	T33	13
	62	M	T33	13
Vehicle 24h treatment	63	F	C34	16
	64	F	C34	16
	65	F	C34	16
	66	F	C34	16
EIDD-2801 24h treatment	67	F	C34	16
	68	F	C34	16

***PMC7979515**: SARS-CoV-2 Infection is Effectively Treated and Prevented by EIDD-2801*

Table (2): Comparison of Laboratory data of Group I& Group II patients one week after starting treatment

Group / Variable	Group I after one week of treatment Mean ±SD	Group II after one week of treatment Mean ±SD	Independent t-test	P-value
Hgb(gm/Dl	14.2 ± 1.8	14.8 ± 2.7	1.85	0.07
TLC (X 103/ mL)	6.4±2.1	7.1±2.3	2.25	<0.05
Lymphocyte (%)	32.4 ± 6.8	28.2 ± 3.9	5.36	<0.001
CRP (mg/l)	4.8 ± 2.1	8.3 ± 3.6	8.4	<0.001
Serum ferritin (ng/ml)	94.8 ± 4	98.4 ± 54.8	0.49	0.62
D dimer (mg/l)	0.54 ± 0.06	0.68 ± 0.21	6.41	<0.001
RT-PCR(days)	5 ±1	10 ± 4	12.13	<0.001

***PPR230896**: Efficacy and Safety of Ivermectin for Treatment and Prophylaxis of COVID-19 Pandemic*

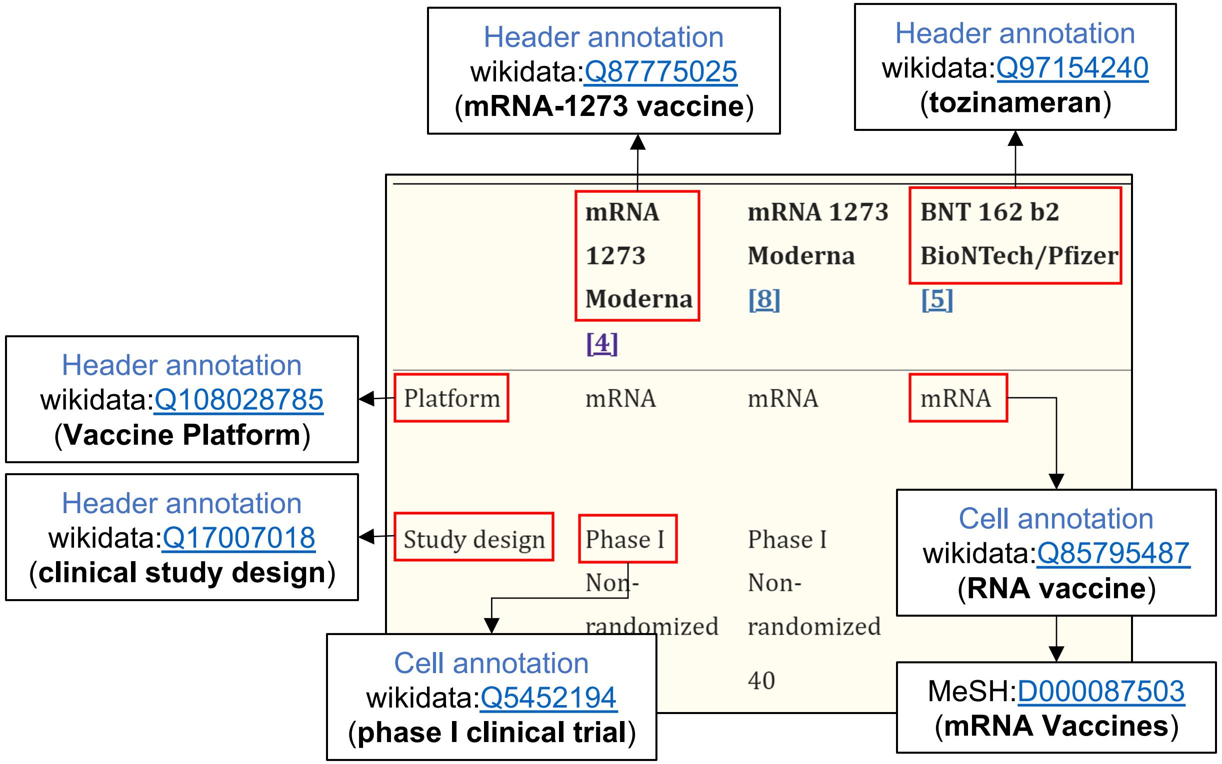
Reliability is a key factor in the discovery and integration of scientific tables (especially in this era of preprints and misinformation)

