

A Practical Entity Linking System for Tables in Scientific Literature

Varish Mulwad¹, Tim Finin², Vijay S. Kumar¹,
Jenny Weisenberg Williams¹, Sharad Dixit¹, Anupam Joshi²

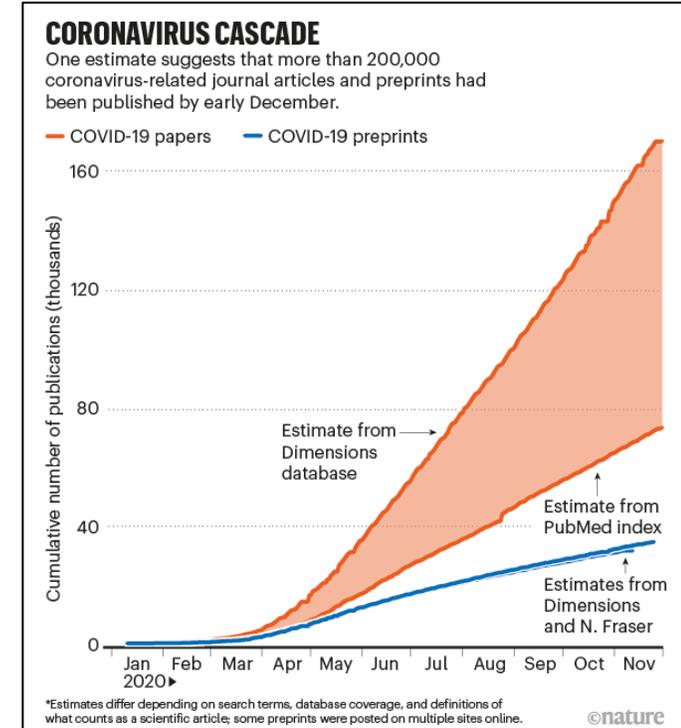
1: GE Research, Niskayuna, USA

2: UMBC, Baltimore, MD



Introduction

- An *infodemic* of scientific literature
 - COVID-19 saw an explosion in scientific publication rates
 - Causes: role of preprint servers, open access publishing
 - Problems: prevalence of misinformation, ‘fake science’
- Most research analyzing publications has focused only on text content
 - [CORD-19](#) dataset of articles on COVID-19 led to tools for search, Q&A, recommendation, summarization over scientific documents
- But significant information is also in tables and data charts
 - Our focus is on understanding information contained in **tables** drawn from scientific documents within specialized domains (e.g., biomedical)



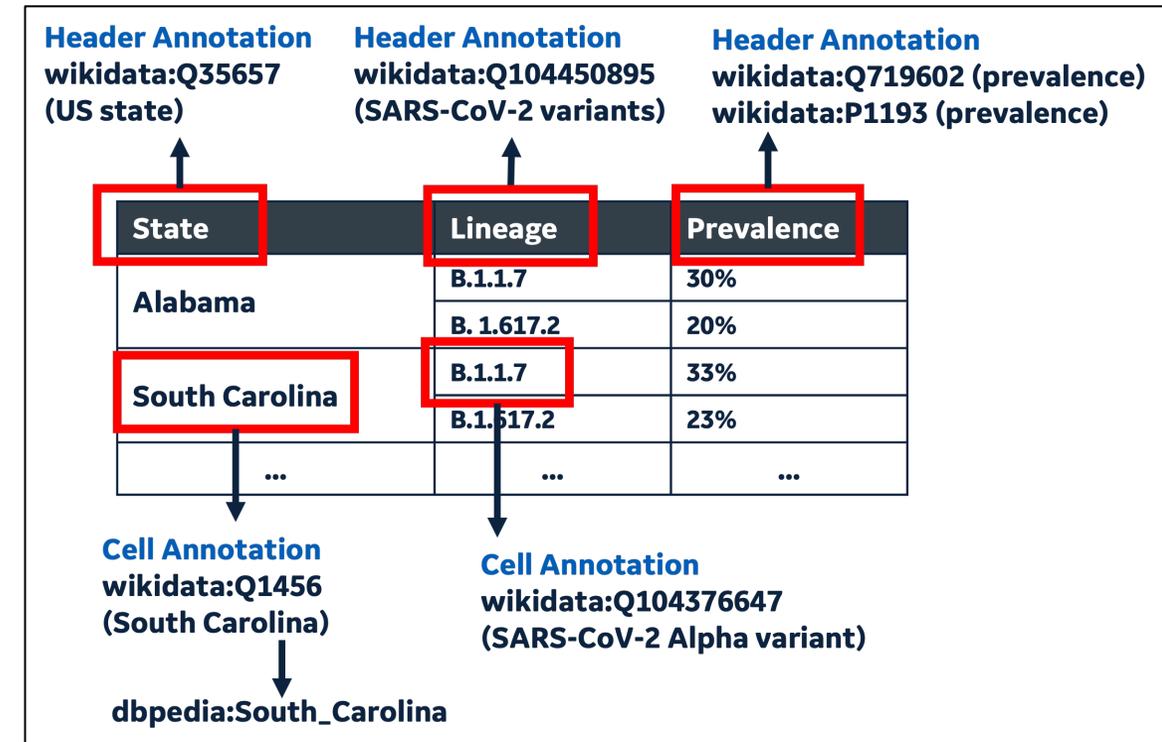
Else, H., 2020. How a Torrent of COVID Science changed Research Publishing—in seven charts. *Nature*, V588 (553).



