



# Grafy Wiedzy: Fundamenty, Techniki i Zastosowania

Kompleksowe streszczenie na podstawie książki  
M. Kejdiwala, C. Knoblocka i P. Szekely'ego (MIT Press)

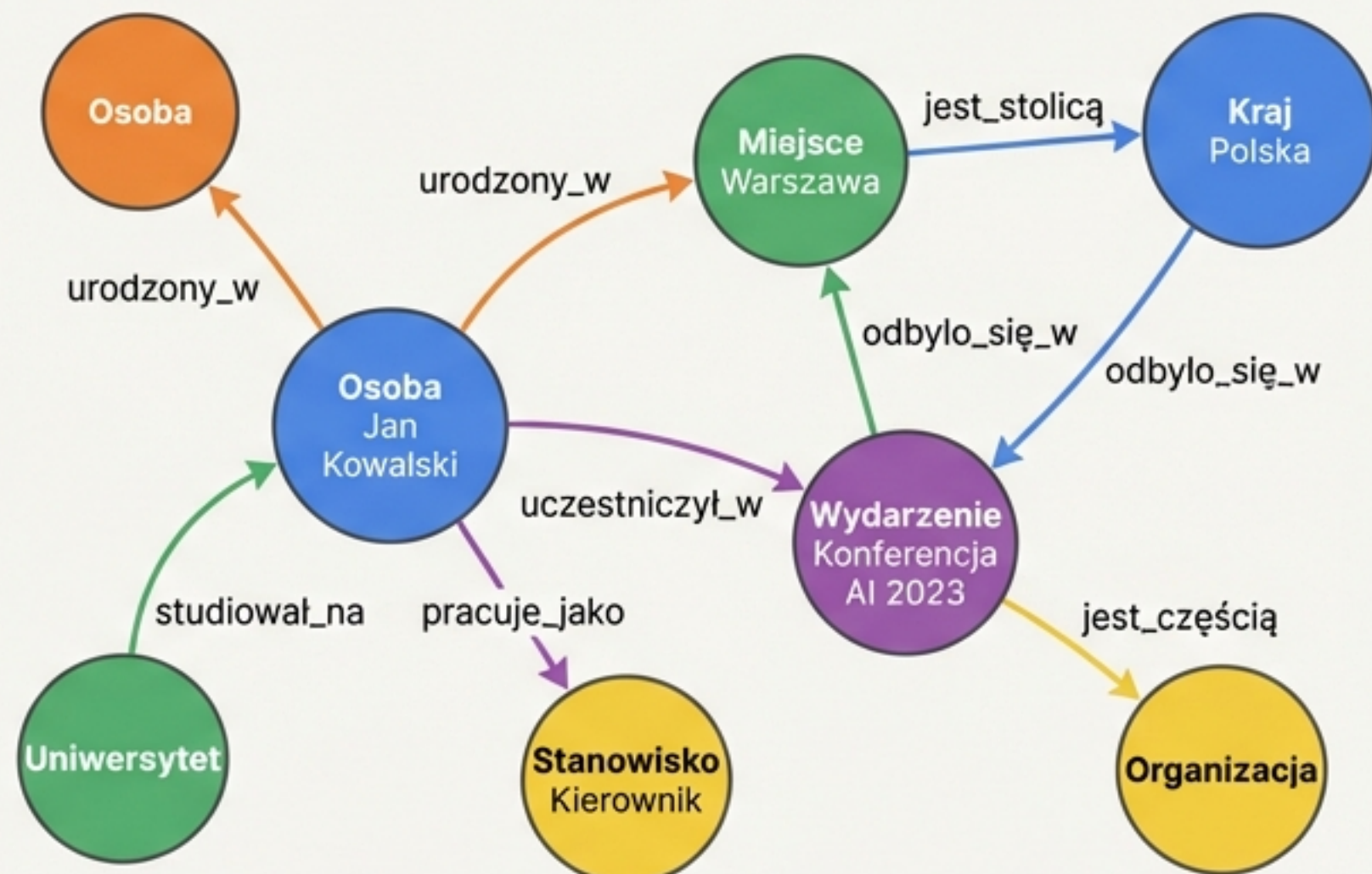
# Wprowadzenie: Czym są Grafy Wiedzy (KG)?

## Tradycyjna Baza Danych

ID	Imię	Nazwisko	Wiek	Adres	Stanowisko
001	Jan	Kowalski	35	ul. Długa 1, Warszawa	Kierownik
002	Kyka	Kowalski	32	ul. Długa 1,	Kierownik
003	Hanegh	Flyrm	59	Ltuct 3	Kierownik
004	Sarkes	Laptop	47	ul. Długa 1, Warszawa	Kierownik
...	...	...	...	...	...

Produkt ID	Nazwa Produktu	Cena
P101	Laptop	4500
...	...	...

## Graf Wiedzy



## Definicja

Reprezentacja informacji o świecie rzeczywistym (encje, relacje, atrybuty) w formie czytelnej dla maszyn.

