Building a PubMed knowledge graph

Simone Colli
simone.colli@studenti.unipr.it

Master's Degree in Computer Science

29-05-2025

Introduction

PubMed is an **essential resource** for the **medical domain**.

Useful **concepts** are **ambiguous** or **difficult** to extract.

Medical experts communicate through **highly specialized languages**, that capture critical biomedical knowledge but remain challenging for automated extraction.

Many studies have been devoted to **building open-access datasets** that solve **bio-entity recognition problems**.

These datasets are **predominatly about bio-entity recognition**, but **researchers have also been interested** in **extracting other types of entities and relationships** such as informations about the authors.

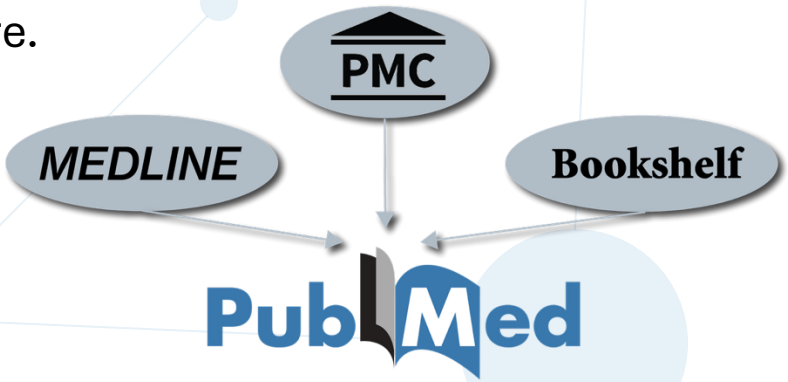
PubMed knowledge graph (PKG) is proposed as **solution**.

PubMed

PubMed is the National Library of Medicine's® (NLM) free, searchable **bibliographic database supporting scientific and medical research**. It contains more than 37 million citations and abstracts of biomedical and life sciences literature.

The **citation information** featured in PubMed is derived from **three main sources**:

- **MEDLINE**, is the primary database of citations and abstracts that PubMed searches.
- **PubMed Central**, is a full text archive that includes articles from journals reviewed and selected by NLM for archiving.
- **Bookshelf**, is a full text archive of books, chapters, reports, databases, and other documents related to biomedical, health, and life sciences.



Knowledge graph

A **knowledge graph represents** a **network of real-world entities** (such as objects, events, situations or concepts), and **illustrates the relationship** between them.

This information is usually stored in a **graph database** and visualized as a graph structure, prompting the term knowledge “graph.”

NIH ExPORTER

National institutes of health (NIH) is the **primary government agency** in the United States responsible for **biomedical and public health research**.

It offers **free databases and resources** such as PubMed and NIH ExPORTER.

Export expenditures and results tool (NIH ExPORTER) is an online database that **contains detailed information about research projects** funded by the NIH and other major U.S. government health agencies.

The database includes data such project titles and descriptions, names and affiliations of principal investigators, funding amounts and duration, etc..

Generally NIH ExPORTER is **widely used to track funding patterns**, **evaluate research productivity**, and **analyze scientific collaboration**.

ORCID

Open researcher and contributor ID (ORCID) is an international, non-profit organization that provides a **persistent digital identifier** (ORCID ID).

ORCID ID allows:

- **Disambiguation of researchers** with similar or identical names.
- **Reliable tracking of individual contributions** to research outputs, such as publications, grants, datasets, and peer reviews.
- **Collection and management** of researchers' affiliation history, educational background, and other professional activities.

MapAffil dataset

MapAffil dataset is a dataset that links **author affiliation strings** from PubMed articles to **geographic locations** (such as cities, countries, and geocodes). It provides:

- **Fine-grained geographic** data for author affiliations (e.g., city, state, country).
- A way to **resolve and standardize affiliation strings** that are often written in inconsistent formats.
- **Author-to-location mapping** at the article level, which is especially useful for analyzing research output by region or institution.

NLP, NER & BERT

Natural language processing (NLP) is a **subfildof AI**, concerned with the interaction between computers and **human language**.