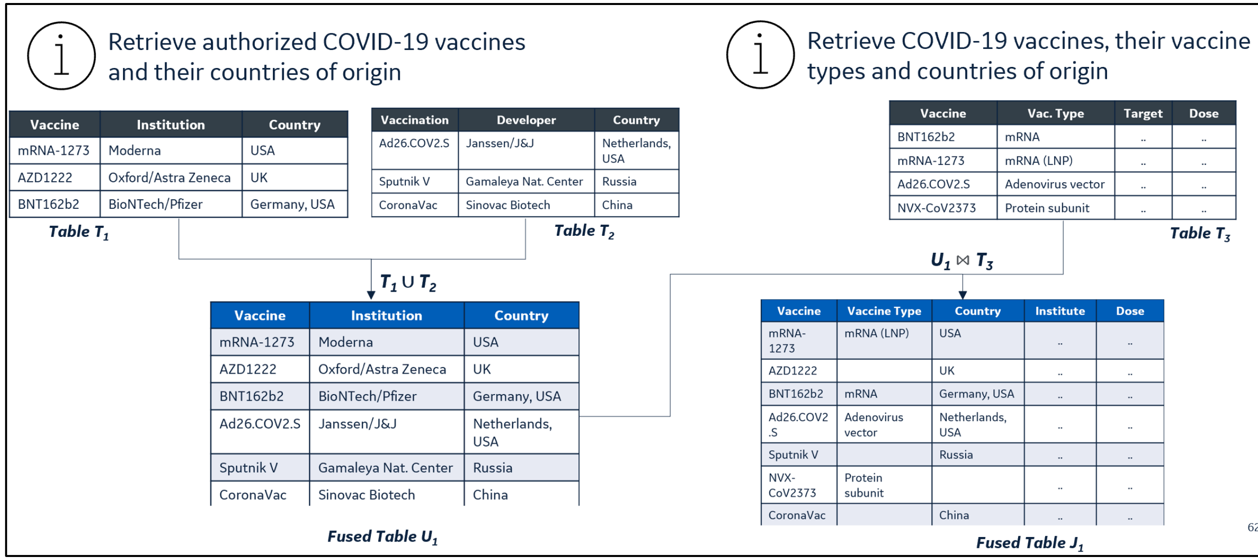
Research Goals

1. **Understand scientific tables**
   - infer the semantics of tables and their relevance to search queries
   - analyze scientific tables in the broader context of their structure and information reliability

2. **Enable discovery of relevant tabular data**
   - Systematically explore collections of scientific tables via rich semantic / contextual search
   - Discovery ➔ generate tabular response on the fly by fusing information from multiple tables