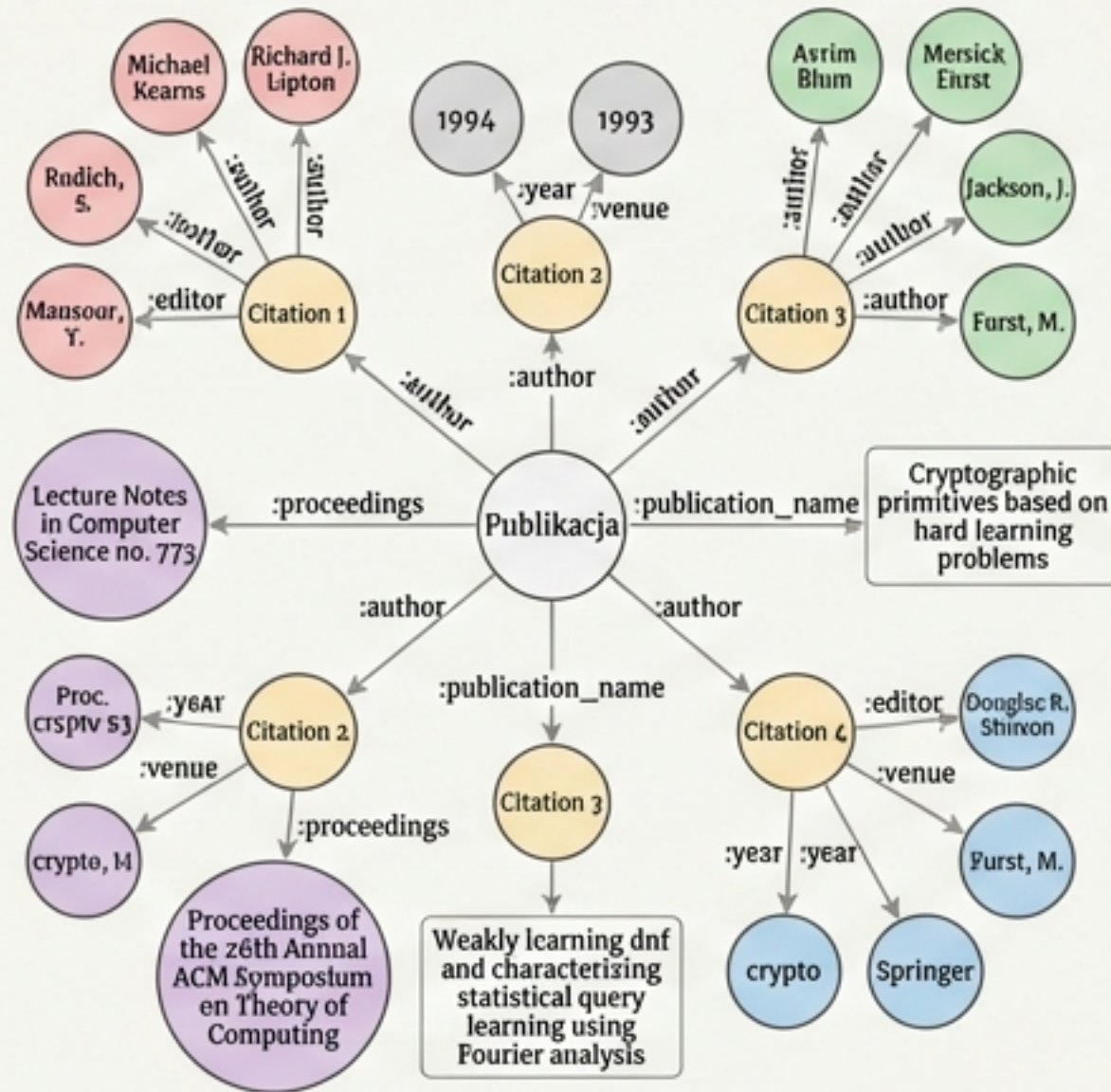
## Kontekst historyczny

Termin zyskał popularność po ogłoszeniu Google Knowledge Graph w 2011 roku, choć fundamenty (AI, bazy wiedzy) istniały od dekad.

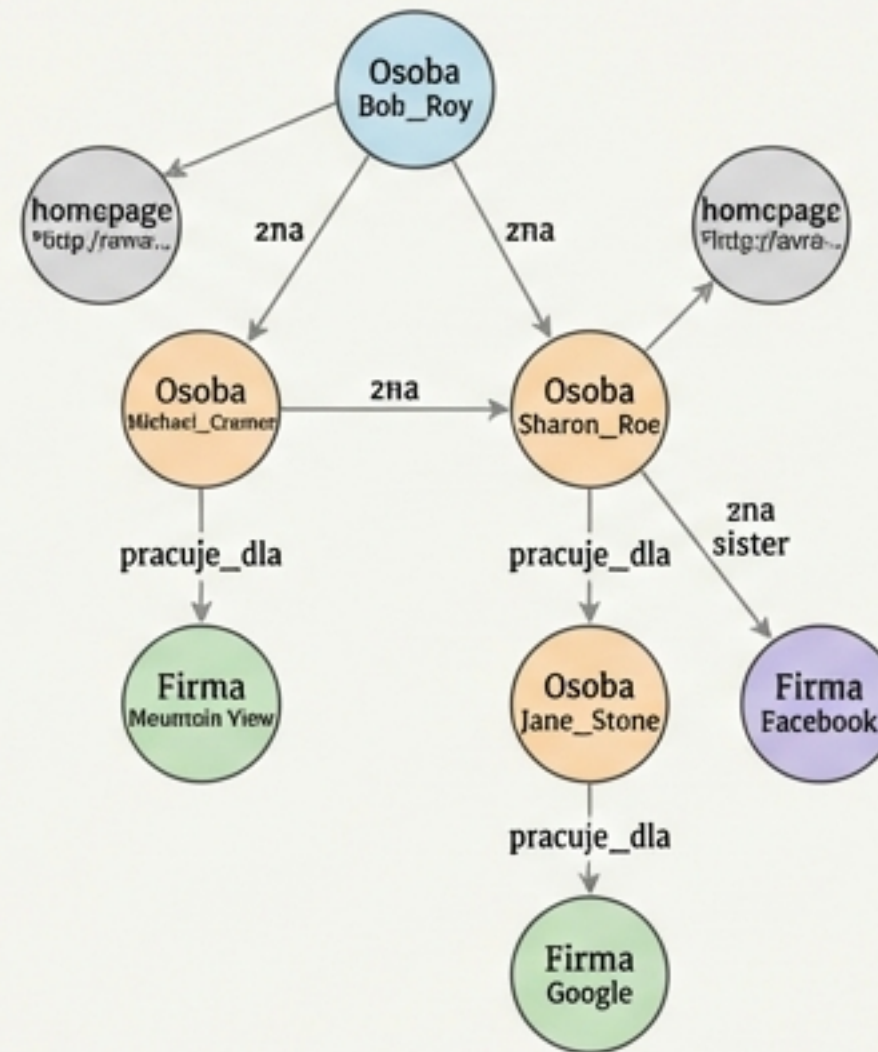
## Dlaczego grafy?

Elastyczność w obsłudze danych częściowo ustrukturyzowanych i skala webowa ("Web scale").