Research Goal

- Information in scientific tables critical to new knowledge-driven applications
 - e.g., [genomic surveillance](#) to track spread of virus variants
- Need to search relevant tables for vital knowledge nuggets
 - Possibly fusing information from multiple tables on the fly
- Our goal: understand scientific tables to infer their semantics and relevance to search queries

Linking table cells to concepts in a background knowledge graph like Wikidata is an important step in understanding



“Treatment efficacies against the top prevalent COVID-19 variants in southern US states”

Tables in Scientific Documents

- Significant advances in pre-trained / representational models for **well-structured** tabular data, e.g. [TaBERT](#), [Sherlock](#), [TABBIE](#)
- But scientific tables bring additional **challenges** and **opportunities**
- For example, these three aspects

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 for IgM and 93.3% for IgG	Hangzhou Biotest Biotech Company, 2020

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
Elecsys® Anti-SARS-CoV-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
LIAISON® SARS-CoV-2 S1/S2 IgG Negative	5	157				

Less text ... more numbers ... sub columns

Table 2
Performance of serological assays in dependence of time after onset of symptoms.

	n	IgA				p	κ	IgG				p
		S1-assay		N-assay				S1-assay		N-assay		
		pos.	% (CI95%)	pos.	% (CI95%)			pos.	% (CI95%)	pos.	% (CI95%)	
Sensitivity _{0-3 d}	10	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.
Sensitivity _{4-7 d}	25	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.
Sensitivity _{8-10 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.
Sensitivity _{11-13 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.

Row and Column Headers

Tables in Scientific Documents

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 for IgM and 93.3% for IgG	(Hangzhou Biotest Biotech Company, 2020)

Similar to web tables ... with domain specific entities



Tables in Scientific Documents

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				

Less text ... more numbers ... sub columns



Tables in Scientific Documents

Table 2

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

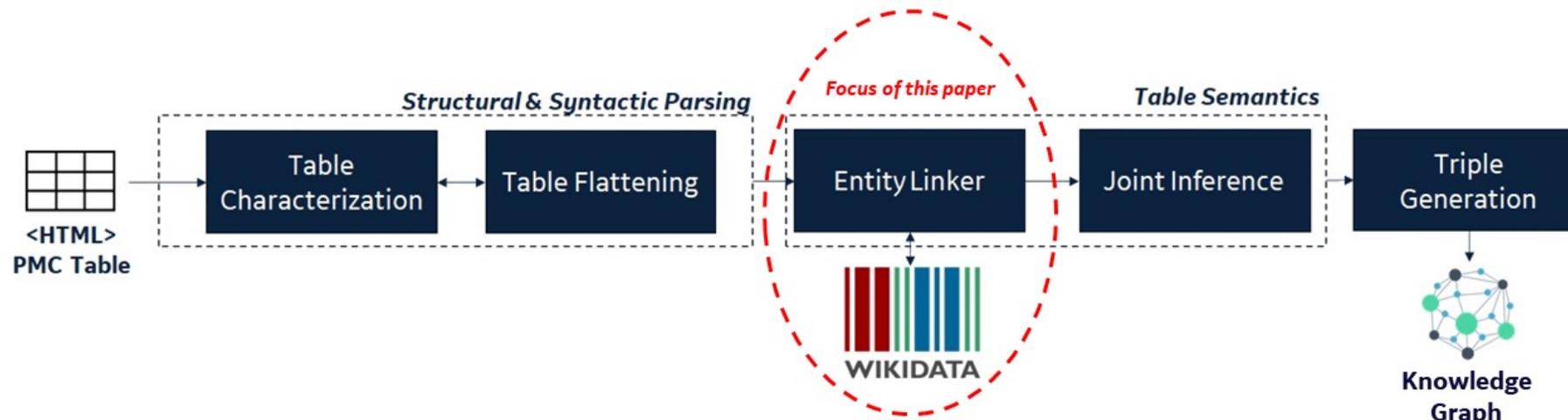
	n	IgA				p	κ	IgG				p
		SI-assay		N-assay				SI-assay		N-assay		
		pos.	%	pos.	%			pos.	%	pos.	%	
		(CI _{95%})		(CI _{95%})				(CI _{95%})		(CI _{95%})		
Sensitivity _{0-3 d}	16	5	31.2 (12.1–58.5)	2	12.5 (2.2–39.6)	n.s.	0.470	2	12.5 (2.2–39.6)	2	12.5 (2.2–39.6)	n.s.
Sensitivity _{4-7 d}	23	12	52.2 (31.1–72.6)	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.
Sensitivity _{8-10 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.
Sensitivity _{11-13 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.

Row and Column Headers



Approach

- Represent scientific tables as semantically annotated **linked data**
 - complementary approach accounting for structural heterogeneity & specialized terms
 - link with knowledge extracted from other information types (text, charts) and sources (e.g., document provenance)
 - leverage this additional context, structural characterization, and explicit reasoning over tabular content to infer table semantics
- We use a complex pipeline to construct a **knowledge graph** from tables in PubMed Central (PMC) documents
- This paper focuses on linking table header/body cells to Wikidata items



Linking entities to Wikidata's Knowledge Graph



- Recognizing and understanding entities and concepts in text, tables, and graphs essential
- System links text mentions to items in Wikidata's graph of ~100M entities
 - Including biomedical concepts like genes, diseases, drugs, symptoms and their names, aliases, and descriptions in multiple languages
- Supported by an ultra-fine-grained ontology of >2.5M types for entities, properties & constraints
- Items also linked to other knowledge graphs: DBpedia, Google Knowledge Graph, NCI Thesaurus, Medical Subject Headings thesaurus and more.