Focus of this paper

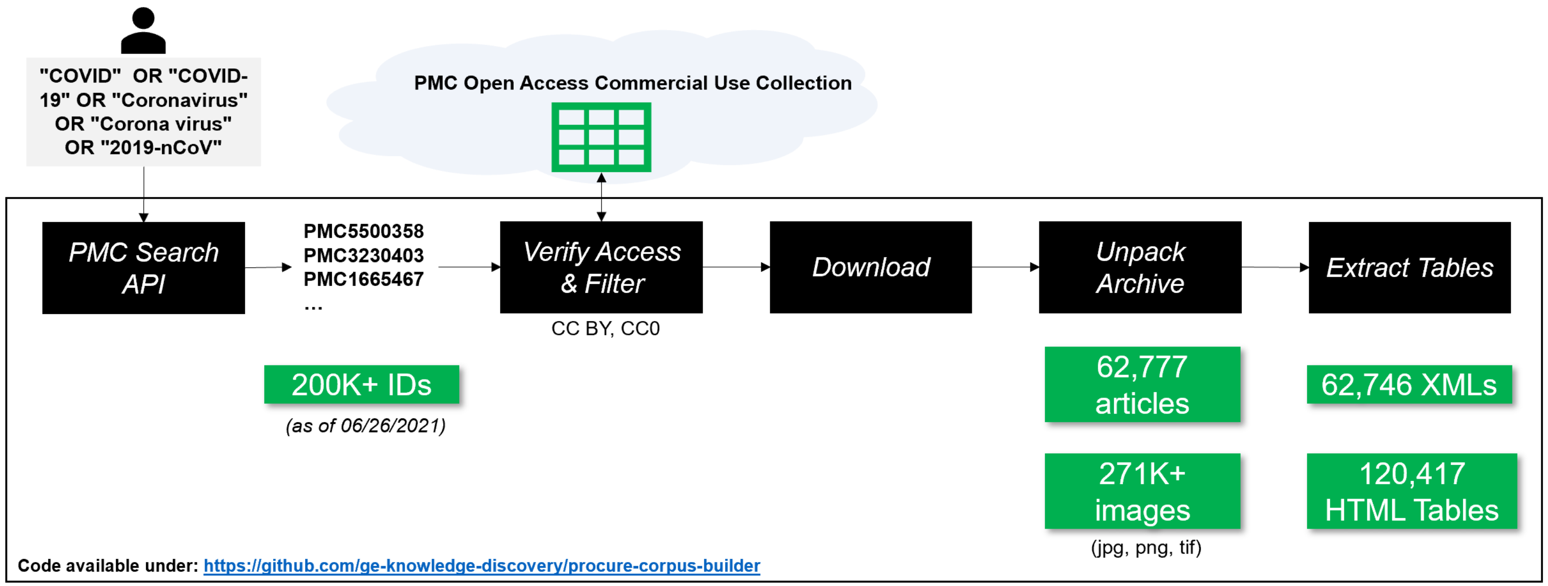




End-to-end Approach

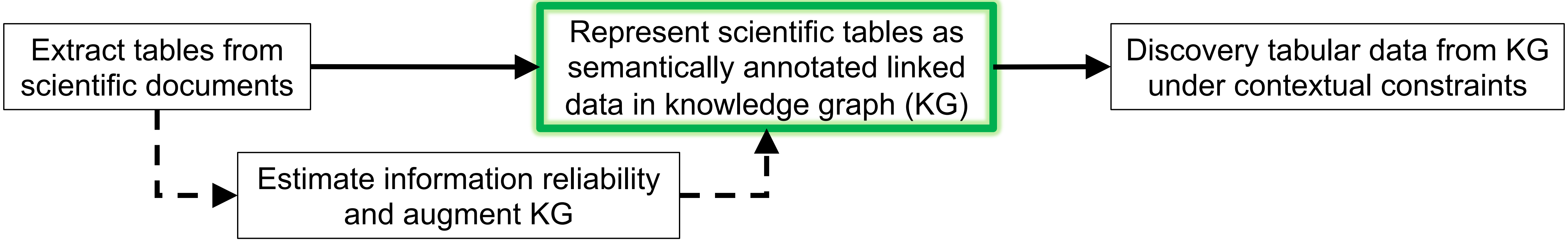
Extract tables from scientific documents → Represent scientific tables as semantically annotated linked data in knowledge graph (KG) → Discover tabular data from KG under contextual constraints