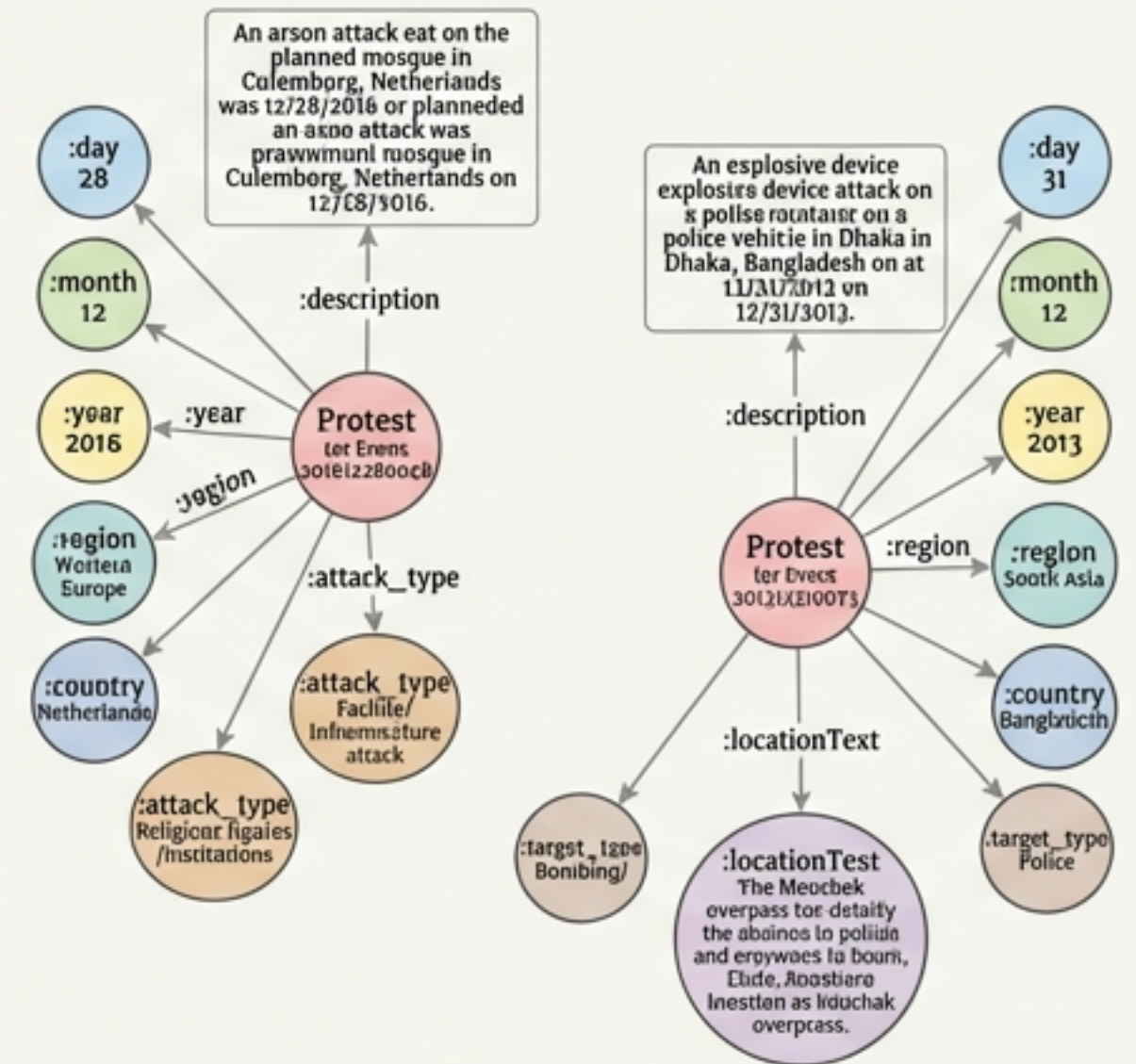
# Reprezentacja Wiedzy w Praktyce



Nauka i Publikacje  
(Autorzy, Cytowania)



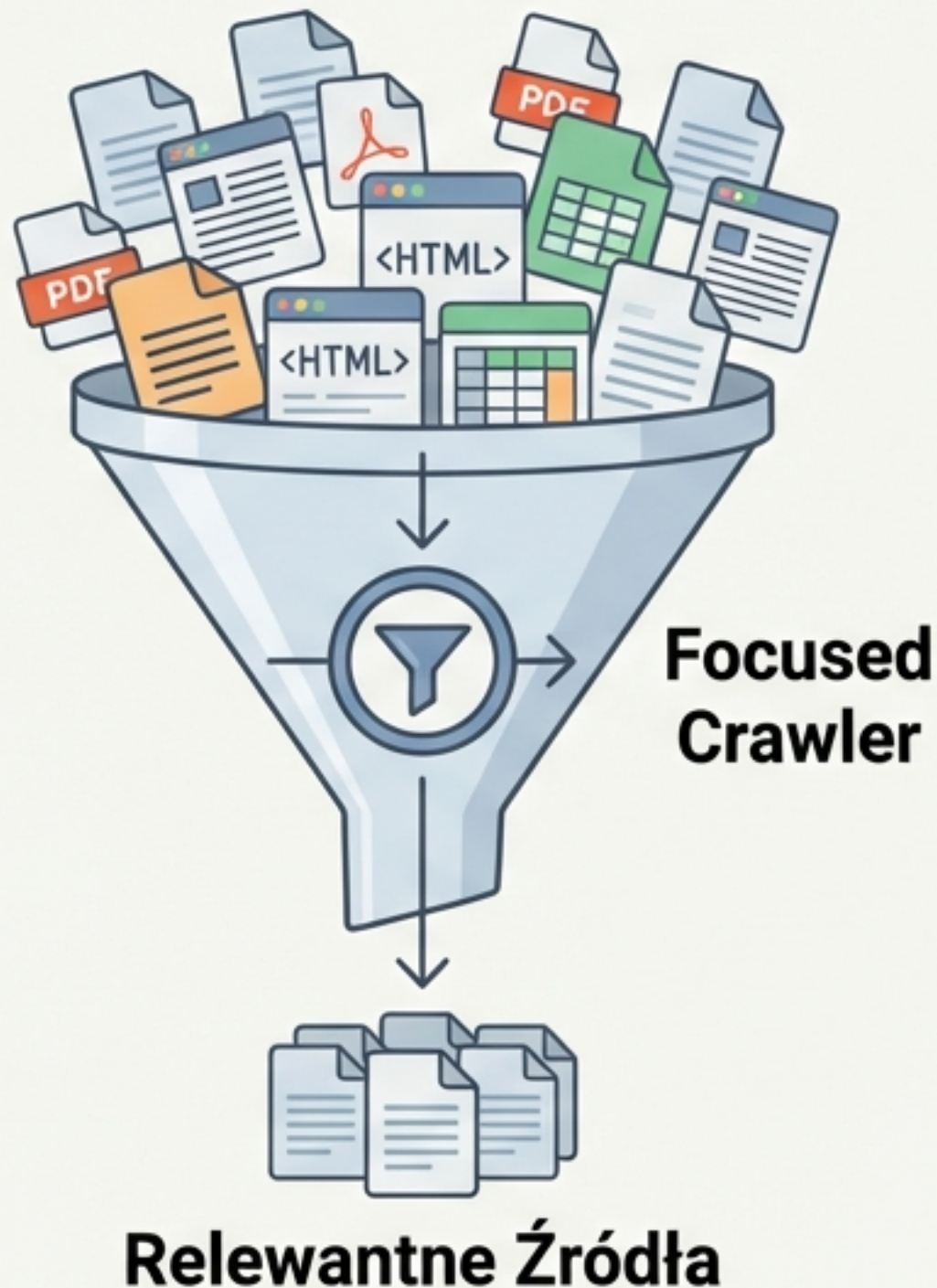
Sieci Społecznościowe  
(FOAF)




Wydarzenia Geopolityczne  
(GDELT)