The screenshot shows the Wikidata page for 'pneumonia' (Q12192). The page includes a search bar, navigation tabs (Discussion, Read, View history, More), and a description: 'inflammatory condition of the lung'. Below the description is a table with columns for Language, Label, Description, and Also known as. The table lists translations for English, Spanish, Traditional Chinese, and Chinese. Below the table is a 'Statements' section with two rows: 'instance of' (cause of death) and 'subclass of' (lung disease). Each row has an 'edit' link and a reference count.

Language	Label	Description	Also known as
English	pneumonia	inflammatory condition of the lung	acute pneumonia lung inflammation inflammation of lung tissue inflammation of lungs Pneumonia
Spanish	neumonía	enfermedad del sistema respiratorio	pneumonía pulmonía pneumonia pulmonia inflamación del pulmón
Traditional Chinese	肺炎	肺部疾病	
Chinese	肺炎	肺部疾病	

Statements

instance of	cause of death	edit
	- 0 references	+ add reference
	infectious disease	edit
	1 reference	+ add value
subclass of	lung disease	edit



Linking entities to Wikidata's Knowledge Graph



The linked items can be

- Proper *entities* like **pneumonia** ([Q12192](#)), which is an instance of an **infectious disease** ([Q18123741](#))
- General *concepts*, like **hospitalization rate** ([Q107527870](#)) and **blood pressure** ([Q82642](#)) that may represent properties
- *Wikidata properties* like **symptoms and signs** ([P780](#)) that connect a medical condition with its possible symptoms

The screenshot shows the Wikidata page for COVID-19 (Q84263196). The page title is "COVID-19 (Q84263196)". Below the title, there is a description: "respiratory syndrome and infectious disease in humans, caused by SARS coronavirus 2". The page lists various symptoms and signs, each with a Wikidata ID and an edit button. The symptoms and signs listed are: cough (Q12192), fever (Q12192), respiratory failure (Q12192), headache (Q12192), myalgia (Q12192), fatigue (Q12192), hemoptysis (Q12192), diarrhea (Q12192), dyspnea (Q12192), lymphopenia (Q12192), and anemia (Q12192). The page also shows the number of references for each symptom and sign.

Symptoms and signs	Wikidata ID	References
cough	Q12192	2 references
fever	Q12192	2 references
respiratory failure	Q12192	1 reference
headache	Q12192	1 reference
myalgia	Q12192	1 reference
fatigue	Q12192	1 reference
hemoptysis	Q12192	1 reference
diarrhea	Q12192	2 references
dyspnea	Q12192	2 references
lymphopenia	Q12192	1 reference
anemia	Q12192	0 references

Linking noun chunks in table reference



Sentence referencing a table in an article on COVID-19:

Trials involving vaccines, antiviral drugs, immunotherapies, monoclonal antibodies, stem cells, and nitric oxide are summarized in **Table 1**.

Without additional adaptation or training for the medical domain, our linker finds good links for phrases using the whole sentence as context:

- **Trials** ('[Q1436668](#)', 'randomized controlled trial', 'experimental ...')
- **vaccines** ('[Q134808](#)', 'vaccine', 'a substance used to stimulate ...')
- **antiviral drugs** ('[Q40207875](#)', 'antiviral agent', 'substance that destroys ...')
- **immunotherapies** ('[Q1427096](#)', 'immunotherapy', 'therapy to elicit or ...')
- **monoclonal antibodies** ('[Q422248](#)', 'monoclonal antibody', 'monospecific ...')
- **stem cells** ('[Q48196](#)', 'stem cell', 'undifferentiated biological cells ...')
- **nitric oxide** ('[Q14916164](#)', 'nitric oxide biosynthetic process', 'The chemical