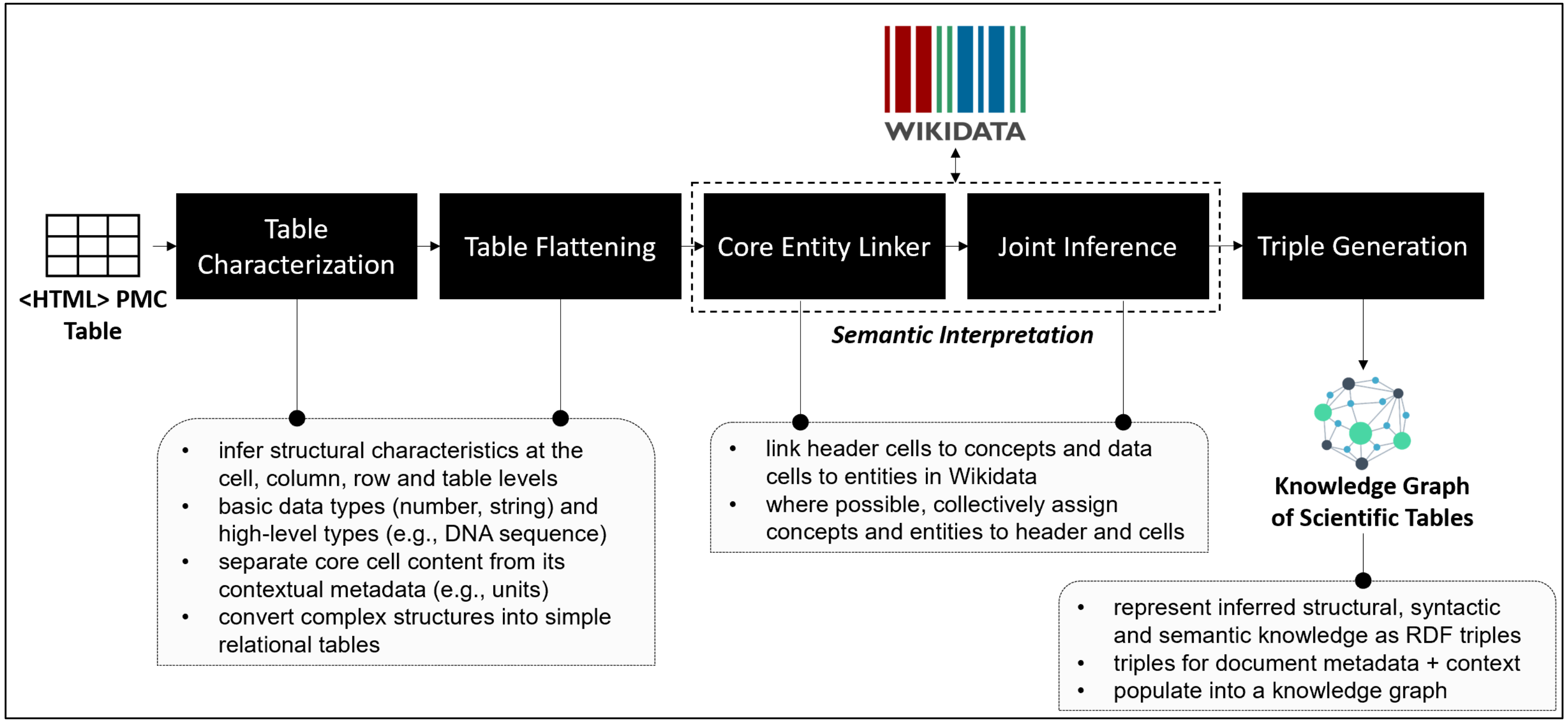
Estimate information reliability and augment KG



Pipeline to harvest tables and associated metadata on-demand from open-access articles hosted at PubMed Central (PMC)

End-to-end Approach





- *Mulwad et al.* "Towards Semantic Exploration of Tables in Scientific Documents". SemTech4STLD workshop at ESWC 2023
- *Mulwad et al.* "A Practical Entity Linking System for Tables in Scientific Literature". SDU workshop at AAAI 2023

Knowledge Graph of Scientific Tables

Table	Figure
Document	
W3C PROV*	
W3C OWL, RDF, RDFS, …	

Ontology for tabular data, metadata, inferred semantics

	mRNA 1273 Moderna [4]	mRNA 1273 Moderna [8]	BNT 162 b2 BioNTech/Pfizer [5]
Platform	mRNA	mRNA	mRNA
Study design	Phase I Non-randomized	Phase I Non-randomized	Phase I Randomized
Participants	45	40	195

<PMC8114590_Table_1> <type> <HtmlTable> .
<PMC8114590_Table_1> <caption> “*<caption><p>Main data from phases I/II clinical trials of the 4 vaccines available in Europe.</caption></p>*” .
<PMC8114590_Table_1> <numBodyRows> “9"^^xsd:integer .
<PMC8114590_Table_1> <numBodyCols> “6"^^xsd:integer .
<PMC8114590_Table_1> <numHeaderRows> "1"^^xsd:integer .
<PMC8114590_Table_1> <numHeaderCols> "1"^^xsd:integer .
<PMC8114590_Table_1> <identifier> “Table 1” .
<PMC8114590_Table_1> <mainClassification> “vertical" .

<PMC8114590> <type> <Document> .
<PMC8114590> <journal> “Therapies” .
<PMC8114590> <title> “*Efficacy of COVID-19 vaccines: From clinical trials to real life*” .
<PMC8114590> <publicationDate> “2021-05-12”^^xsd:date .
<PMC8114590> <table> <PMC8114590_Table_1> .