# Konstrukcja Grafu (I): Odkrywanie i Ekstrakcja


Nieuporządkowana Sieć WWW



## Wyzwania Big Data (3V)

 **Volume:** Ogromna ilość danych do przetworzenia.

 **Velocity:** Szybkość zmian – dane stają się nieaktualne.

 **Variety:** Różnorodność formatów (HTML, PDF, tabele).

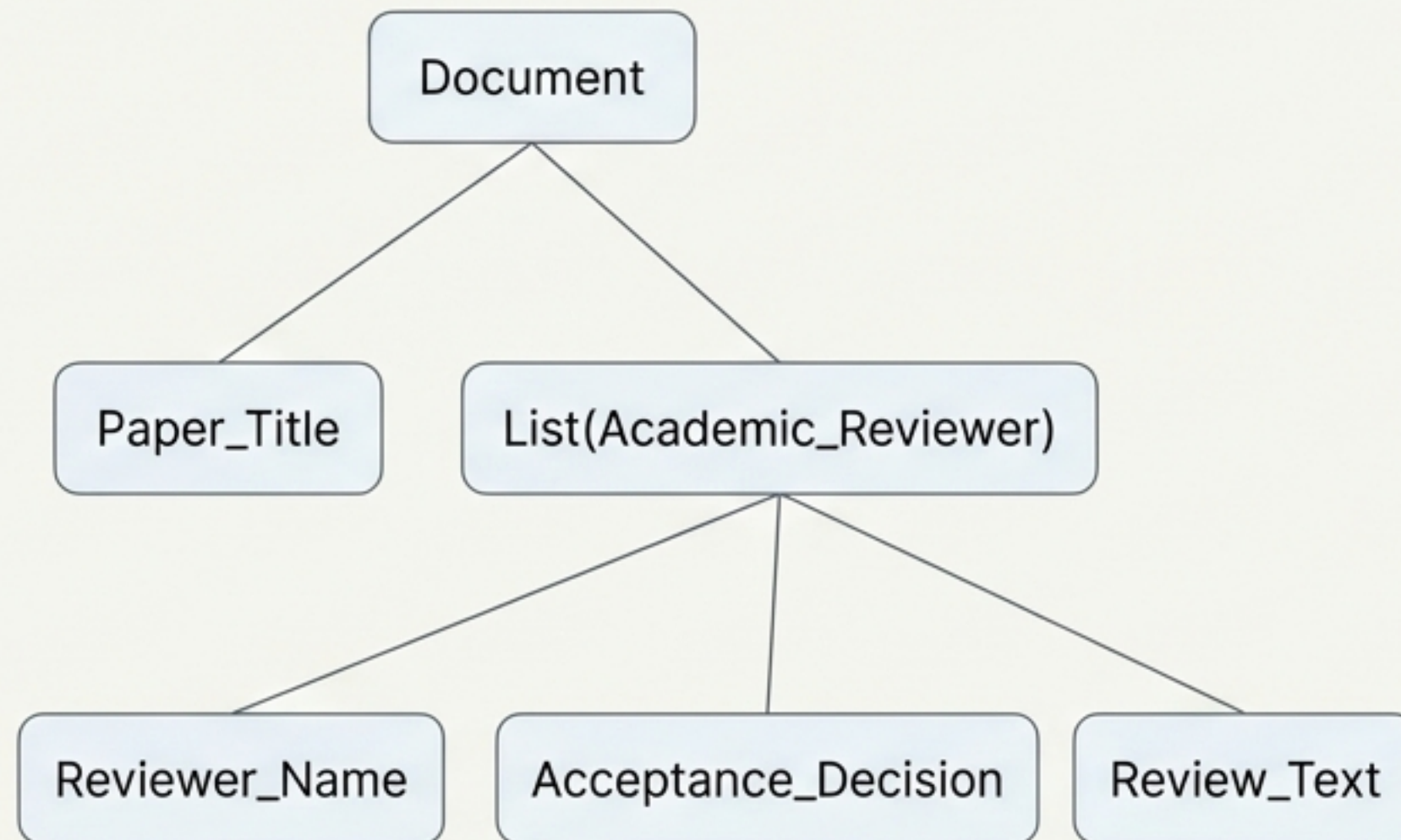
---

Domain Discovery: Proces identyfikacji źródeł specyficznych dla danej dziedziny.

# Konstrukcja Grafu (II): Wrappery i Przetwarzanie Stron

## Web Information Extraction (Web IE)

Przekształcanie stron HTML w ustrukturyzowane rekordy.



## Wrappery

Programy parsujące drzewo DOM (znaczniki HTML) w celu precyzyjnego wycięcia danych.

# Rozpoznawanie Encji Nazwanych (NER)

Elon Musk założył SpaceX w Kalifornii.