Named entity recognition (NER) is an NLP task that focuses on the identification and **classification of named entities** (objects, numbers, organizations, dates, etc..) in **unstructured text**.

Bidirectional encoder representation from transformers (BERT), is a state-of-the-art deep learning model for **language understanding introduced** by Google in 2018.

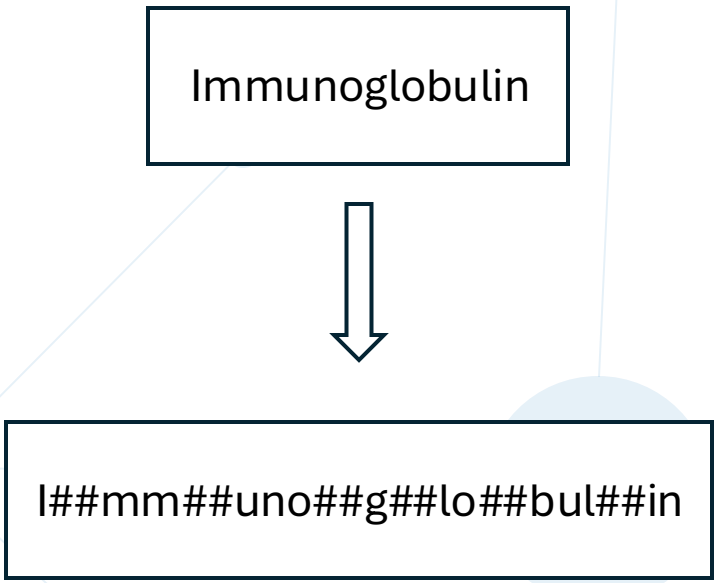
The BRCA1 gene has been extensively studied in relation to breast cancer.

The BRCA1 gene has been extensively studied in relation to breast cancer.

WordPiece tokenization

WordPiece tokenization is a **technique** used to **split words into smaller subword units**, allowing models to **handle rare or unseen words** more effectively.

Tokenization is a **fundamental pre-processing** step for most NLP applications.



Model evaluation

To **evalute models** some common **metrics** involves:

- **Precision** (1), measures how many of predicted positive instances are actually correct. Higher precision means fewer false positives.
- **Recall** (2), measures how many of the actual positive instances have been correctly identified. Higher recall means less false negatives.
- **F1 score** (3), measures the harmonic mean of precision and recall.

1) $$\frac{True\ positives}{True\ positives + False\ positives}$$

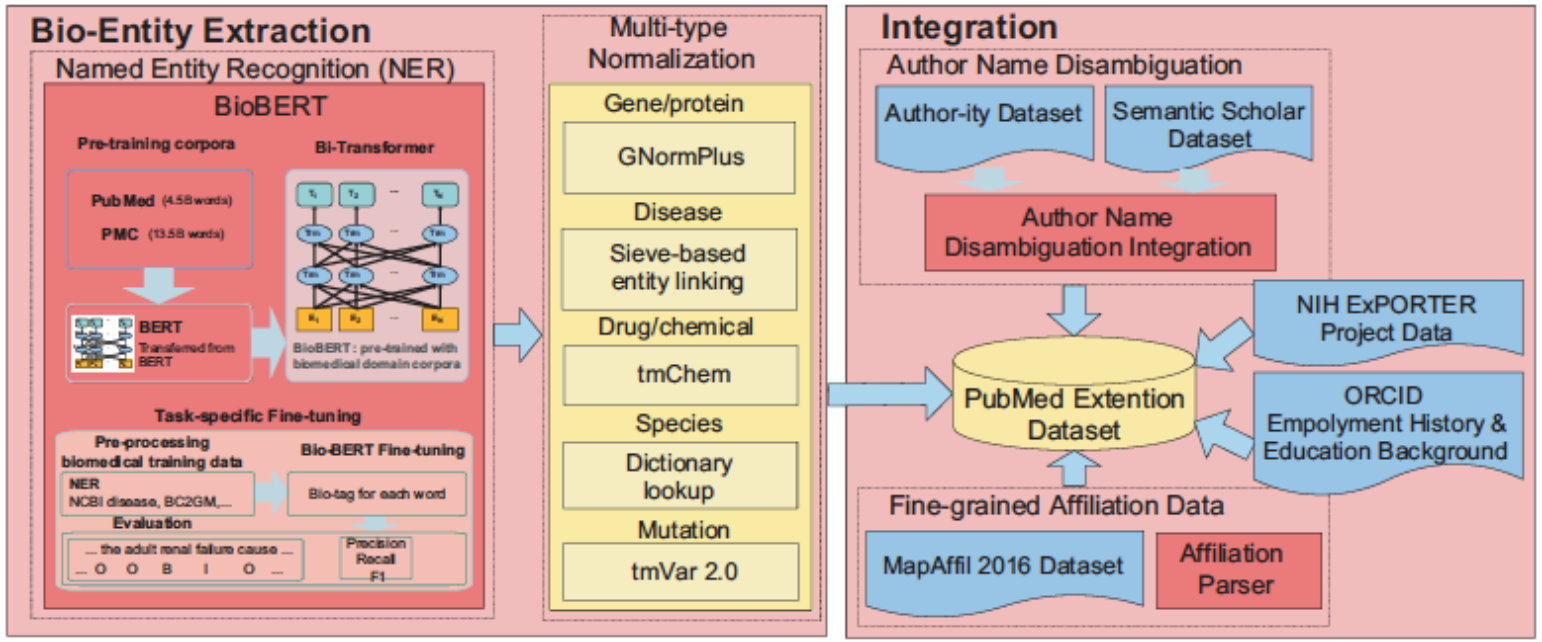
2) $$\frac{True\ positives}{True\ positives + False\ negatives}$$