Online Linking Approach

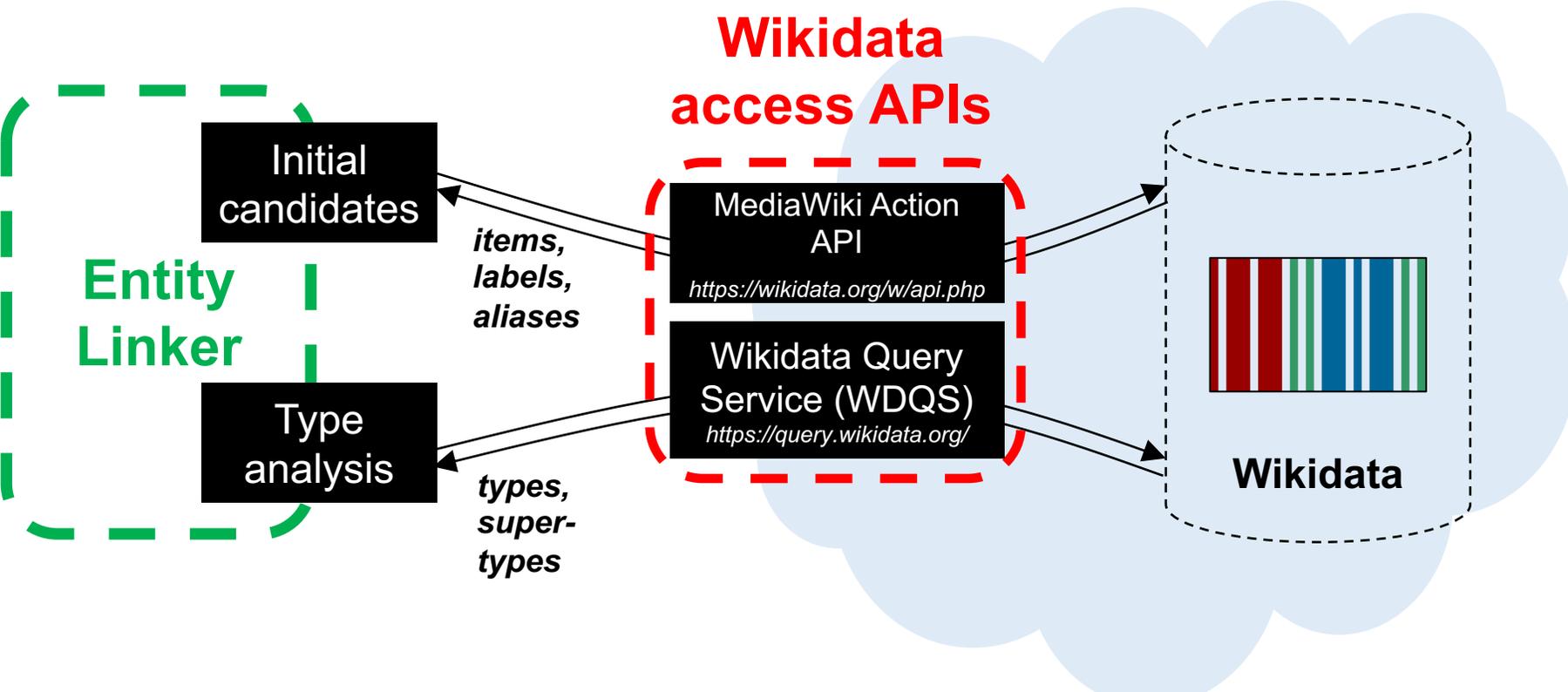


Wikidata's search APIs produce ranked candidates which we filter by **provenance, inherited types, and context**

- Use a mapping from [SpaCy](#) ontonotes types to Wikidata types, augmented with additional biomedical types
- *WD bad types* eliminate items unlikely to be good, e.g., fictional characters, scholarly papers, songs, musical groups, etc.
- The **SpaCy NLP** system makes type errors, so we employ a weighted mapping from target type (e.g., LOC) to *near misses* (e.g., FAC, GPE), and fallback *ok types* (e.g., ENTITY)
- **Embeddings** are used to compare a candidate link's short description with a mention sentence



Efficient Entity Linking at Large Scale



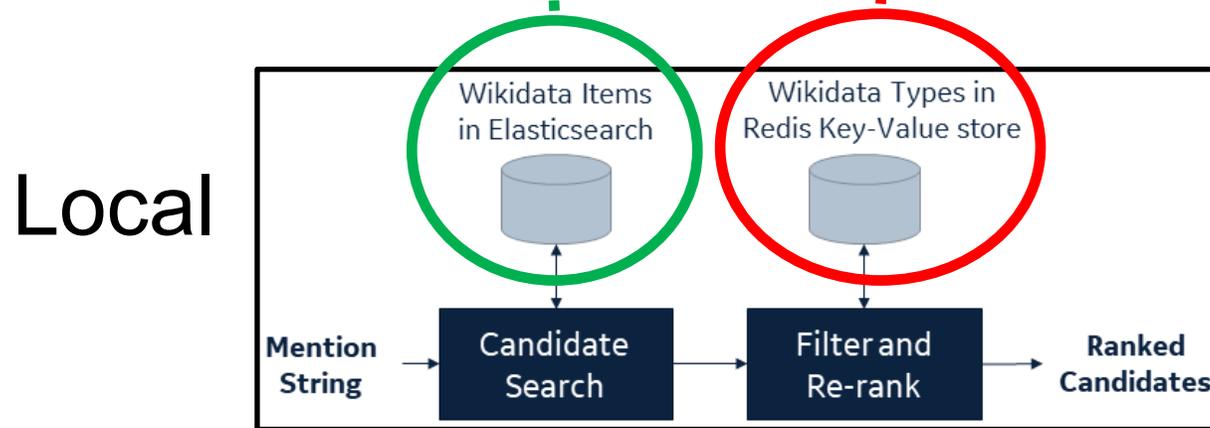
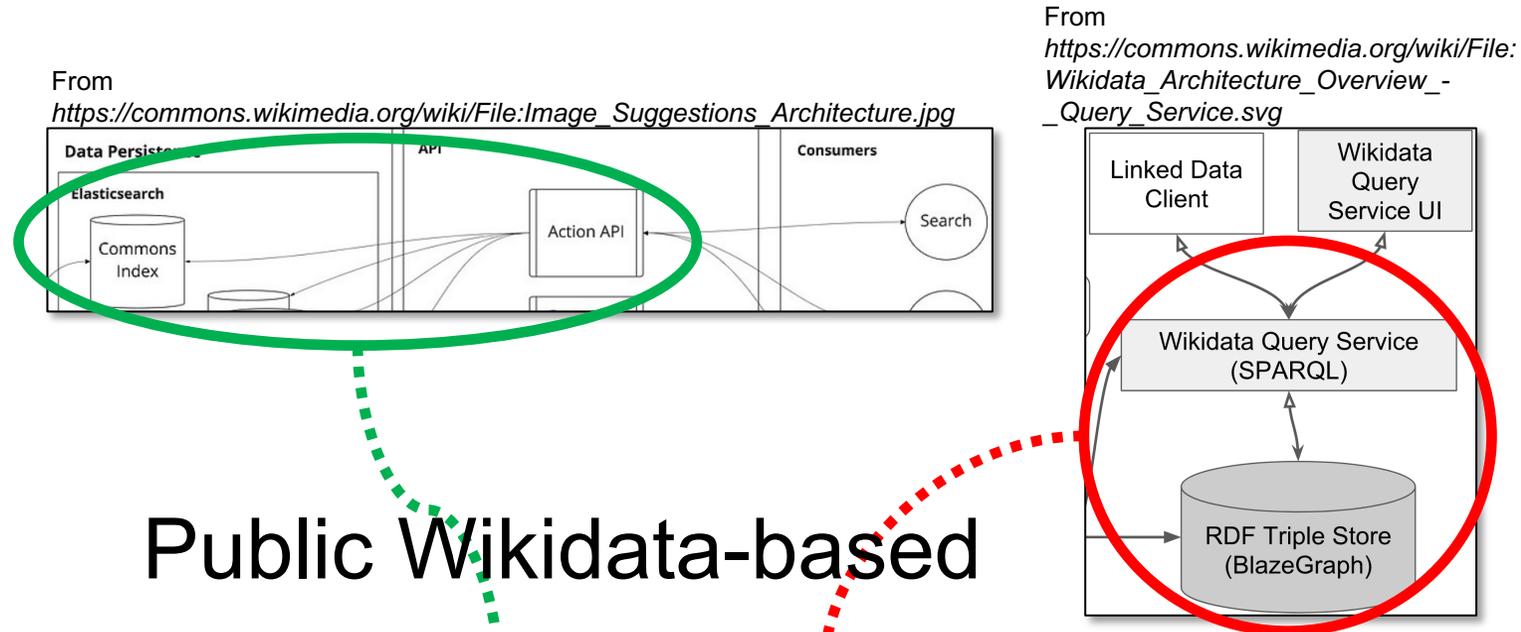
- Rate-limited public Wikidata access APIs a major scaling bottleneck in our entity linker
- Cannot sustain modern scientific publication rates

