OD2WD: From Open Data to Wikidata through Patterns

Muhammad Faiz, Gibran M.F. Wisesa, **Adila Krisnadhi**, and Fariz Darari

Faculty of Computer Science, Universitas Indonesia, Depok, Indonesia

Outline

- Motivation
- The OD2WD system
- Emerging patterns
- Discussion and Future Work

Motivation

- Worldwide open data adoption
- Indonesia: several open data portals with total of >50,000 CSV/Excel tables

Motivation

- Many portal stops at publishing CSV files hence preventing FAIR
- Linked Data is a solution but difficult due to technical, budgetary, or policy reasons

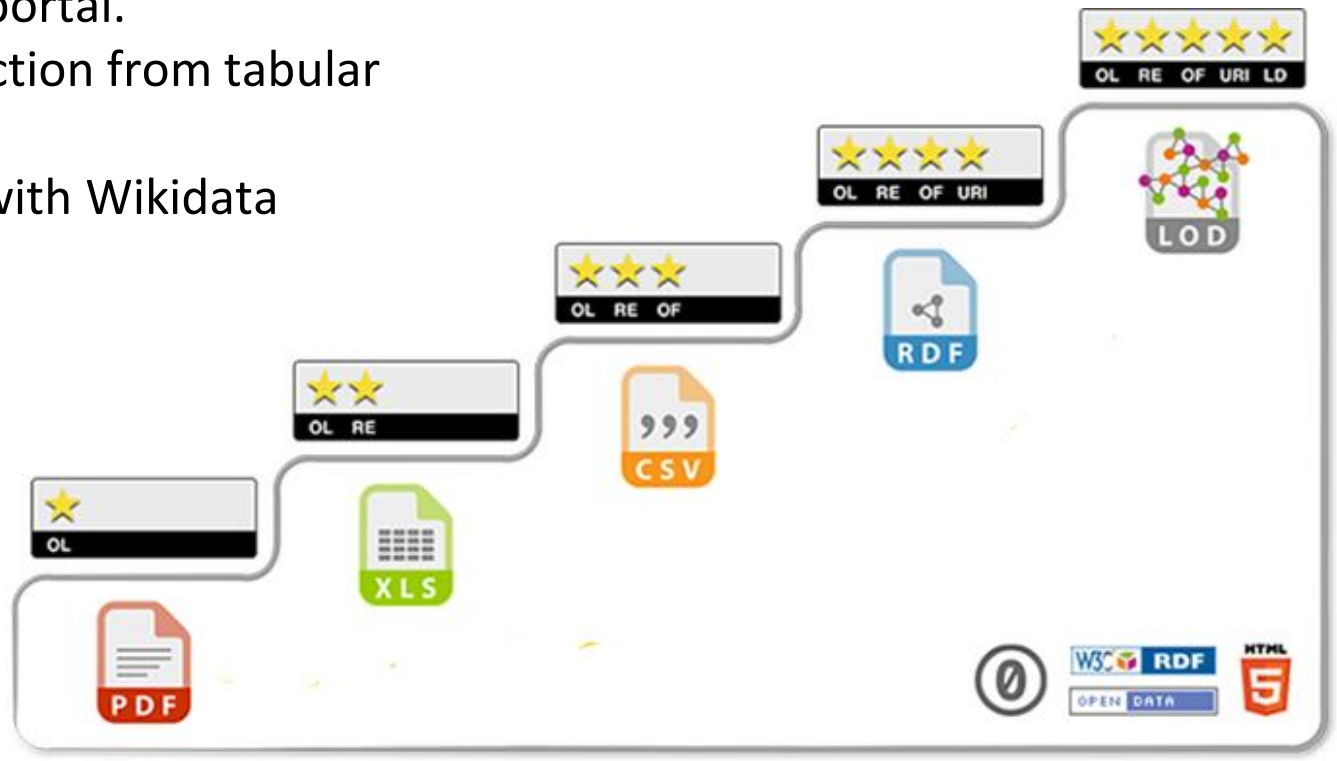
Proposed Solution

Idea: Make use of infrastructure of **existing** linked data infrastructure

- Transform and republish tabular data to repository of choice: Wikidata
- Upside #1: Allows further edits by public
- Upside #2: Wikidata is enriched further