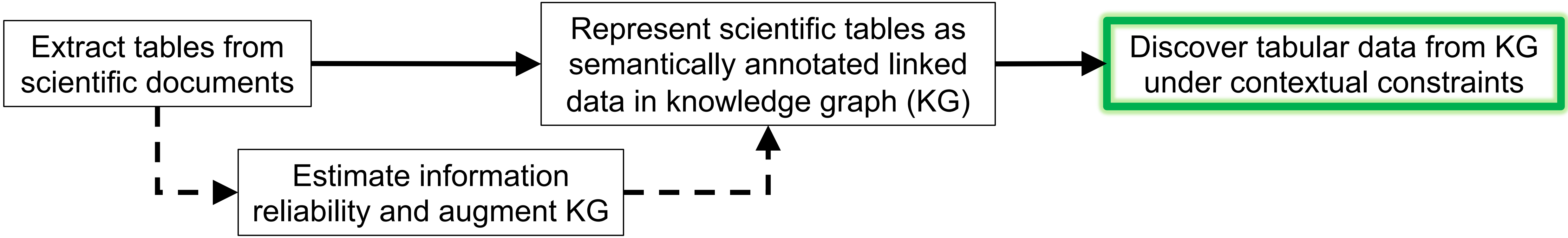
<PMC8114590> <docReliability> <PMC8114590_doc_reliability> .
<PMC8114590_doc_reliability> <provenanceScore> “0.75e0”^^xsd:double .
…

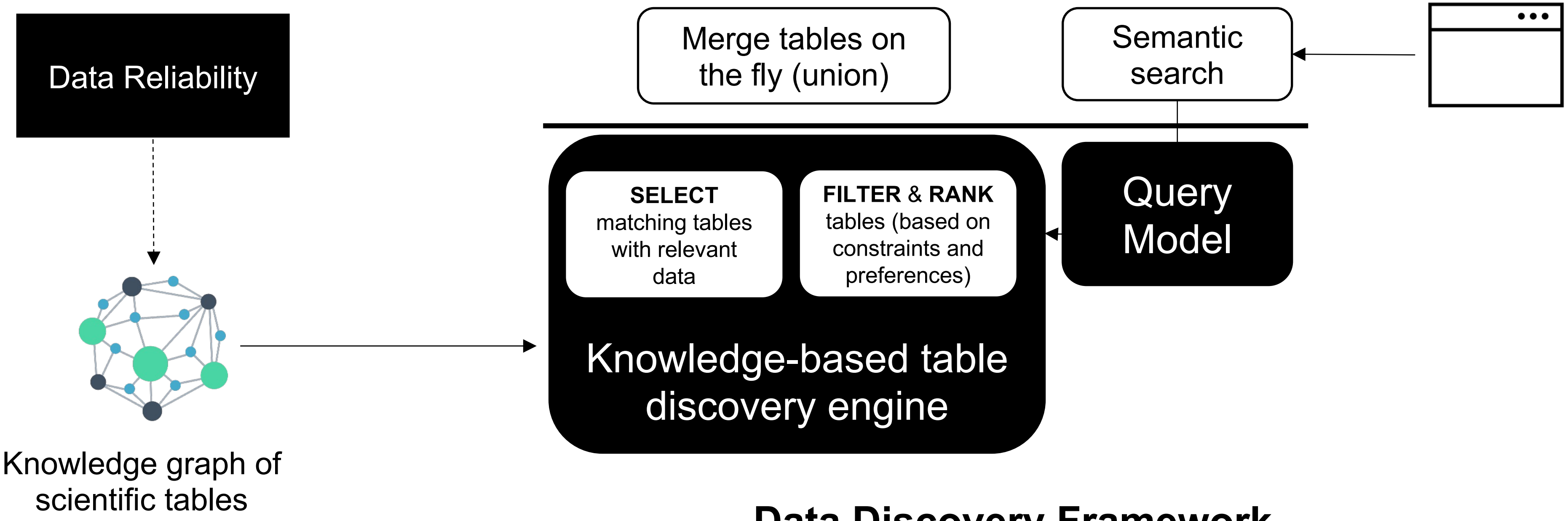
<PMC8114590_Table_1> <cell> <PMC8114590_Table_1_data_cell_1_1> .
<PMC8114590_Table_1_data_cell_1_1> <rowIndex> “1”xsd:integer .
<PMC8114590_Table_1_data_cell_1_1> <colIndex> “1”xsd:integer .
<PMC8114590_Table_1_data_cell_1_1> <type> <DataCell> .

<PMC8114590_Table_1_data_cell_1_1> <type> <PMC8114590_Table_1_data_cell_1_1_value> .
<PMC8114590_Table_1_data_cell_1_1_value> <type> <StringCellValue> .
<PMC8114590_Table_1_data_cell_1_1_value> <rawCellValue> “mRNA” .

<PMC8114590_Table_1_data_cell_1_1_value> <cellAnnotation>
<http://www.wikidata.org/entity/Q85795487> .