Average time taken by Entity Linker for mention string: "coronavirus"	
Retrieve initial ranked list of candidate items via Action API (top 20-50 candidates)	12 sec.
Retrieve all types and supertypes per candidate via WDQS . Filter & rank candidates based on type analysis (good/OK/bad) for domain	18 sec.



Efficient Entity Linking at Large Scale

- Avoiding API invocations via caching results helped, but not enough
- So, we built an “offline” linker to eliminate all reliance on Wikidata APIs
- We replicated *just enough* Wikidata functionality locally



Functional Architecture of local “offline” Entity Linker



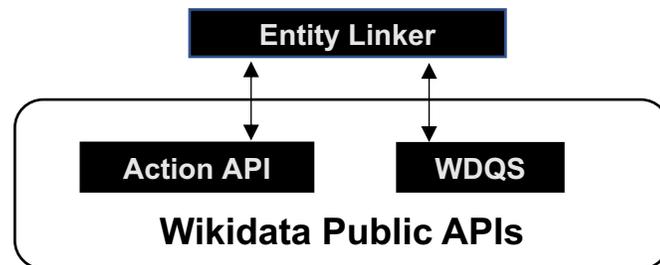
Efficient Entity Linking at Large Scale

Phased approach to transition to an offline entity linker:

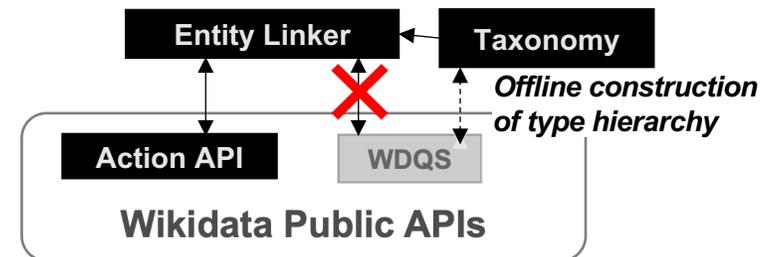
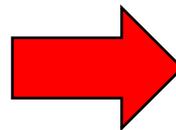
1

Offline, we use WDQS to retrieve current **taxonomy** of Wikidata types and their supertypes.

Resulting dictionary mapping each of **2.6M types** to their supertypes is loaded into **Redis**, an efficient key-value store.



Retrieve types via WDQS + type analysis: **~18 sec.**



Retrieve types via Redis lookup + type analysis: **~9.5 sec.**

- Entity linker result remains unaffected
- Local taxonomy can be updated at some predefined frequency (e.g., daily)



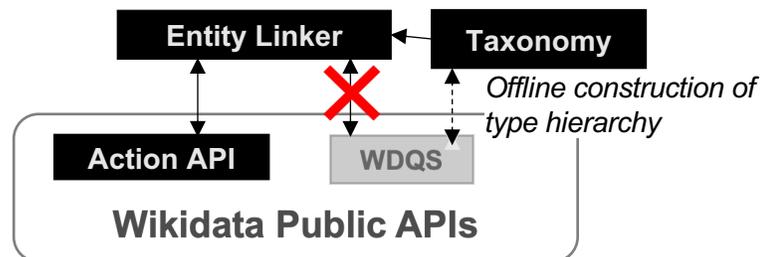
Efficient Entity Linking at Large Scale

Phased approach to transition to an offline entity linker:

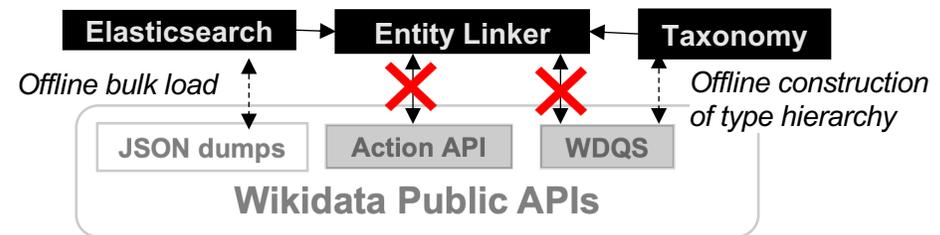
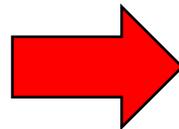
2

Offline, we upload partial (JSON) dump of Wikidata entities into local [Elasticsearch](#), an efficient, scalable search engine

Resulting searchable index contained **95.8M items**, their attributes (label, aliases, description) and sitelinks counts



Get top 20 candidate items via Action API: **~12 sec.**



Get top 20 candidate items from Elasticsearch: **~0.08 sec.**

- sitelinks count used to rank items. Initial candidate list has impact on linker quality.
- Trade-off: Tolerable reduction in quality for 3x faster linking.