3) $$2 * \frac{Precision * Recall}{Precision + Recall}$$

Bio-entity integration framework for PKG

The **bio-entity integration framework** consists of **two parts**.

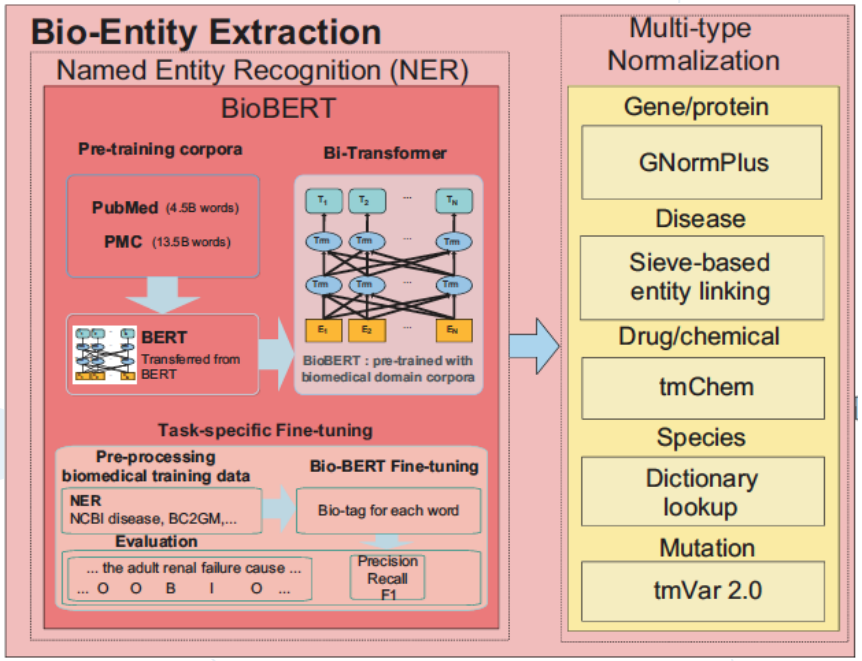
- **Bio-entity extraction**, which contains entity extraction, named entity recognition (NER), and multi-type normalization.
- **Integration**, which connects authors, ORCID, and funding information



Bio-entity extraction

Bio-entity extraction component involves **2 models**:

- **Named Entity Recognition** (NER).
- **Multi-type normalization**.