OSOBA

ORGANIZACJA

LOKALIZACJA

## Ewolucja Metod



### Podejścia Regułowe

Listy, wzorce gramatyczne

### Uczenie Nadzorowane

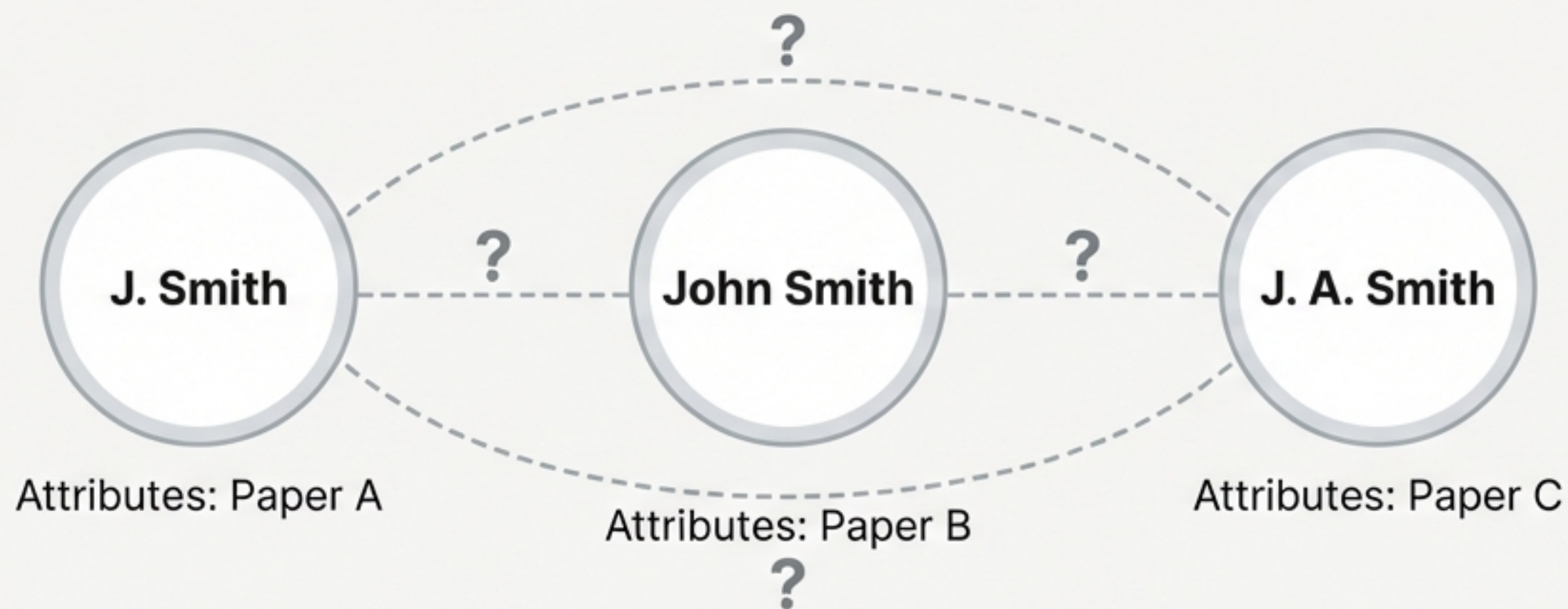
Korpusy treningowe

### Deep Learning

Sieci neuronowe, brak sztywnych reguł

# Kompletność i Jakość: Problem "Brudnych Danych"

## Duplicate Problem



## Typowe błędy w surowych grafach:

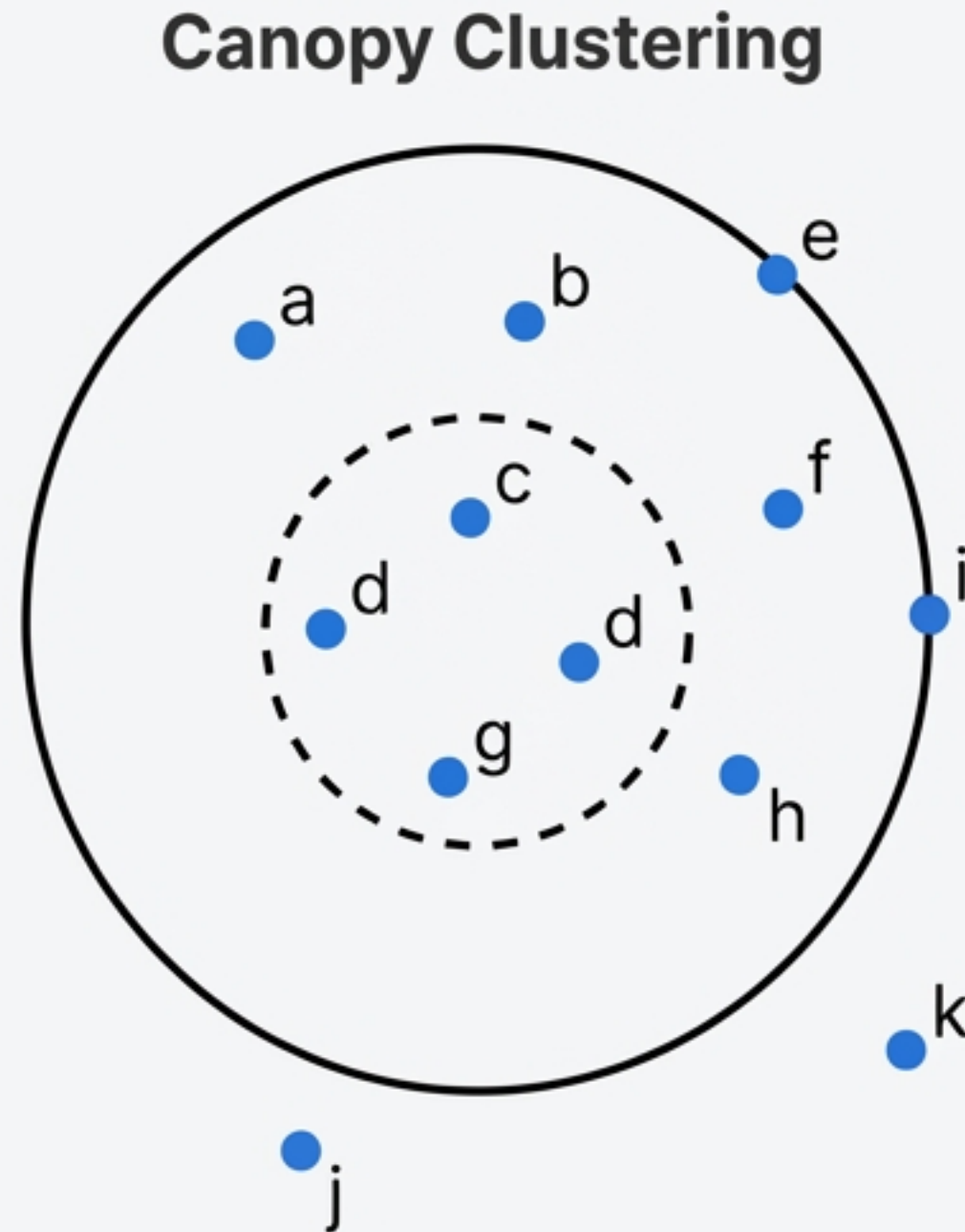
- **Duplikaty:** Ta sama encja występująca pod różnymi nazwami.
- **Braki danych:** Puste atrybuty i brakujące relacje.
- **Błędy:** Literówki i przestarzałe informacje.

**Konsekwencja: Błędne wyniki analiz (np. zaniżona liczba publikacji autora).**

# Uzgadnianie Instancji (Instance Matching - IM)

## Definicja:

Proces klastrowania wzmiarek tak, aby każdy klaster odnosił się do jednej unikalnej encji.



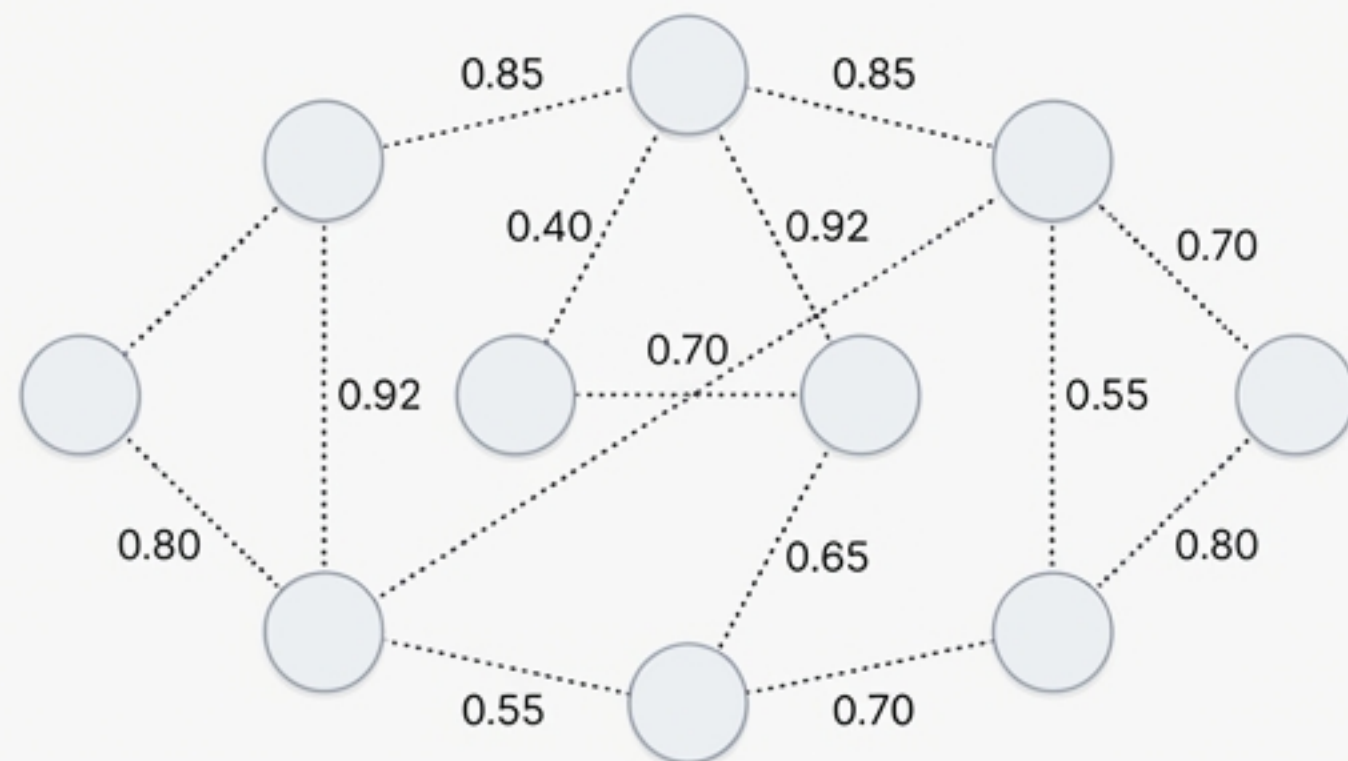
## Blocking & Canopies:

**Problem:** Porównywanie 'każdego z każdym' jest zbyt kosztowne obliczeniowo.

**Rozwiązanie:** Grupowanie obiektów w bloki i porównywanie tylko wewnątrz nich (złożoność zredukowana).

# Wnioskowanie w Warunkach Niepewności (SRL)

Statistical Relational Learning (SRL)



## Koncepcja

Łączenie logiki (reguł) z prawdopodobieństwem. Dane w grafach nie są niezależne.