Inferring Semantics of Tables

To improve overall entity linking performance, we first structurally & syntactically characterize scientific tables using two approaches

The diagram illustrates several table structures and their semantic characterizations:

- Horizontal Table:** A table with columns for Haplotype, Marker, and Species. Example: H1 | 594^A | 718 | 797 | 176 | 961 | T. urartu (30,30.30%)² | T. caucasicum (37,38.92%)² | T. caucasicum (117,54.42%)²
- Header Row:** A row at the top of a table, e.g., Haplotype, Marker, Species.
- Header Column:** A column at the top of a table, e.g., Haplotype.
- Vertical Table:** A table with rows for Gender (%), Age at inclusion, Ethnicity (%), and Caucasian.
- Matrix Table:** A table with rows for Virus category (PST, LLV-A, LLV-B, HELV) and columns for n, %.
- Simple Header:** A table with columns for Developer, Platform, Target, Target, Other features, and References.
- Concise Header:** A table with columns for Haplotype, Marker, and Species.
- Multilevel Header:** A table with columns for Variables and rows for 1k², 3, 4, 5, 6, 7, 9, Missing, N.
- Splitted (Header):** A table with columns for Sample type, Method, No. of positive sample, No. of positive sample by method, Positivity rate (%), and Positivity rate (%).
- Multi-dimensional Table:** A table with columns for IQ test rates, No IQ test rates, Combined rates, and Beta by years of age.
- Basic type: string:** IGD9-SF (-)
- Basic type: number:** 60,418.0
- Basic type: number with tolerance:** -0.0051 ± 0.0024
- Basic type: number with range:** 1.28 (0.30-5.43)

Rule-based characterizations to categorize tables into types based on their structure

Semantic Type REFERENCE

Study	Location	Total cases
CDC (12)	United States	149,760
Livingston and Bucher (18)	Italy	22,512
Tagarro et al. (19)	Spain (Madrid)	4,695

Subset of PMC7347905 Table 1

HTML `<xref>` tag containing attribute `ref-type="bibr"`

Semantic Type DNA/RNA SEQUENCE

^a AA	Codon	^b RSCU	AA	Codon	RSCU
Ala	GCA	1.25	Leu	CUA	0.22
	GCC	0.59		CUC	0.20
	GCG	0.02		CUG	0.64
	GCU	2.12		CUU	1.12

Subset of PMC3087699 Table 2

Regex based: 3+ characters from the set {G, A, T, C, U}

Semantic Type CLINICAL TRIAL ID

Virus	Location	Phase	Year	Identifier
SARS-CoV	United States	I	2004	NCT00099463
SARS-CoV	United States	I	2007	NCT00533741
SARS-CoV	United States	I	2011	NCT01376765
MERS	United Kingdom	I	2018	NCT03399578
MERS	Germany	I	2018	NCT03615911

Subset of PMC7239068 Table 1

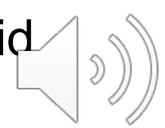
Regex based: NCT followed by 8 digits

Basic Type QUANTITY

Variable studied	Median	Standard deviation
IL-6 (normal range 0.0-15.5 pg/ml) [†]	65 pg/ml	446.6 pg/ml
D-dimer (normal range 0.20-0.28 µg/ml)	2.1 µg/ml	4.7 µg/ml
Ferritin (normal range 20-450 ng/ml)	446 ng/ml	1822 ng/ml
CRP (normal range 0.0-9.0 mg/l)	90 mg/l	132.9 mg/l

subset of PMC7493720 Table 1

Specialists (pattern-based or ML-based) assess commonly encoded data types in cells to avoid linking those with specific kinds of literals



Inferring Table Semantics: Rules



Haplotype	Marker					Species
	<i>Xrj5</i>	<i>Xid3</i>	<i>Xrj6</i>	<i>Xid4</i>	<i>Xrj7</i>	
H1	594 ^a	718	797	176	961	<i>T. urartu</i> (30, 30.30%) ^c <i>T. turgidum</i> (37, 38.95%) <i>T. aestivum</i> (117, 54.42%)
H2	594	808	797	176	961	<i>T. urartu</i> (2, 2.02%)
H3	594	655	797	176	961	<i>T. urartu</i> (1, 1.01%)

Horizontal Table

	Viremia category				p value
	PSV n = 716 (70.6%)	LLV-I n = 46 (4.6%)	LLV-II n = 52 (5.1%)	HLV n = 201 (20%)	
Gender [n (%)]					0.011
Male	468 (65%)	39 (85%)	36 (69%)	148 (74%)	
Female	248 (35%)	7 (15%)	16 (31%)	53 (23%)	
Age at inclusion [median years (IQR)]	39 (33–46)	43 (35–50)	39 (34–49)	39 (33–47)	0.28
Ethnicity [n (%)]					0.0027
Caucasian	307 (43%)	25 (54%)	29 (56%)	96 (48%)	
African	172 (24%)	12 (26%)	12 (23%)	68 (34%)	
Asian	81 (11%)	2 (4.3%)	3 (3.8%)	9 (4.5%)	
Other/unknown	156 (22%)	7 (15%)	8 (15%)	28 (14%)	