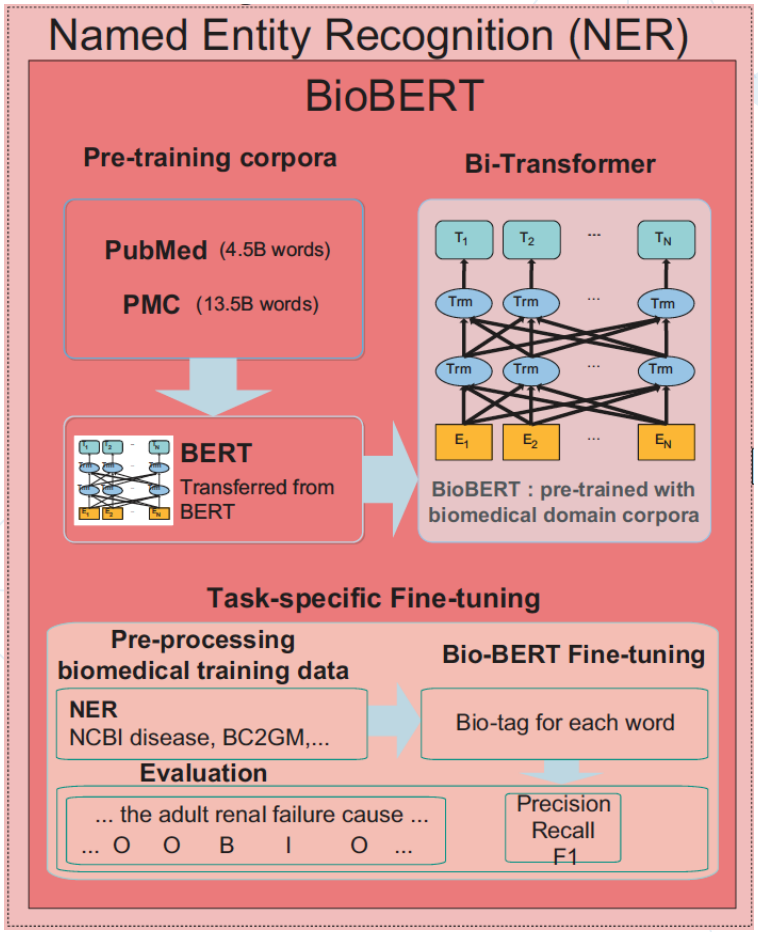
Named entity recognition

NER task recognizes a variety of domain-specific terms/entities in a biomedical corpus and is perceived as one of the **most notable biomedical text mining tasks**.

Case-sensitive BERT, designed as general-purpose language representation model, has been used to initialize **BioBERT**.

PubMed and PubMed Central articles were used to **pre-train weights**.

Then BioBERT has been **fine-tuned** using **WordPiece tokenization**.

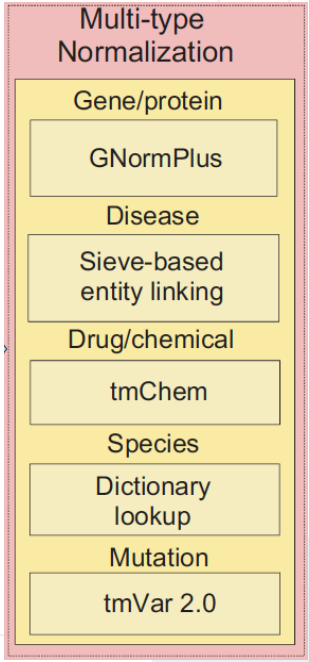


Multi-type normalization

Entity may be **referred** to by **several synonymous terms** (synonyms). A term can be **polysemous** if it **refers to multiple entity types** (polysemy).

A **normalization process is required**.

Single normalization tool for multiple entity types is difficult to build. So **multiple NER normalization models have been combined** into one **multi-type normalization** model that assigns IDs to extracted entities.