OD2WD: Open Data to Wikidata

- Online at: http://od2wd.id
- Currently implemented for Satu Data Indonesia portal, Jakarta Open Data portal, and Bandung Open Data portal.
- Challenge #1: triple extraction from tabular cell values
- Challenge #2: alignment with Wikidata vocabulary



SATU DATA INDONESIA

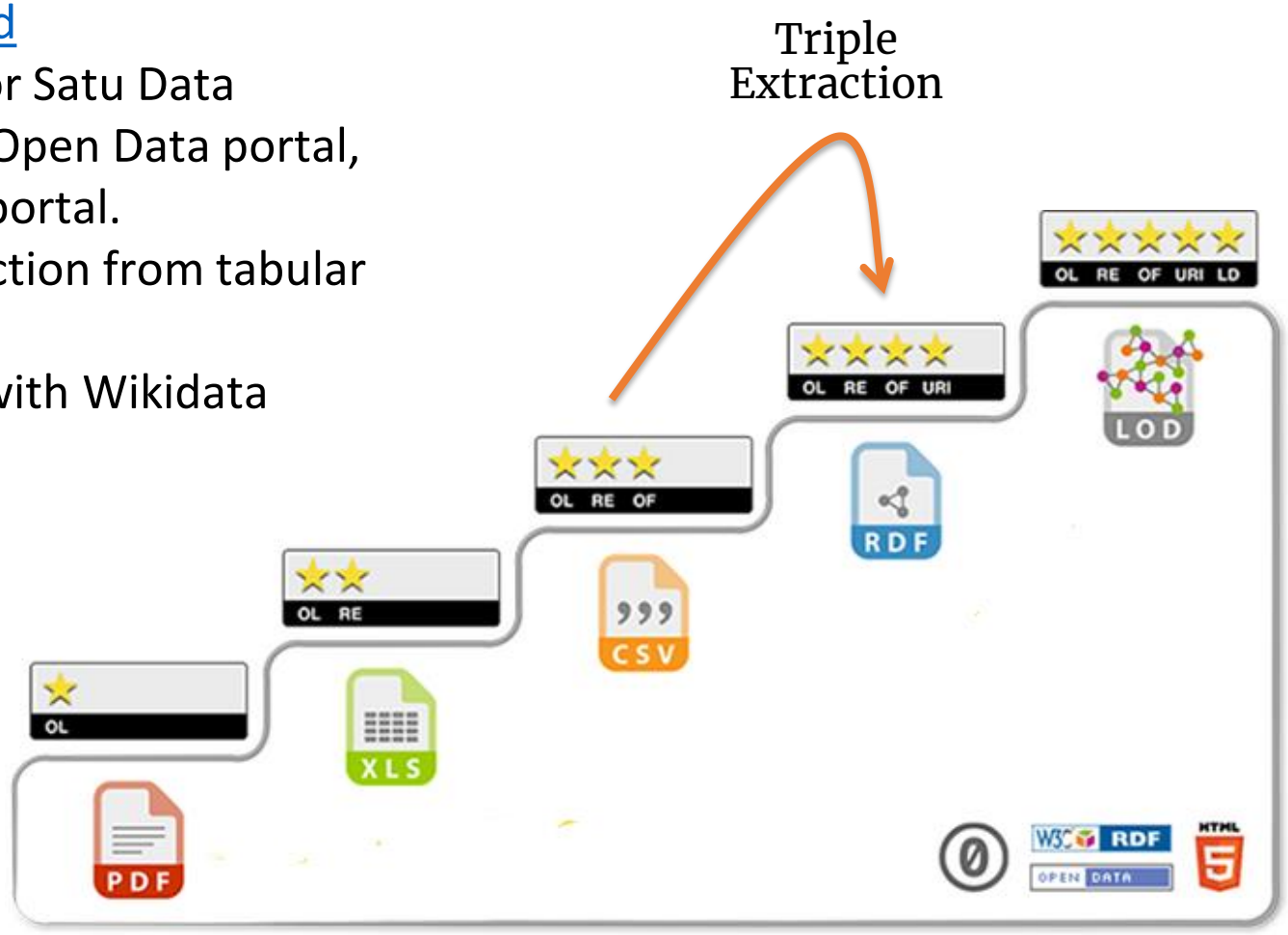
OPEN DATA KOTA BANDUNG

JAKARTA open data
Berbagi Data untuk Transparansi

WIKIDATA

OD2WD: Open Data to Wikidata

- Online at: http://od2wd.id
- Currently implemented for Satu Data Indonesia portal, Jakarta Open Data portal, and Bandung Open Data portal.
- Challenge #1: triple extraction from tabular cell values
- Challenge #2: alignment with Wikidata vocabulary