## Frameworki

Markov Logic Networks (MLN) oraz Probabilistic Soft Logic (PSL).

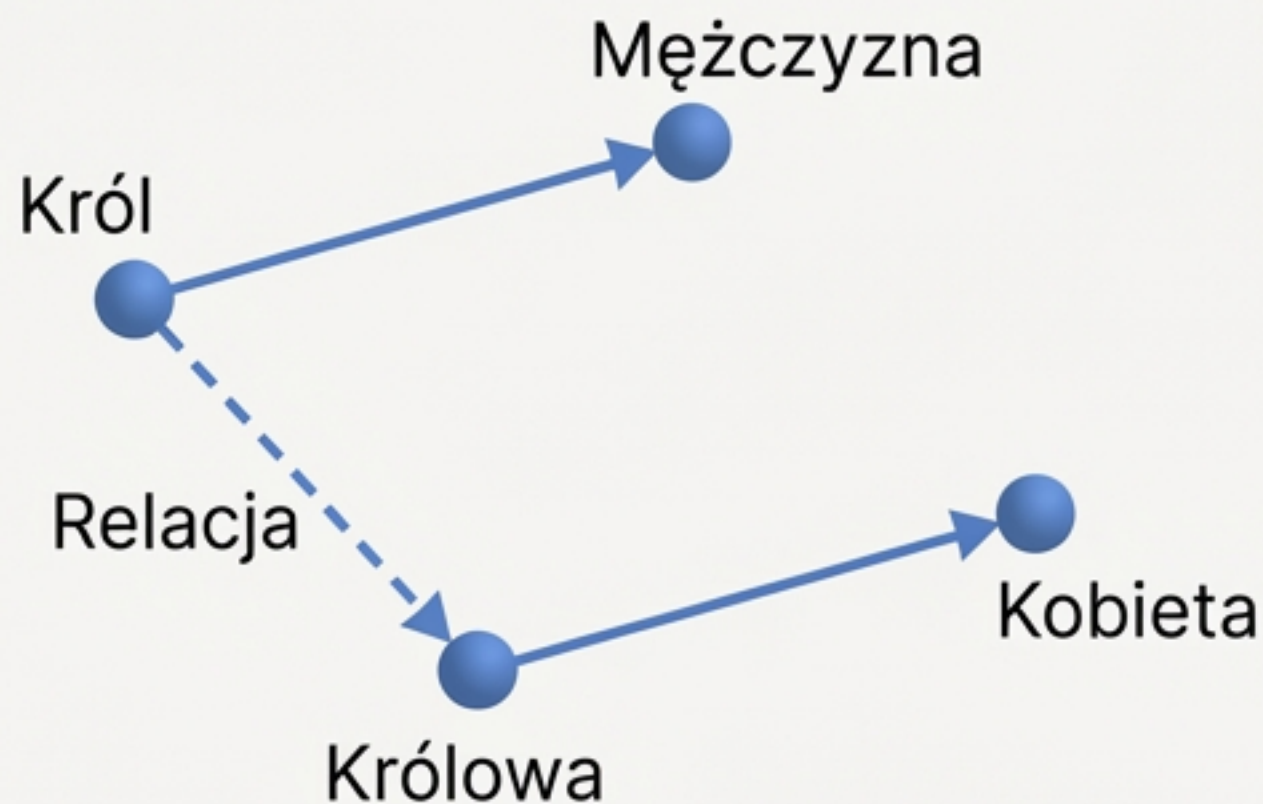
## Zastosowanie

Link Prediction (przewidywanie brakujących połączeń) i klasyfikacja zbiorowa.

# Uczenie Reprezentacji (Knowledge Graph Embeddings)

## Idea:

Zanurzenie węzłów i relacji w niskowymiarowej przestrzeni wektorowej.



## Cel:

Umożliwienie sieciom neuronowym "rozumienia" semantyki i uzupełniania grafu.

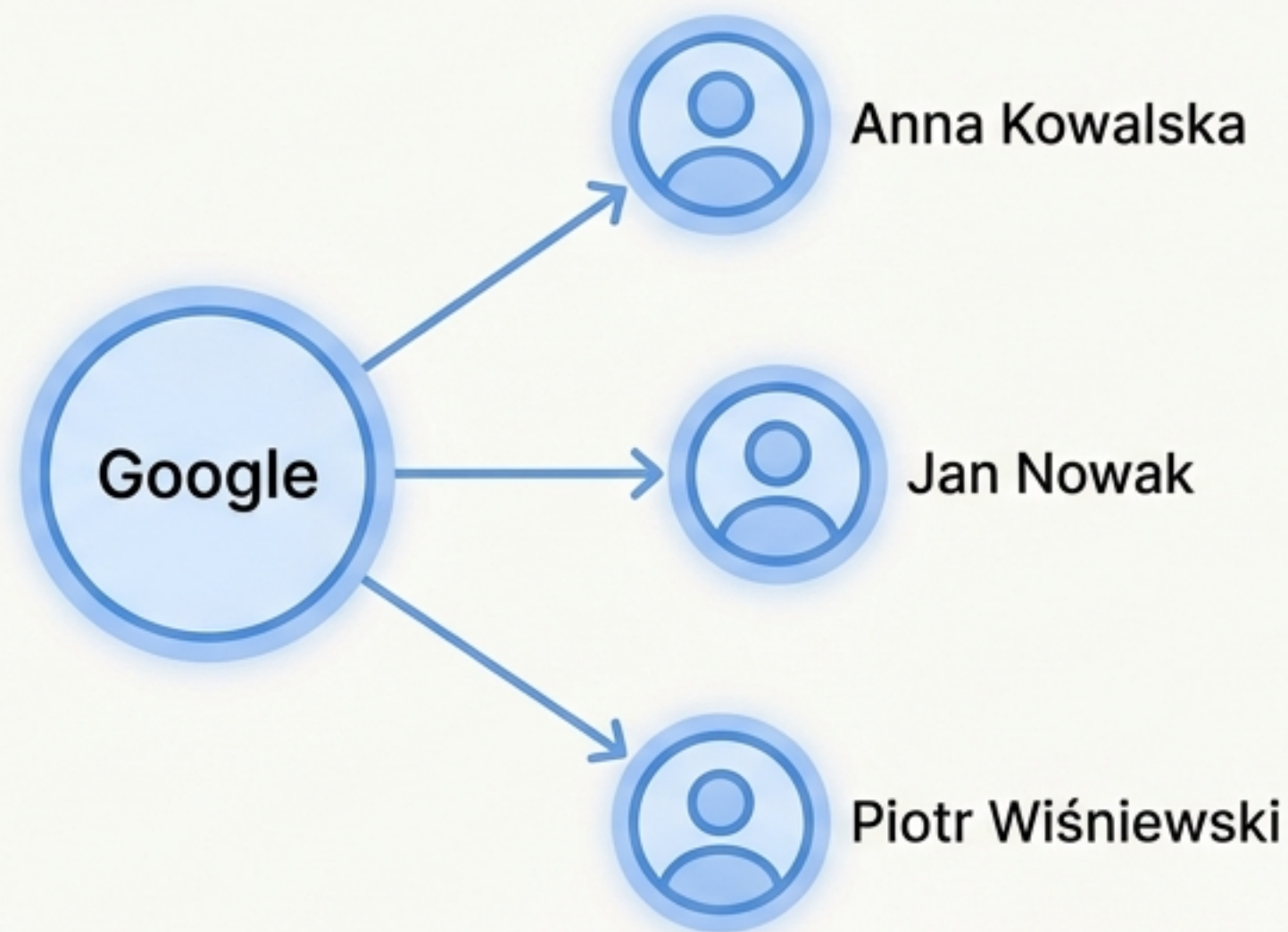
$$V(\text{Królowa}) \approx V(\text{Król}) - V(\text{Mężczyzna}) + V(\text{Kobieta})$$