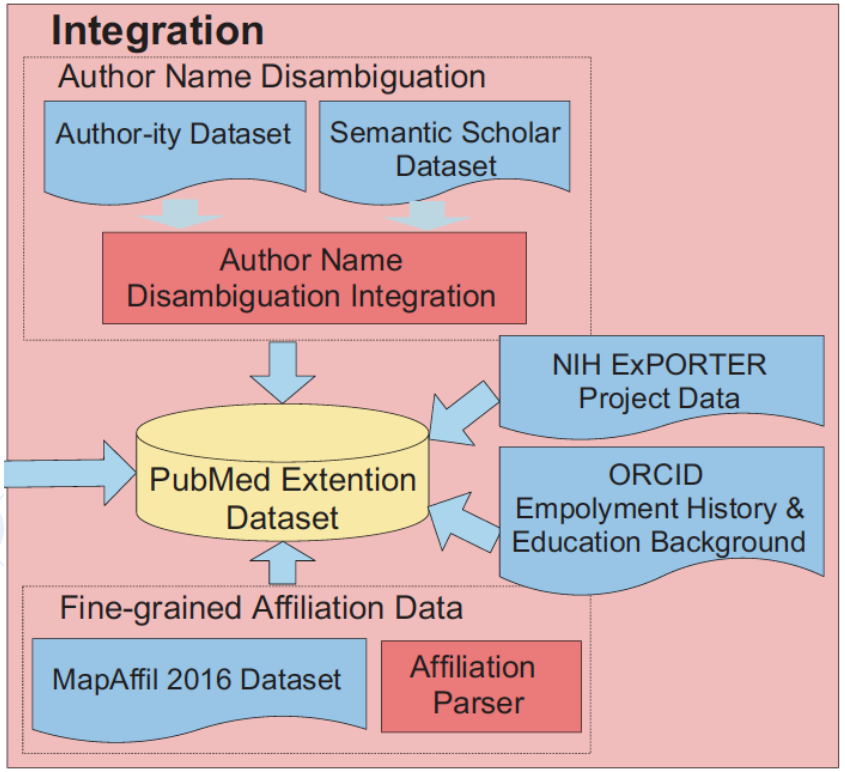


Entity types	Normalization models	Dictionaries	# of IDs	# of names	Avg. # of names per ID
Gene/Protein	GNormPlus	Entrez Gene[46]	139,375	248,581	1.8
Disease	Sieve-based entity linking[47]	MeSH[48], OMIM[49], SNOMED-CT[50], PolySearch2[51]	32,954	172,650	5.2
Drug/Chemical	tmChem without Ab3P	MeSH[48], ChEBI[52], DrugBank[18], US FDA-approved drugs	518,223	2,571,570	5.0
Species	Dictionary lookup	NCBI Taxonomy	398,037	3,119,005	7.8
Mutation	tmVar 2.0	dbSNP[53], Clin Var[54]	208,474	302,498	1.5
Total			1,297,063	6,414,304	4.9

Integration

The integration part **combines** extracted **bio-entity** with:

- **Author data**
- **Affiliations**
- **Funding information**



Author name disambiguation

In recent decades, researchers have made several attempts to **solve the AND problem**.

Researchers used **three types of methods**:

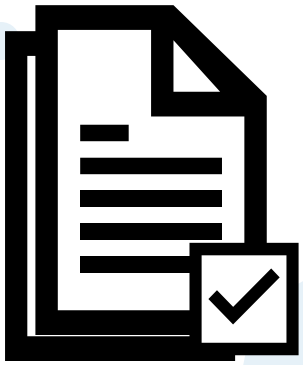
- **Manual matching** of articles with authors by surveying scientists or consulting curricula vitae (CVs) gathered from the Internet.
- **Publicly-accessible registry platforms**, such as ORCID or Google Scholar.
- **Automated approach** to estimate the similarity of author instance feature combinations and identify whether they refer to the same person.

AND manual matching

First method relies on **manual matching of articles** with authors by surveying scientists or consulting curricula vitae (CVs) gathered from the Internet.

This type of method ensures **high accuracy**, but requires a considerable **amount of investment** in labor to collect and code the data.

This method is **impractical for huge datasets**.

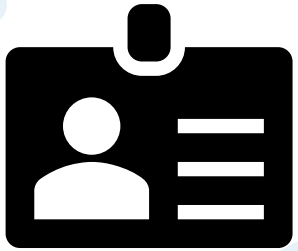


AND registry platforms

Second method relies on **publicly-accessible registry platforms**, such as ORCID or Google Scholar, to help researchers identify their own publications.

This type of method produces a **source of highly accurate** and **low-cost accessible disambiguation** of authorship for large numbers of authors.