WIKIDATA

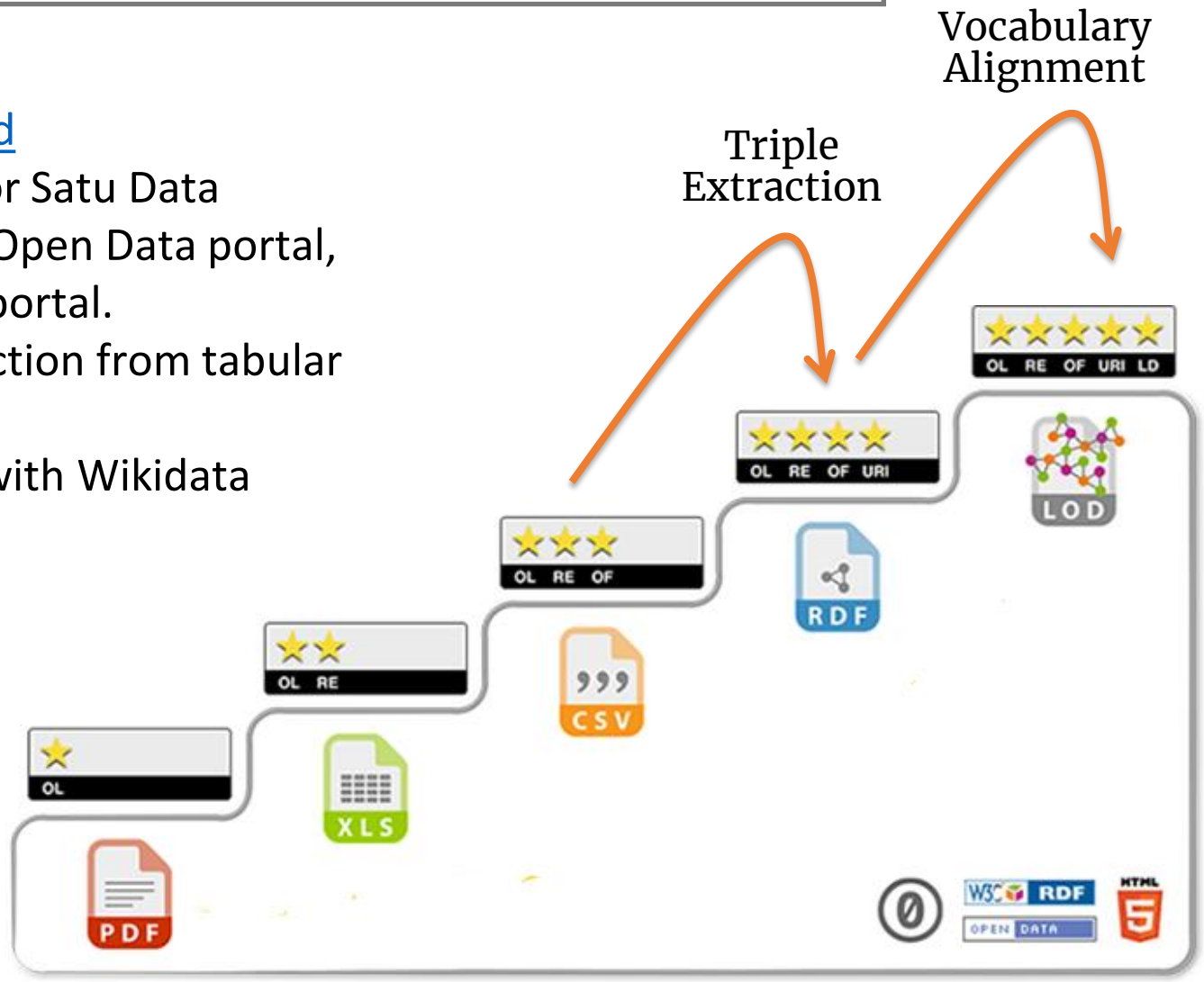
SATU DATA INDONESIA

OPEN DATA KOTA BANDUNG

JAKARTA open data
Berbagi Data untuk Transparansi

OD2WD: Open Data to Wikidata

- Online at: http://od2wd.id
- Currently implemented for Satu Data Indonesia portal, Jakarta Open Data portal, and Bandung Open Data portal.
- Challenge #1: triple extraction from tabular cell values
- Challenge #2: alignment with Wikidata vocabulary