# Dostęp do Wiedzy: Zapytania i SPARQL

## Zapytanie SPARQL

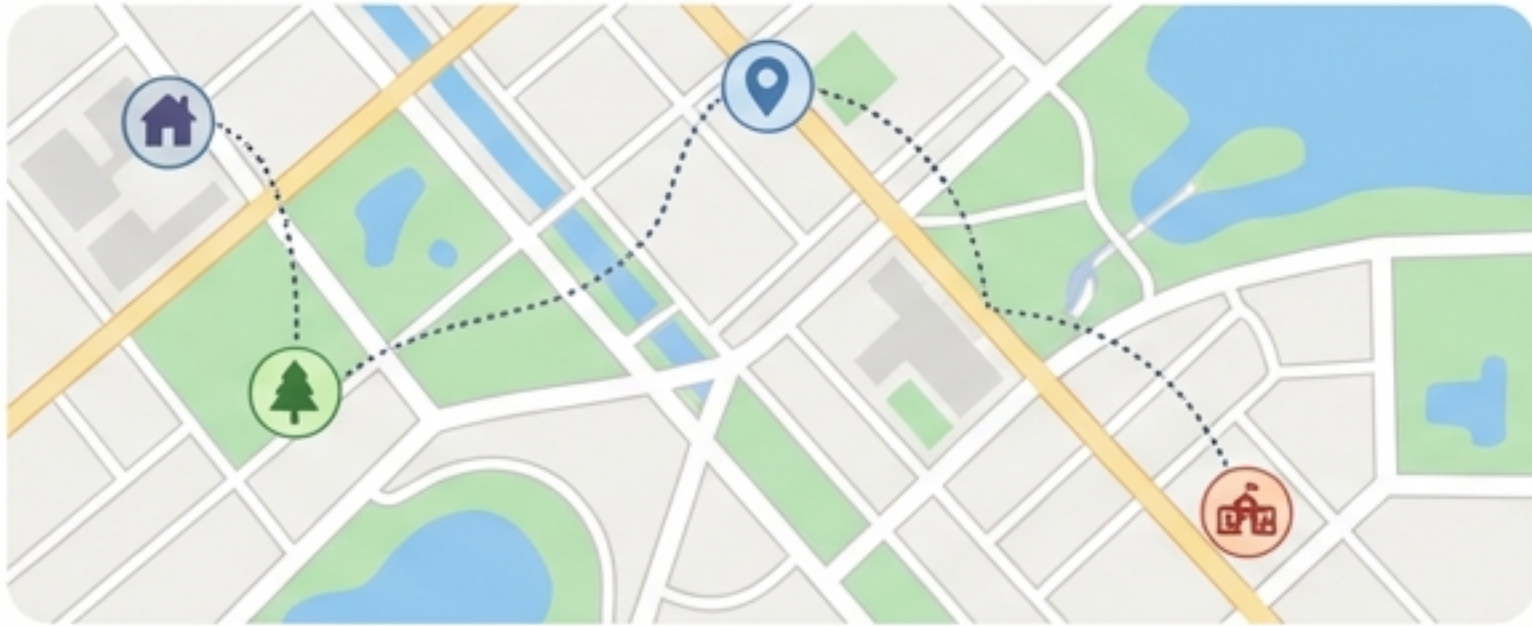
```
1 SELECT ?name
2 WHERE {
3   ?x works_for Google .
4   ?x has_name ?name
5 }
```

## Wynik Semantyczny



SPARQL to standard W3C dla grafów RDF – odpowiednik SQL dla baz relacyjnych. Umożliwia precyzyjne odpytywanie o wzorce grafowe.

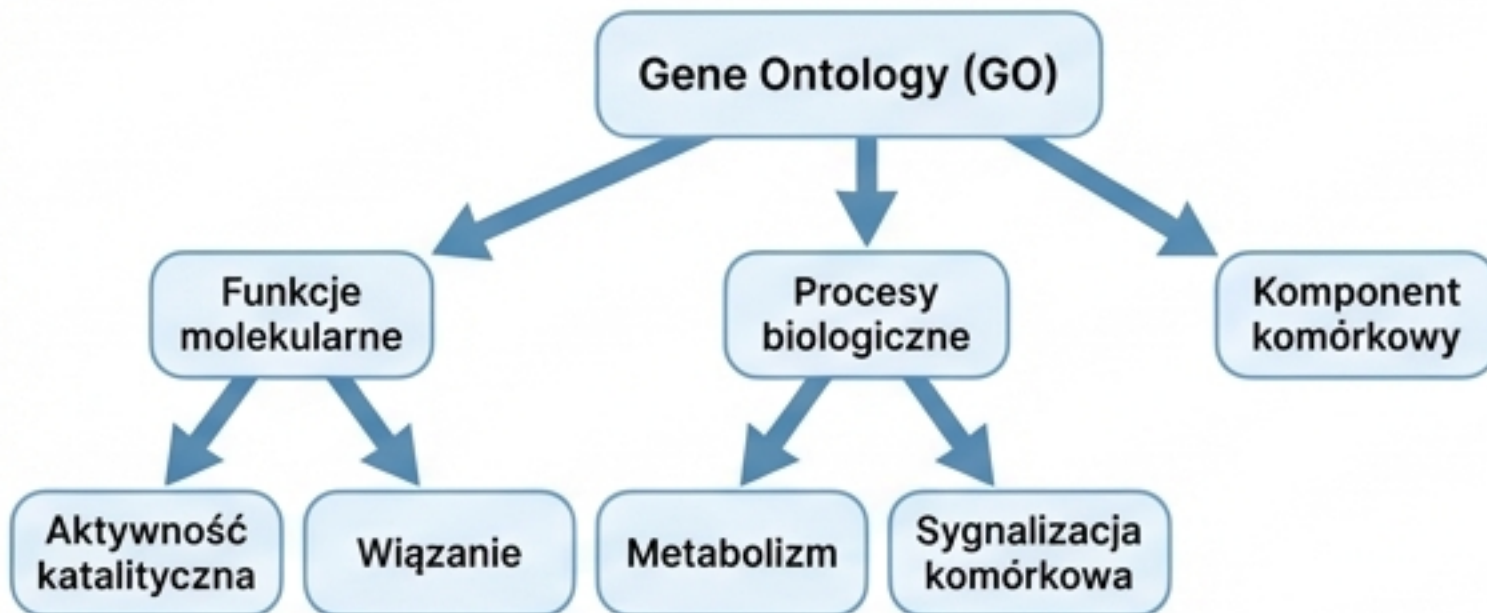
# Ekosystemy Danych: Geografia i Nauka



## Geografia (Crowdsourcing)

OpenStreetMap (OSM) & GeoNames.