Registries cover only a **small proportion of researchers**.



AND automated approach

Third method relies on **automated approach** to **estimate** the **similarity** of author instance **feature combinations** and identify whether they refer to the same person.

The features for automated AND include:

- Author name.
- Author affiliation.
- Article keywords.
- Journal names.
- Coauthor information.
- Citation patterns.

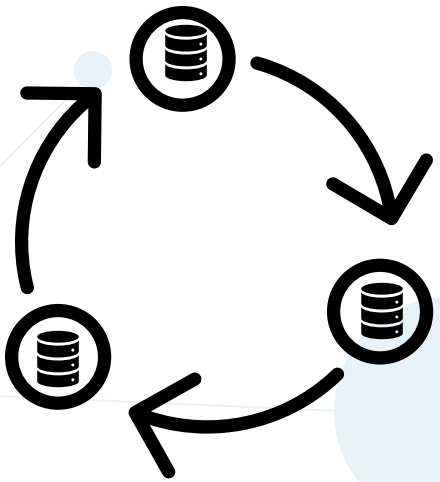
Automated methods typically rely on **supervised or unsupervised machine learning**.

AND PubMed

For PubMed, automated methods are the optimal choice.

A **high-quality PubMed AND dataset** with **complete coverage** can be obtained through the **integration of two existing AND datasets**:

- **Author-ity**, which uses diverse information about authors and publications to determine whether two or more instances of the same name (or of highly similar names) on different papers represent the same person.
- **Semantic Scholar**, which trains a binary classifier to merge a pair of author names and use the pair to create author clusters incrementally.

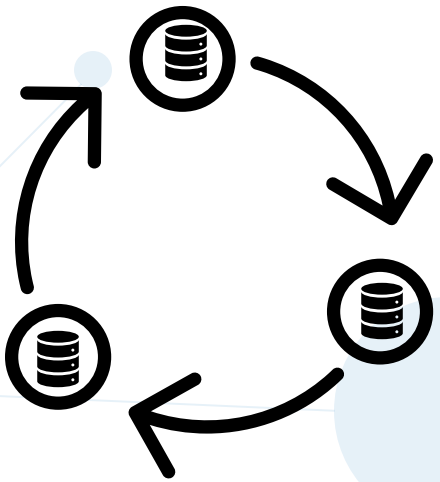


AND PubMed

Author-ity dataset has a higher F1 score than the Semantic Scholar dataset.

So author's unique ID of the Author-ity dataset as the **primary AND_ID**. These are **limited by time range**, containing only PubMed papers before 2009.

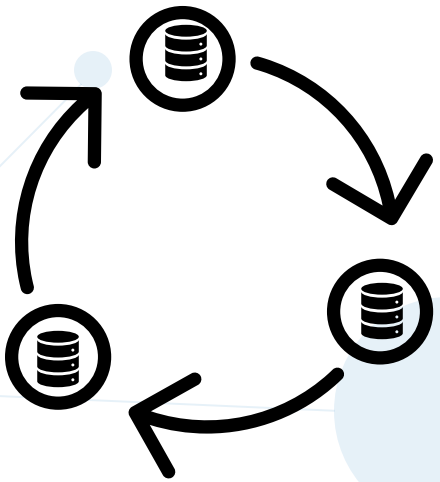
So authors after 2009 are supplemented using AND result from Semantic Scholar.



AND PubMed

To generate AND_IDs :