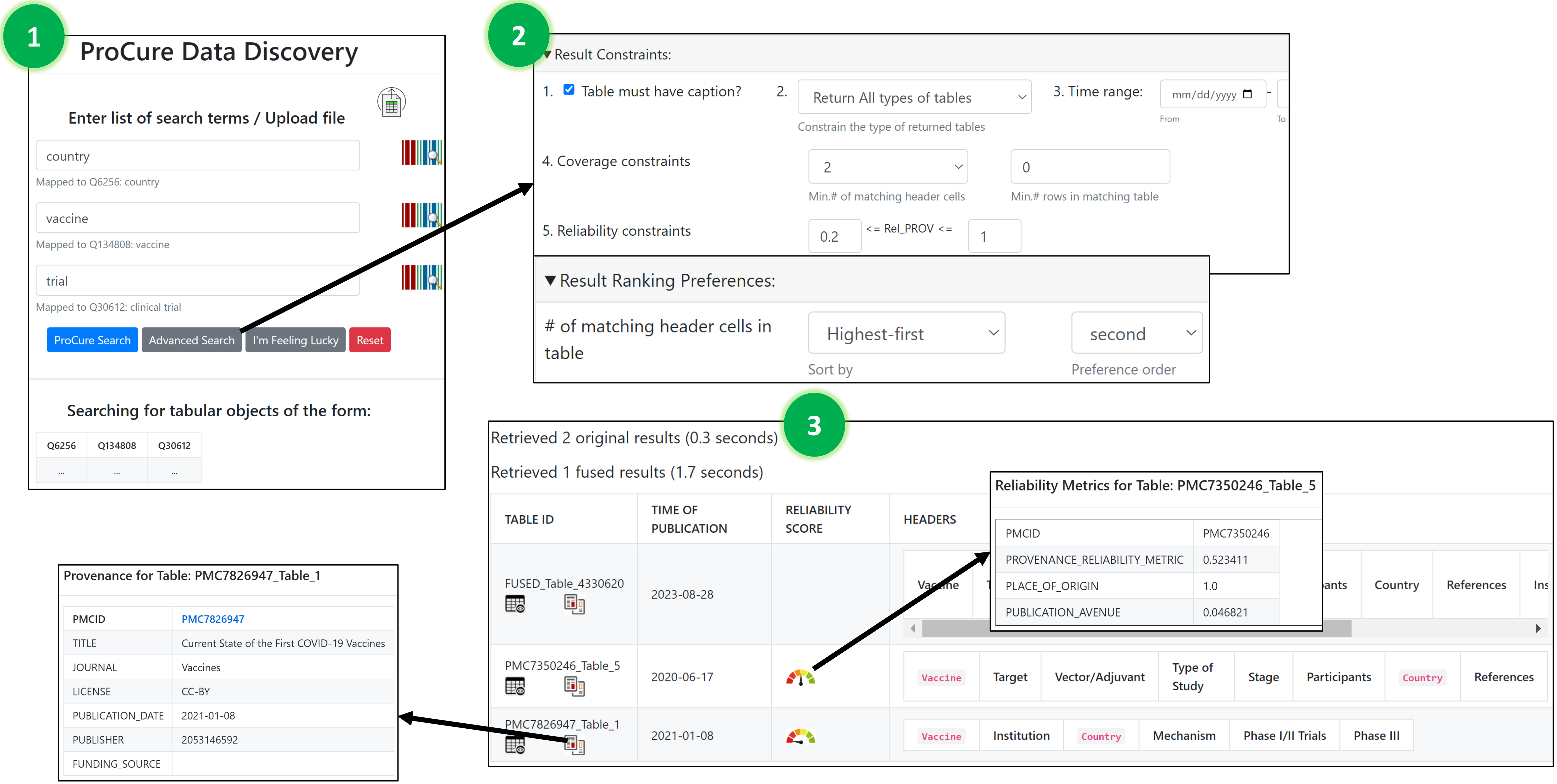
End-to-end Approach



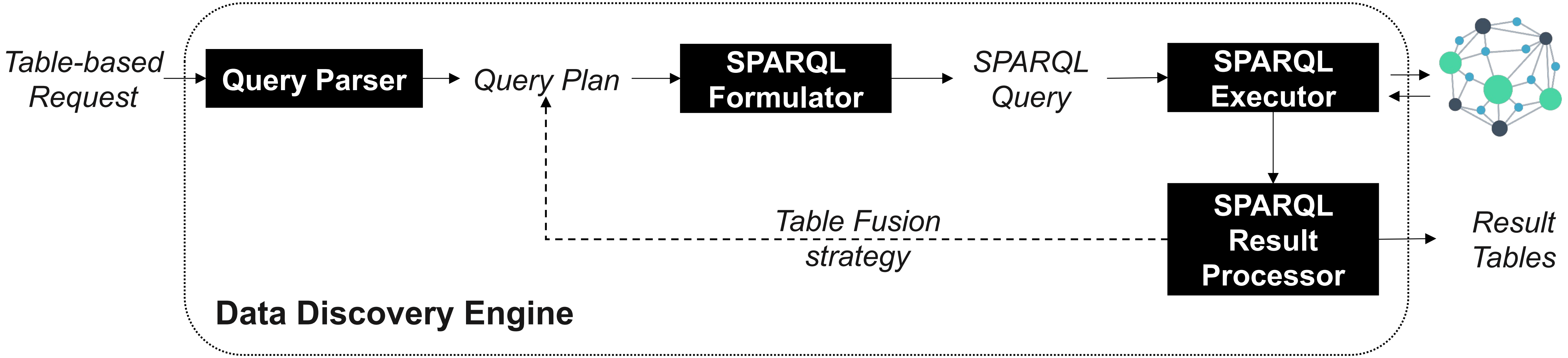


Data Discovery Framework

Table discovery prototype system



(Preliminary) Discovery Engine



A search request is composed of a set of logical primitives to SELECT matching tables, FILTER matching tables based on constraints, RANK filtered tables based on preferences, etc.

Implementation of each primitive driven by knowledge graph.
Discovery engine systematically compiles a query model into a SPARQL query

Implementation Details

```json
"conf": {
  "kg": {…
  },
  "result_prefs": {
      "return_captions": false,
      "return_footers": false,
      "return_headers": true,
      "return_reliability_scores": true,
      "return_time_info": true
  }
}
```

```sparql
SELECT DISTINCT (?table AS ?TABLE_ID) (?date AS ?TIME_OF_PUBLICATION) (?provenanceScore
AS ?RELIABILITY_SCORE) (GROUP_CONCAT(DISTINCT ?header;separator="|") AS ?HEADERS)
WHERE {
  ?document ns1:docReliability ?rel_uri .
  ?table_uri rdf:type ns1:HtmlTable .
  ?rel_uri ns1:provenanceScore ?provenanceScore_uri .
  ?table_uri ns1:cell ?cell_uri .
  ?cellValue_uri ns1:rawCellValue ?header .
  ?document ns1:publicationDate ?date_uri .
  ?cell_uri ns1:cellValue ?cellValue_uri .
  ?cell_uri rdf:type ns1:HeaderCell .
  ?document ns1:table ?table_uri .

  BIND (str(?provenanceScore_uri) AS ?provenanceScore)
  BIND (str(?date_uri) AS ?date)
}
GROUP BY ?table ?date ?provenanceScore
```

Bootstrapping SPARQL query with initial triple patterns and basic graph patterns based on return preferences in query model

Implementation Details

SELECT


```json
"terms": [
    {
        "entity_classes": [],
        "entity_embeddings": [],