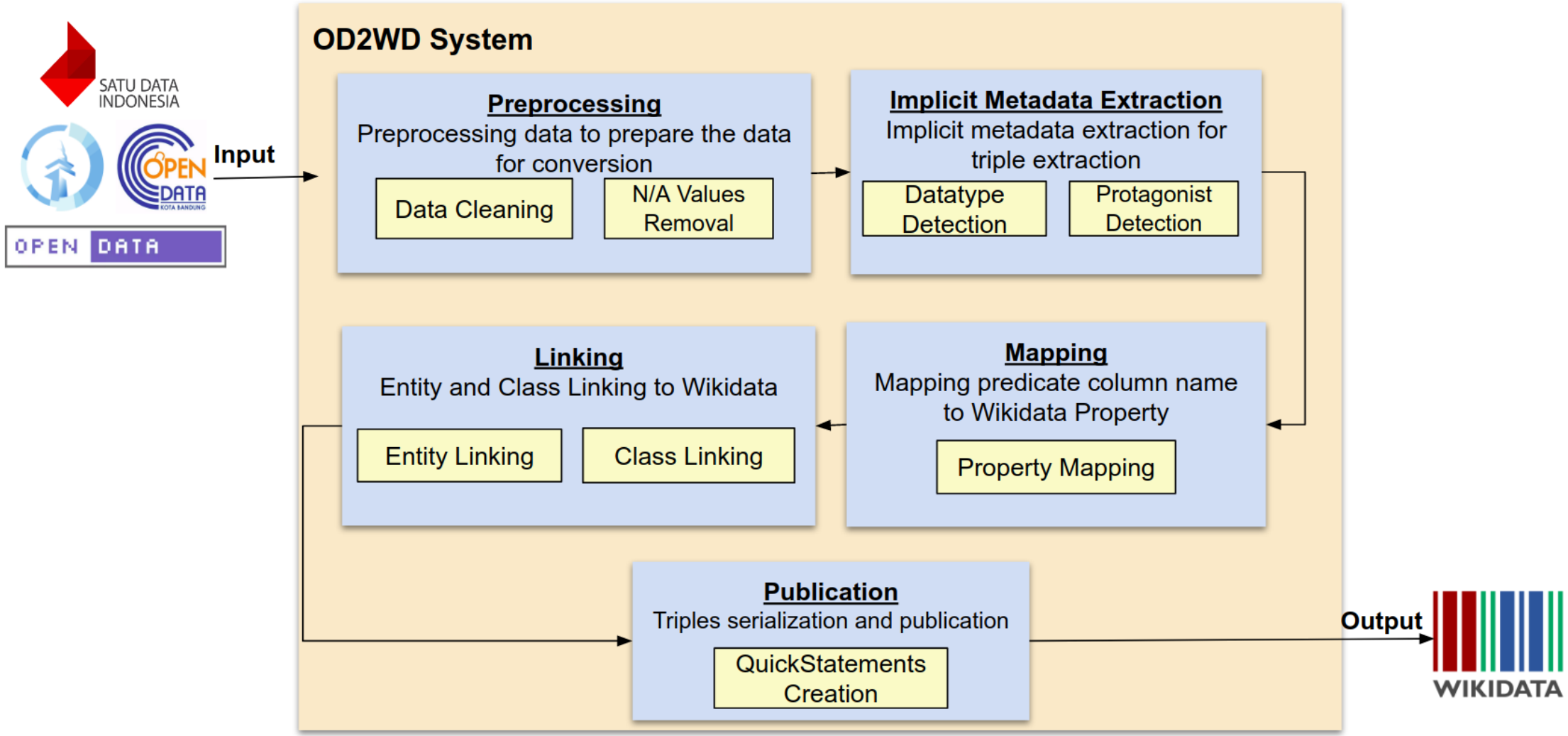
WIKIDATA

SATU DATA INDONESIA

OPEN DATA KOTA BANDUNG

JAKARTA open data
Berbagi Data untuk Transparansi

OD2WD Architecture



Reengineering Pattern

Given: Schema tuple $T = (C_1, \ldots, C_m)$, $t = (c_1, \ldots, c_m)$ is a row in table T, and for a particular value of k, $1 \leq k \leq m$, $C_k = Prot(T)$, the protagonist of T.