Vertical Table

From:	Employment ^a	Sick leave	Vocational rehabilitation	Medical rehabilitation	Time limited disability pension	Disability pension	Emigrated	Dead (6)	Alive and at risk ^b
Employment ^a (1)	0	1 657 895	45 865	48 607	3 736	20 314	28 421	10 860	729 891
Sick leave (2)	1 595 592	133 909	53 398	48 525	559	3 309	454	726	105 714
Vocational rehabilitation (3)	101 605	983	2 009	962	3 816	9 351	175	85	288
Medical rehabilitation (4)	82 070	130	0	14 381	1 320	5 164	93	81	10 566
Time limited disability benefits (5)	1 153	68	148	84	0	7 930	7	23	191
Disability benefits (6)	7 729	307	75	26	37	0	1 404	819	45 265
Emigrated (7)	26 487	317	32	51	0	46	53	13	5
Total	1 814 636	1 793 609	101 527	112 436	9 468	46 114	30 607	12 607	891 920

Matrix Table

Developer	Platform	Target antigen	Target antibody	Other features	References

Simple Header

IQ test taken	No IQ test taken			Combined score				
	Stanine Score	N	Percent	N	Percent	Score		
1	15 709	1,9%	C (assumed below average)	3 769	14,9%	Assumed below average ^a	57 092	6,6%
2	37 614	4,5%	B (assumed average)	20 318	80,6%	Assumed average ^b	691 580	80,1%
3	75 441	9,0%	A (assumed above average)	1 124	4,5%	Assumed above average ^c	114 697	13,3%

Splitted (Header)

Dead		Before 35 years of age		After 35 years of age	
		HRR (95% CI)	p	HRR (95% CI)	P
No mental problems	High IQ	0.60 (0.53–0.69)	<0.001	0.71 (0.62–0.80)	<0.001
	Average IQ	1	ref	1	Ref
Mental problems	Low IQ	1.72 (1.57–1.89)	<0.001	2.11 (1.93–2.31)	<0.001
	High IQ	1.05 (0.56–1.95)	0.89	1.27 (0.72–2.26)	0.41
	Average IQ	1.94 (1.72–2.20)	<0.001	2.21 (1.97–2.49)	<0.001
	Low IQ	2.60 (2.21–3.06)	<0.001	2.99 (2.56–3.49)	<0.001

Multi-dimensional Table

Haplotype	Marker					Species
	<i>Xrj5</i>	<i>Xid3</i>	<i>Xrj6</i>	<i>Xid4</i>	<i>Xrj7</i>	

Concise Header

Sample type (n)	Method	No. of positive sample	No. of positive sample by any method	Positivity rate ^a [% (95% CI)]	No. of positive patients by any methods in any sample types	Positivity rate ^b [% (95% CI)]
OPS (n=68)	qRT-PCR.A	12	24	50.0 (29.6–70.3)	56 ^c	21.4 (12.0–34.8)
	qRT-PCR.B	18		75.0 (52.9–89.4)		32.1 (20.6–46.1)
	RT-RAA	20		83.3 (61.8–94.5)		35.7 (23.7–49.7)

Concise Body

IGDS9-SF (-)

60,418.0

- 0.051 ± 0.024^{*}

1.28 (0.30–5.43)

...

Basic type: string
Basic type: number
Basic type: number with tolerance
Basic type: number with range

Variables	1&2 ^a	3	4	5	6	7	9	Missing	N
	n	n	n	n	n	n	n	n	N

Multilevel Header

Rule-based characterizations classify tables into types based on their structure



Inferring Table Semantics: ML



Semantic Type REFERENCE

Study	Location	Total cases
CDC (17)	United States	149,760
Livingston and Bucher (18)	Italy	22,512
Tagarro et al. (19)	Spain (Madrid)	4,6,95

Subset of PMC7347905 Table 1

HTML `<xref>` tag containing attribute `ref-type="bibr"`

Semantic Type DNA/RNA SEQUENCE

^a AA	Codon	^b RSCU	AA	Codon	RSCU
Ala	GCA	1.25	Leu	CUA	0.22
	GCC	0.59		CUC	0.20
	GCG	0.02		CUG	0.64
	GCU	2.12		CUU	1.12

Subset of PMC3087699 Table 2

Regex based: 3+ characters from the set {G, A, T, C, U}

Semantic Type CLINICAL TRIAL ID

Virus	Location	Phase	Year	Identifier
SARS-CoV	United States	I	2004	NCT00099463
SARS-CoV	United States	I	2007	NCT00533741
SARS-CoV	United States	I	2011	NCT01376765
MERS	United Kingdom	I	2018	NCT03399578
MERS	Germany	I	2018	NCT03615911

Subset of PMC7239068 Table 1

Regex based: NCT followed by 8 digits

Basic Type QUANTITY

Variable studied	Median	Standard deviation
IL-6 (normal range 0.0–15.5 pg/ml) [†]	65 pg/ml	446.6 pg/ml
D-dimer (normal range 0.20–0.28 µg/ml)	2.1 µg/ml	4.7 µg/ml
Ferritin (normal range 20–450 ng/ml)	446 ng/ml	1822 ng/ml
CRP (normal range 0.0–9.0 mg/l)	90 mg/l	132.9 mg/l