        "entity_id": "Q6256",
        "entity_label": "country",
        "must_have": false,
        "present_in_table": true,
        "qualifiers": {},
        "source": "Wikidata",
        "string": "country"
    },
    {
        "entity_classes": [],
        "entity_embeddings": [],
        "entity_id": "Q134808",
        "entity_label": "vaccine",
        "must_have": false,
        "present_in_table": true,
        "qualifiers": {},
        "source": "Wikidata",
        "string": "vaccine"
    },
    {
        "entity_classes": [],
        "entity_embeddings": [],
        "entity_id": "Q30612",
        "entity_label": "clinical trial",
        "must_have": false,
        "present_in_table": true,
        "qualifiers": {},
        "source": "Wikidata",
        "string": "trial"
    }
]
```


```
SELECT DISTINCT ( … )
WHERE {
  …
  ?table_uri rdf:type ns1:HtmlTable .
  ?table_uri ns1:numBodyRows ?num_body_rows .
  ?table_uri ns1:numCols ?num_body_cols .
  {
    SELECT DISTINCT ?table1 (COUNT(*) AS ?coverage)
    WHERE     {
      ?cell1 rdf:type ns1:HeaderCell .
      ?table1 ns1:cell ?cell1 .
      OPTIONAL {
        ?cell1 ns1:cellAnnotation ?annotation .
      }
      BIND (strafter(str(?annotation), "http://www.wikidata.org/entity/") AS
?annotation_str)
      FILTER (?annotation_str IN ("Q6256","Q134808","Q30612"))
    }
    GROUP BY ?table1
    HAVING (?coverage >= 2)
    ORDER BY ASC(?coverage)
  }
  …
  FILTER (?date >= "0001-01-01")
  FILTER (EXISTS {?table_uri ns1:caption ?c})
  FILTER (?num_body_rows >= 0)
  FILTER (?table_uri = ?table1)
  FILTER (?date <= "2023-08-28")
  FILTER (?num_body_cols >= 0)
}
GROUP BY … ?coverage ?num_body_cols
```