Generate: Graph with the following form (written in Turtle syntax):

$$\texttt{Subj rdf:type Cls};$$
$$\texttt{Pred}_1\ \texttt{Obj}_1\ ;\ \ldots\ ;\ \texttt{Pred}_{k-1}\ \texttt{Obj}_{k-1}\ ;\ \texttt{Pred}_{k+1}\ \texttt{Obj}_{k+1}\ ;\ \ldots\ ;\ \texttt{Pred}_m\ \texttt{Obj}_m.$$

where for $1 \leq j \leq m$, $j \neq k$, we have:

- $\texttt{Subj} = LinkRes(c_k)$, the Wikidata entity corresponding to c_k according to AP2,
- $\texttt{Obj}_j = LinkRes(c_j)$, the Wikidata entity corresponding to c_j according to AP2;
- $\texttt{Pred}_j = MapRes(C_j)$, the Wikidata property corresponding to column header C_j according to AP1;
- $\texttt{Cls} = ClassRes(C_k)$, the Wikidata class corresponding to the protagonist column header C_k according to AP3

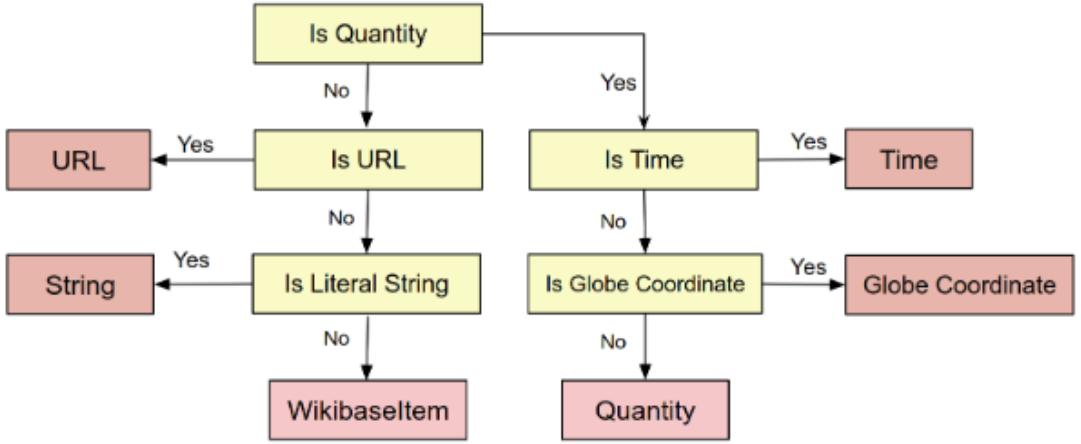
- Currently only handling vertical listing tables.
- Other table types are left as future work, e.g., horizontal listings, enumeration, matrix.
- Protagonist column: the one with the highest number of unique cell values, with leftmost position winning the tiebreaker.

Datatype Detection

Given: A column header C_j of table T containing N rows and $c_j^{(i)}$, $1 \leq i \leq N$ are N (not necessarily unique) values from each row of T at the j-th column.

Generate: A Wikidata datatype `dt` for column C_j if the majority of $c_j^{(i)}$'s satisfy the datatype pattern `dtp` defined as a Boolean combination formed by the four regular expression patterns (Quantity, URL, Literal String, Date, and Globe Coordinate defined by the table below) according to the following conditions:

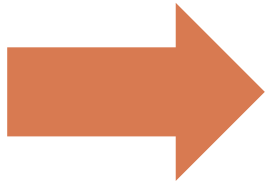
- `dt` is WikibaseItem when `dtp` is neither Quantity, URL, nor Literal String;
- `dt` is String when `dtp` is neither Quantity nor URL, but is Literal String;
- `dt` is URL when `dtp` is URL, but not Quantity;
- `dt` is Quantity when `dtp` is Quantity, but not Date and not Globe Coordinate;
- `dt` is Globe Coordinate when `dtp` is Quantity, not Date and is Globe Coordinate;
- `dt` is Time when `dtp` is Quantity and Date.