"Wikipedia map" tworzona przez społeczność.  
Kluczowa w pomocy humanitarnej i nawigacji.

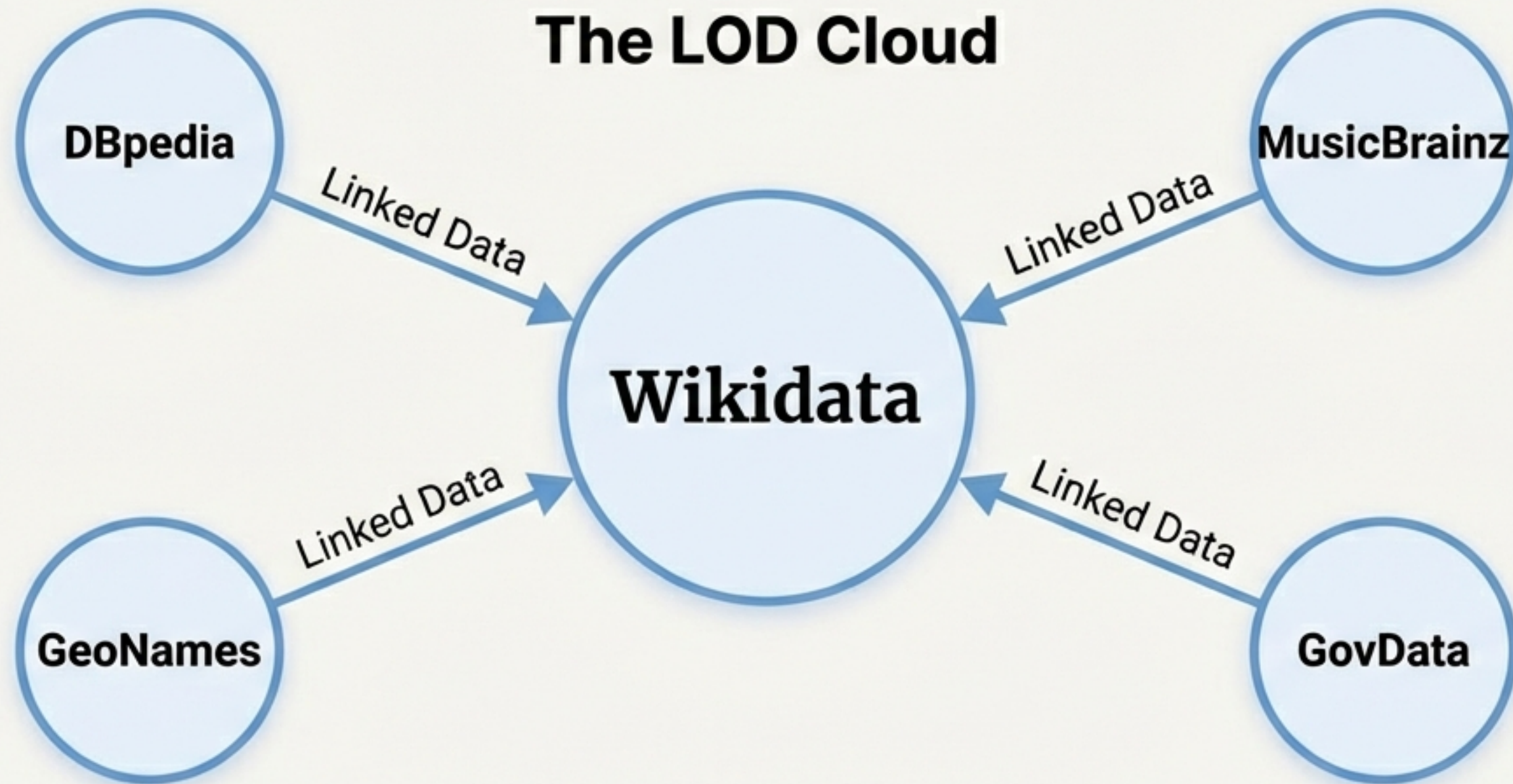


## Nauka (Bioinformatyka)

Gene Ontology (GO).

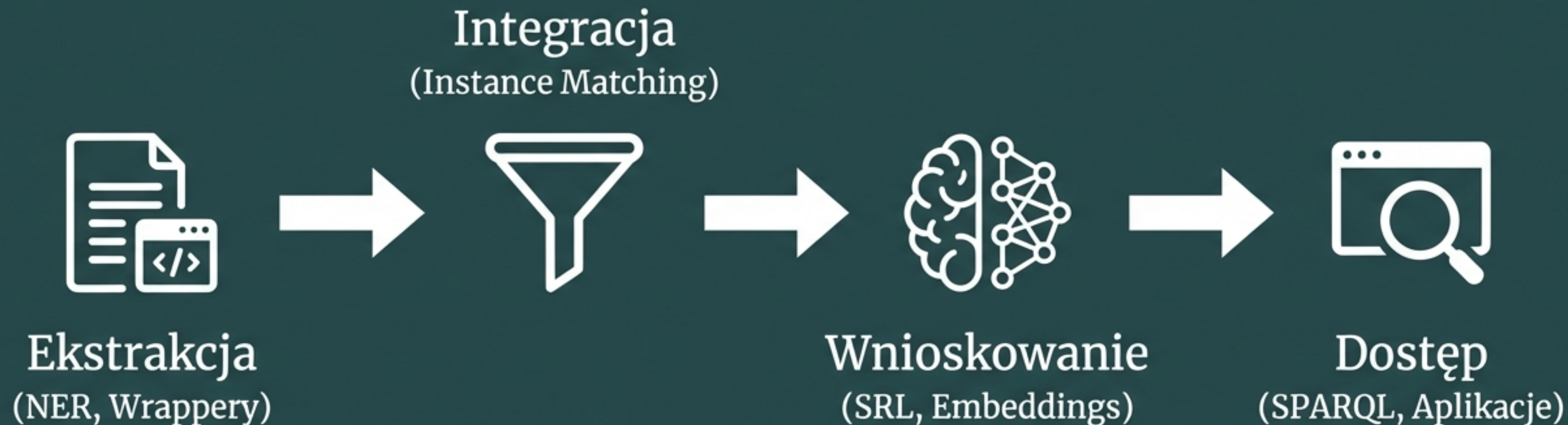
Standaryzacja wiedzy biologicznej (funkcje molekularne, procesy). Niezbędna do integracji danych z eksperymentów.

# Przyszłość: Linked Open Data (LOD)



Wizja: Przejście od 'Internetu Dokumentów' do 'Internetu Danych'. Łączenie niezależnych grafów w jedną globalną sieć za pomocą identyfikatorów URI.

# Podsumowanie: Cykl Życia Wiedzy



Grafy Wiedzy stanowią most między chaotycznymi danymi a rozumieniem maszynowym, napędzając nowoczesne AI.