FILTER


```json
"result_constraints": {
    "has_caption": true,
    "max_coverage": {
        "num_body_cols": 0,
        "num_body_rows": 0,
        "num_matching_body_cells_per_header": [],
        "num_matching_header_cells": 0,
        "percent_matching_body_cells_per_header": [],
        "percent_matching_header_cells": 0.0
    },
    "min_coverage": {
        "num_body_cols": 0,
        "num_body_rows": "0",
        "num_matching_body_cells_per_header": [],
        "num_matching_header_cells": "2",
        "percent_matching_body_cells_per_header": [],
        "percent_matching_header_cells": 0.0
    },
    "relations": {},
    "reliability": {
        "prov_hi": 1.0,
        "prov_lo": 0.0
    },
    "time_of_publication": {
        "after": "0001-01-01",
        "before": "2023-08-28",
        "past_year": false
    },
    "units": {}
}
```


Incrementally adding subqueries and clauses based on query semantics and other contextual constraints

Retrieved 2 original results (0.3 sec

Retrieved 1 fused results (1.7 secon

TABLE ID	TIME OF PUBLICATION
FUSED_Table_4330620	2023-08-28
PMC7350246_Table_5	2020-06-17
PMC7826947_Table_1	2021-01-08

Table: FUSED_Table_6019884

Number of rows: 16

Number of columns: 12

Number of cells: 192

Vaccine	Target	Vector/Adjuvant	Type of Study	Stage	Participants	Country	References	Institution	Mechanism
Viral vector based	S protein	Adenovirus vector	Randomized, double-blinded	Phase II	500	China	[89]		
Viral vector based (ChAdOx1 n-CoV-19)	S protein	Canine adenovirus vector	Randomized, single-blinded	Phase I/II	1112	UK	[90]		
DNA vaccine (INO-4800)	n.e.	Electroporation	Non-randomized	Phase I	40	USA	[91]		
Inactivated whole-virus	n.e.	n.e.	Randomized, double-blinded	Phase I/II	288 (I), 1168 (II)	China	[92]		
Inactivated whole-virus	n.e.	n.e.	Randomized, double-blinded	Phase I/II	744	China	[93]		
RNA vaccines (BNT162a1, BNT162b1 BNT162b2 and BNT162c2)	n.e.	n.e.	Non-randomized	Phase I/II	196	Germany	[94]		
LNP-encapsulated mRNA-vaccine (mRNA-1273)	S protein	Lipid nanoparticles	Non-randomized	Phase I	45	USA	[95]		
BNT162b1/ BNT162b2						Germany/US		BioNTech/ Pfizer	mRNA
mRNA-1273						US		Moderna	mRNA
AZD1222						UK		University Oxford/ Astra Zeneca	Adenovirus vector, chimpanzee
									Adenovirus

- On-the-fly fused table generation
- Union of rows based on semantic compatibility of 'Vaccine' and 'Country' columns
- Row deduplication (e.g., mRNA-1273) can leverage data semantics – not implemented
- Mechanism and Vector/Adjuvant columns are missed opportunity for merging into single column

Conclusions and Future Work

- Tables in scientific documents contain important information
  - Knowledge discovery from scientific tables is as vital as from text
  - Scientific tables bring additional challenges and opportunities
- Preliminary discovery system over knowledge graph of scientific tables
  - Foundational knowledge-guided discovery engine for selecting, filtering and ranking relevant matching tables
  - Feasibility of semantic search querying and generation of on-the-fly fused tables
  - Information reliability integrated into search and table fusion processes
- Discovery performance and experience can be enhanced in multiple ways:
  - Noisy nature of inferred semantics (e.g., incorrect or missing links) – can be addressed by leveraging other information such as raw string content and embeddings representations (of tables, rows, columns, cells).
  - Precision can also be improved by leveraging additional semantics (such as relationships between columns) once they are extracted
  - Along with header cell semantics, data cell semantics and additional context (such as units) can be used to disambiguate rows or columns during on-the-fly fusion

Acknowledgements

This research was based on work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via [2021-21022600004]. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government.