Pattern Name	Regular Expression
Quantity	`[-+.,()0-9]+`
Time	`^([0-2][0-9]|(3)[0-1])([\/,-])(((0)[0-9])|((1)[0-2]))([\/,-])\d{4}$`
Globe Coordinate	`^[-+]?([1-8]?\d(\.\d+)?|90(\.0+)?),\s*[-+]?(180(\.0+)?|((1[0-7]\d)|([1-9]?\d))(\.\d+)?)$`
URL	`^[a-zA-Z0-9_\-\@]+\.[a-zA-Z0-9]_\-\.`
Literal String	`[\.\,\!\?\>\<\/\\\)\(\-\_\+\=\*\&\^\%\$\#\@\!\:\;\~]`

Mapping/Linking: Disambiguation Challenge

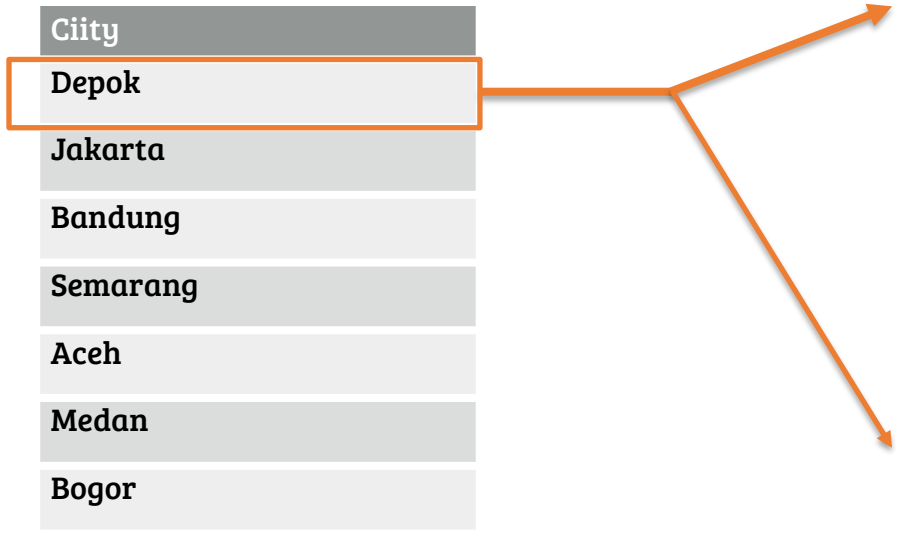
Ciity
Depok
Jakarta
Bandung
Semarang
Aceh
Medan
Bogor



Sumber: (https://wikidata.org)

Mapping/Linking: Disambiguation Challenge

Ciity
Depok
Jakarta
Bandung
Semarang
Aceh
Medan
Bogor



Depok (Q10396)

city in West Java Province, Indonesia

In more languages Configure

Language	Label
English	Depok
Indonesian	Depok
Javanese	Kutha Depok
Sundanese	Kota Dépok

Depok (Q757131)