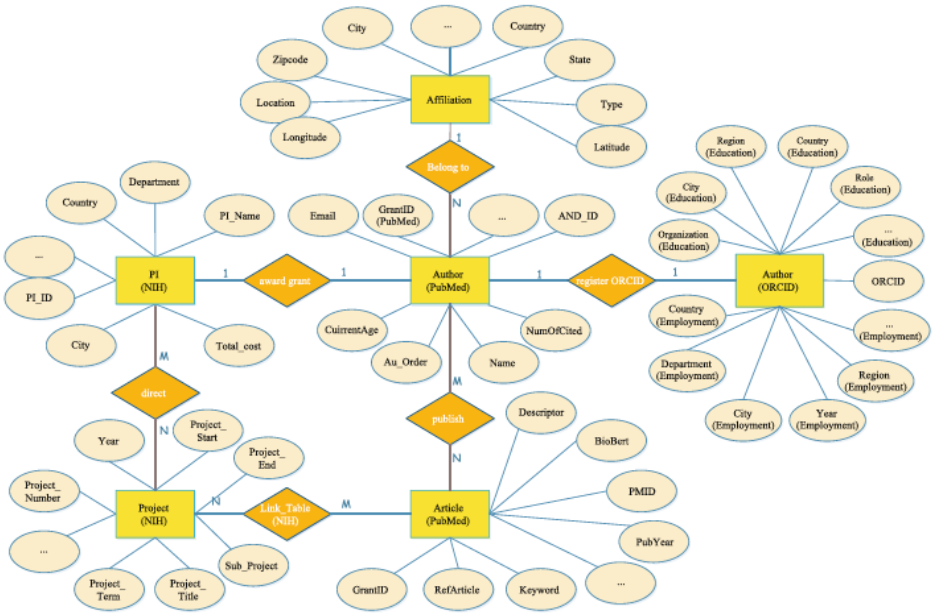
1. Author's unique ID of the Author-ity dataset as the primary AND_ID.
2. Authors that have the same Semantic Scholar AND_ID but never appear in the Author-ity dataset, are **generated new AND_ID to label them**.
3. Authors that have the a unique AND_ID in Semantic Scholar and in which authors had the same Author-ity AND_ID he Author-ity AND_ID is allocated to all author instances as their unique ID.



Extend multi-source information

In addition to bio-entity extraction by BioBERT and AND, PubMed has been integrated by mapping connections between AND_ID and the PubMed identifier (PMID) in order to build relationships between different objects to provide a comprehensive overview of the PubMed dataset.

The **integrations** involves information that **include** the funding data from **NIH ExPORTER**, the affiliation history and educational background of authors from **ORCID**, and the fine-grained region and location information from the **MapAffil**.



Data records

PKG resulting **dataset is freely avaiable** as CSV format.

File	# of Lines	# of Distinct PMIDs	# of Distinct AND_IDs	Short description
Author_List	114,345,178	28,510,300	14,830,461	CSV file containing PubMed authors and AND_IDs.
Bio-entities_Main	330,394,494	18,361,409	—	CSV file containing all types of extracted bio-entities by BioBERT.
Bio-entities_Mutation	1,388,341	312,099	—	CSV file containing additional items of mutations from Bio-entities_Main file.
Affiliations	46,065,099	19,601,383	8,300,984	CSV file containing affiliations and their extracted fine-grained items.
Researcher_Employment	532,356	—	276,483	CSV file containing employment history from ORCID.
Researcher_Education	512,267	—	268,610	CSV file containing educational background from ORCID.
NIH_Porjects	12,340,431	1,790,949	102,070	CSV file containing projects from NIH ExPORTER and mapping relation between PI_ID, PMID, and AND_ID.

Technical validation

PKG has been **validated on different aspects**:

- **Bio-entity extraction**.
- **Multi-type entity normalization**.
- **Author name disambiguation**.

Evaluation metrics involve entity-level **precision**, **recall** and **F1 scores**.

Validity of bio-entity extraction

To **validate the performance** of the bio-entity extraction, BioBERT has been compared to BERT and to the state-of-the-art models.

BERT as pre-trained on the general domain corpus, was highly effective.

State-of-the-art models outperformed BERT.

BioBERT outperformed state-of-the-art.