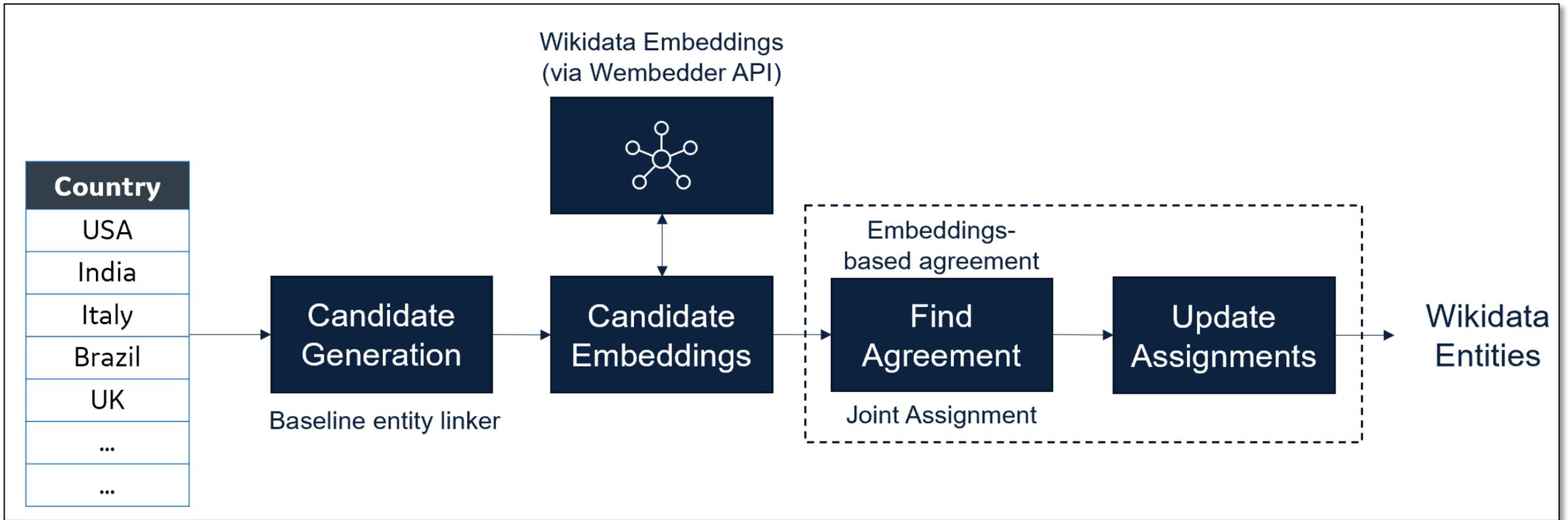
subset of PMC7493720 Table 1

Pattern-based and ML-based specialists assess commonly encoded data types in cells to avoid linking those with specific kinds of literals



Inferring Semantics of Tables

Besides mapping table cells to Wikidata items, inferring semantics of scientific tables is improved using **joint inference** and **embeddings**

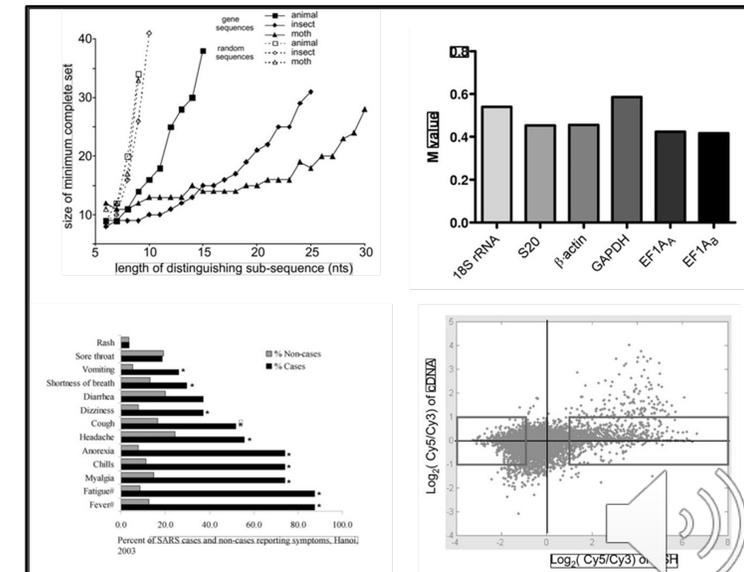
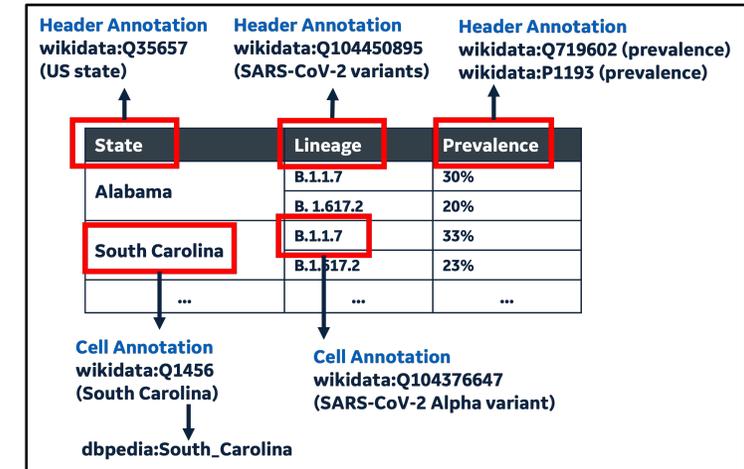


Joint inference using **embeddings** of Wikidata items and embeddings-driven agreement function to compute compatibility between entities and assign entities to cells in a column



Conclusion and Future Work

- **Data Tables** in scientific documents have important information
- Understanding their semantics requires **linking** their elements to concepts and entities in a background knowledge base, like **Wikidata**
- We implemented an initial online system using Wikidata's APIs and a much custom more **efficient offline version**
- Our **future work** will extend and integrate a similar system for **extracting and representing data** found in **data chart images**



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.