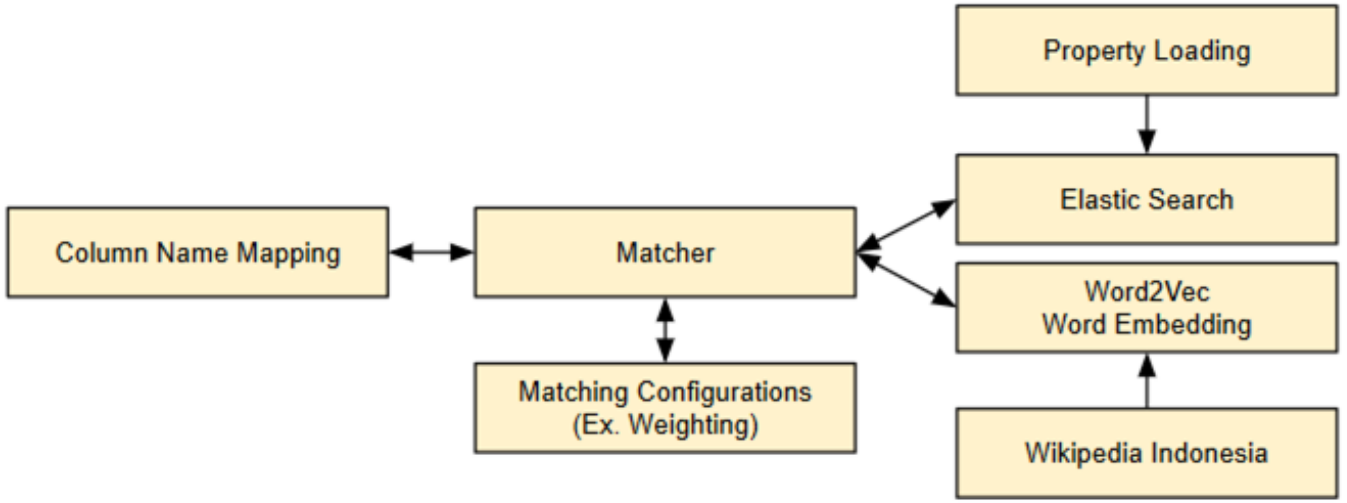
district in Cirebon Regency, Jawa Barat Province, Indonesia