Entity Type	Datasets	Metrics	State-of-the-art	BERT (Wiki + Books)	BioBERT (+PubMed + PMC)
Disease	NCBI disease[55]	P %	86.41	84.12	**89.04**
		R %	88.31	87.19	**89.69**
		F %	87.34	85.63	**89.36**
	2010 i2b2/VA[56]	P %	87.44	84.04	**87.50**
		R %	**86.25**	84.08	85.44
		F %	**86.84**	84.06	86.46
	BC5CDR[57]	P %	85.61	81.97	**85.86**
		R %	82.61	82.48	**87.27**
		F %	84.08	82.41	**86.56**
Drug/Chemical	BC5CDR[57]	P %	**94.26**	90.94	93.27
		R %	92.38	91.38	**93.61**
		F %	93.31	91.16	**93.44**
	BC4CHEMD[58]	P %	91.30	91.19	**92.23**
		R %	87.53	88.92	90.61
		F %	89.37	90.04	**91.41**
Gene/Protein	BC2GM[59]	P %	81.81	81.17	**85.16**
		R %	81.57	82.42	83.65
		F %	81.69	81.79	**84.40**
	JNLPBA[60]	P %	**74.43**	69.57	72.68
		R %	**83.22**	81.20	83.21
		F %	**78.58**	74.94	77.59
Species	LINNAEUS[61]	P %	92.80	91.17	**93.84**
		R %	**94.29**	84.30	86.11
		F %	**93.54**	87.6	89.81
	Species-800[62]	P %	**74.34**	69.35	72.84
		R %	75.96	74.05	**77.97**
		F %	74.98	71.63	**75.31**
Average		P %	85.38	82.61	**85.82**
		R %	85.79	84.00	**86.40**
		F %	85.53	83.25	**86.04**

Validity of multi-type entity normalization

Performance **varied** by **entity type** and **dataset specificity**.

Most cells are empty because the data were not be available.

Entity type	Normalization model	Test sets	Precision %	Recall %	F1 score %	Accuracy %
Gene/Protein	GNormPlus	BC2 Gene Normalization, human species[63]	87.1	86.4	86.7	—
		BC3 Gene Normalization, multispecies[64]	—	—	50.1	—
Disease	Sieve-based entity linking	ShARe/CLEF eHealth Challenge corpus[65]	—	—	—	90.75
		NCBI disease	—	—	—	84.65
Mutation	tmVar 2.0	OSIRISv1.2[66]	97.20	80.62	88.14	—
		Thomas[67]	89.94	88.24	89.08	—
Species	Dictionary lookup of SR4GN[68]	BioCreative III GN[69]	—	—	46.91	—

Validity of AND

The validation of AND remains a **challenge** because there is a **lack of abundant validation sets**.

Validation has been **done on NIH ExPORTER-provided** information on NIH-funded researchers.

After integrating the AND results of Author-ity and Semantic Scholar, has been obtained a **high-quality integrated AND** result that outperformed Semantic Scholar by 1.15% in terms of the F1 score and had more comprehensive coverage (until 2018) than Author-ity (until 2009).

	Precision	Recall	F1 score
Author-ity	99.43%	96.92%	98.16%
Semantic Scholar	96.24%	97.66%	96.94%
AND Integration	98.62%	97.56%	98.09%

Conclusions

The **PubMed Knowledge Graph** (PKG) offers a comprehensive and integrated biomedical resource by **combining** advanced **bio-entity extraction**, high-quality **author name disambiguation**, and multi-source **data integration**.

PKG effectively **addresses challenges of ambiguity and data fragmentation** within PubMed literature, enabling enhanced knowledge discovery, scientific collaboration, and research profiling.

With its open accessibility and extensive coverage, PKG paves the way for new applications in biomedical informatics, including expert finding, trend analysis, and collaborative recommendations.

References

[1] *XU, Jian, et al. Building a PubMed knowledge graph. Scientific data, 2020, 7.1: 205.*

[2] *EHRLINGER, Lisa; WÖß, Wolfram. Towards a definition of knowledge graphs. SEMANTiCS (Posters, Demos, SuCCESS), 2016, 48.1-4: 2.*

[3] *TURC, Iulia, et al. Well-read students learn better: On the importance of pre-training compact models. arXiv preprint arXiv:1908.08962, 2019.*

[4] *https://research.google/blog/a-fast-wordpiece-tokenization-system/*

[5] *https://www.ibm.com/think/topics/natural-language-processing*

[6] *https://www.ibm.com/think/topics/knowledge-graph*

[7] *Wood EH. MEDLINE: the options for health professionals. J Am Med Inform Assoc. 1994 Sep-Oct;1(5):372-80. doi: 10.1136/jamia.1994.95153425. PMID: 7850561; PMCID: PMC116219.*

Thanks