In more languages Configure

Language	Label
English	Depok
Indonesian	Depok
Javanese	Depok
Sundanese	Dépok

Wikidata Allignment

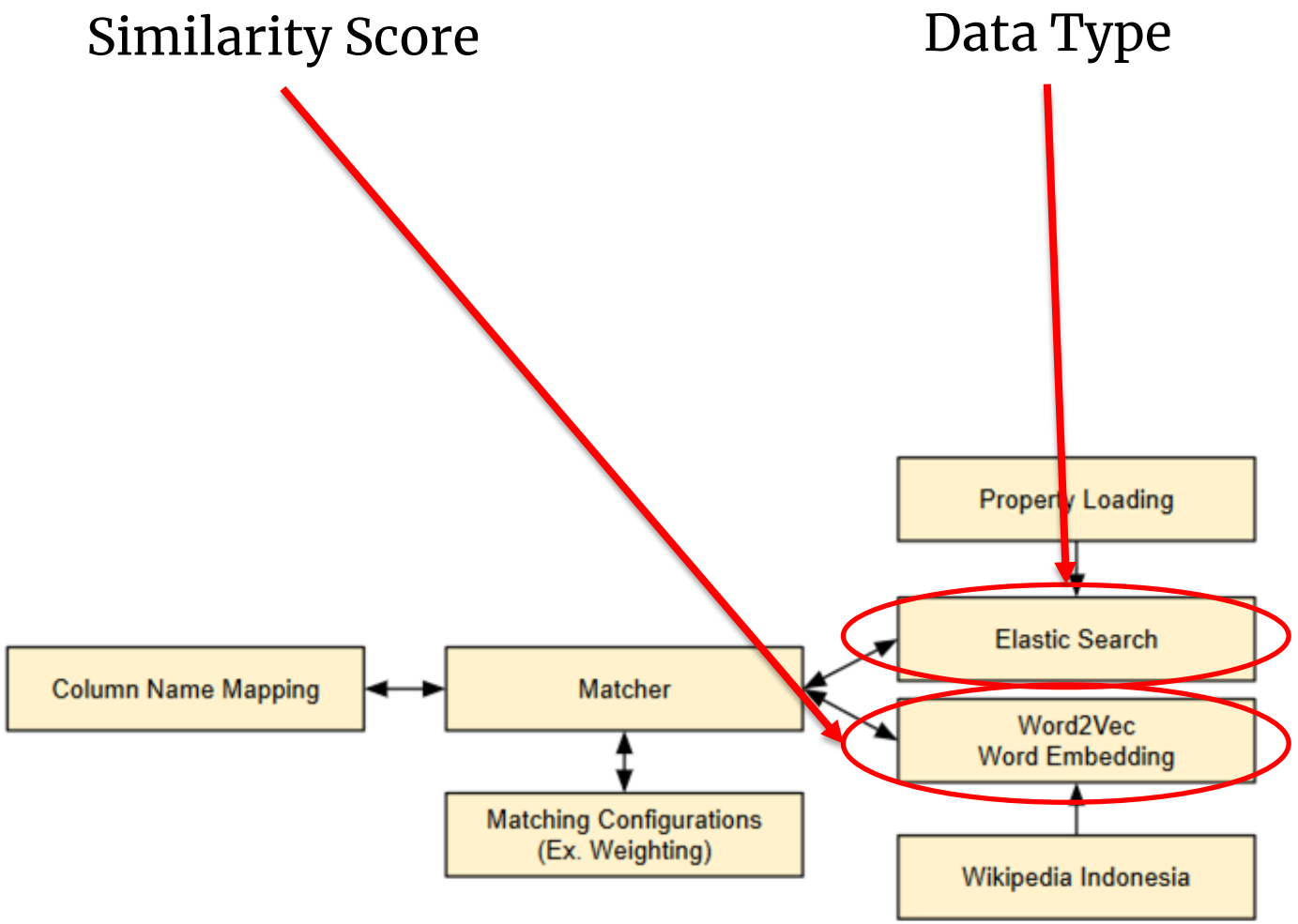
Mapping



Wikidata Allignment

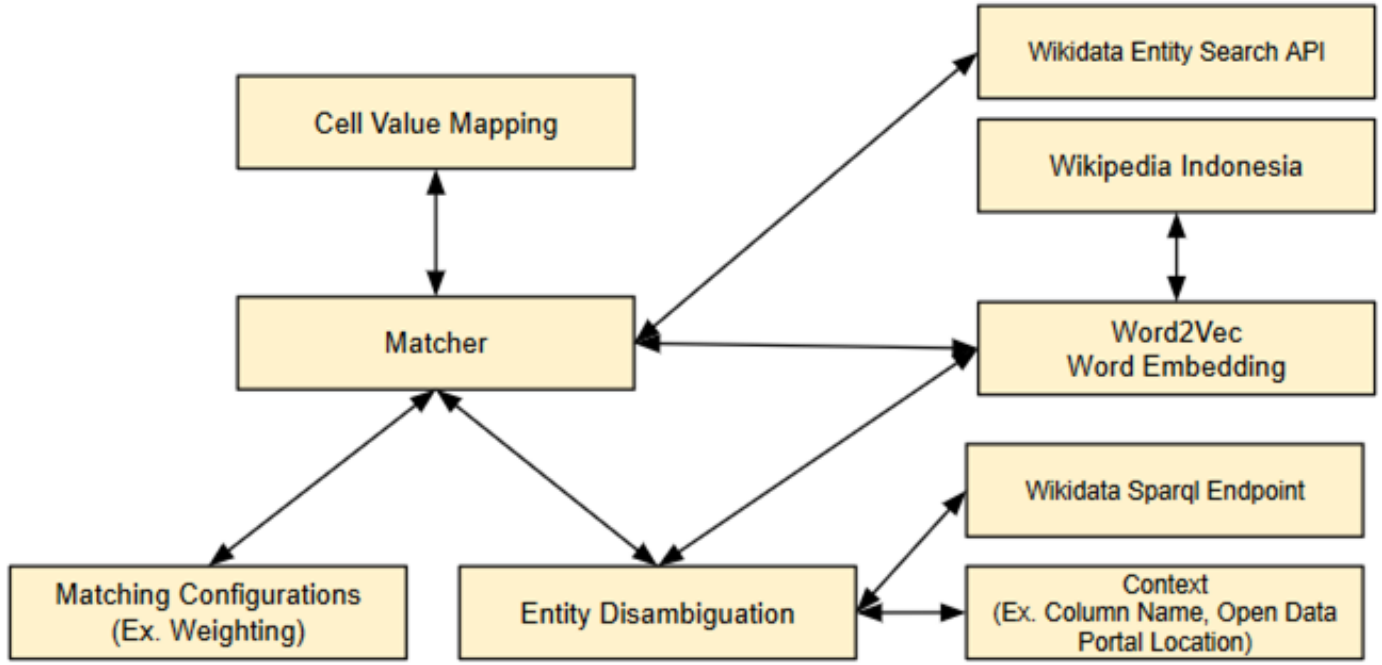
Mapping

Disambiguation



Wikidata Allignment

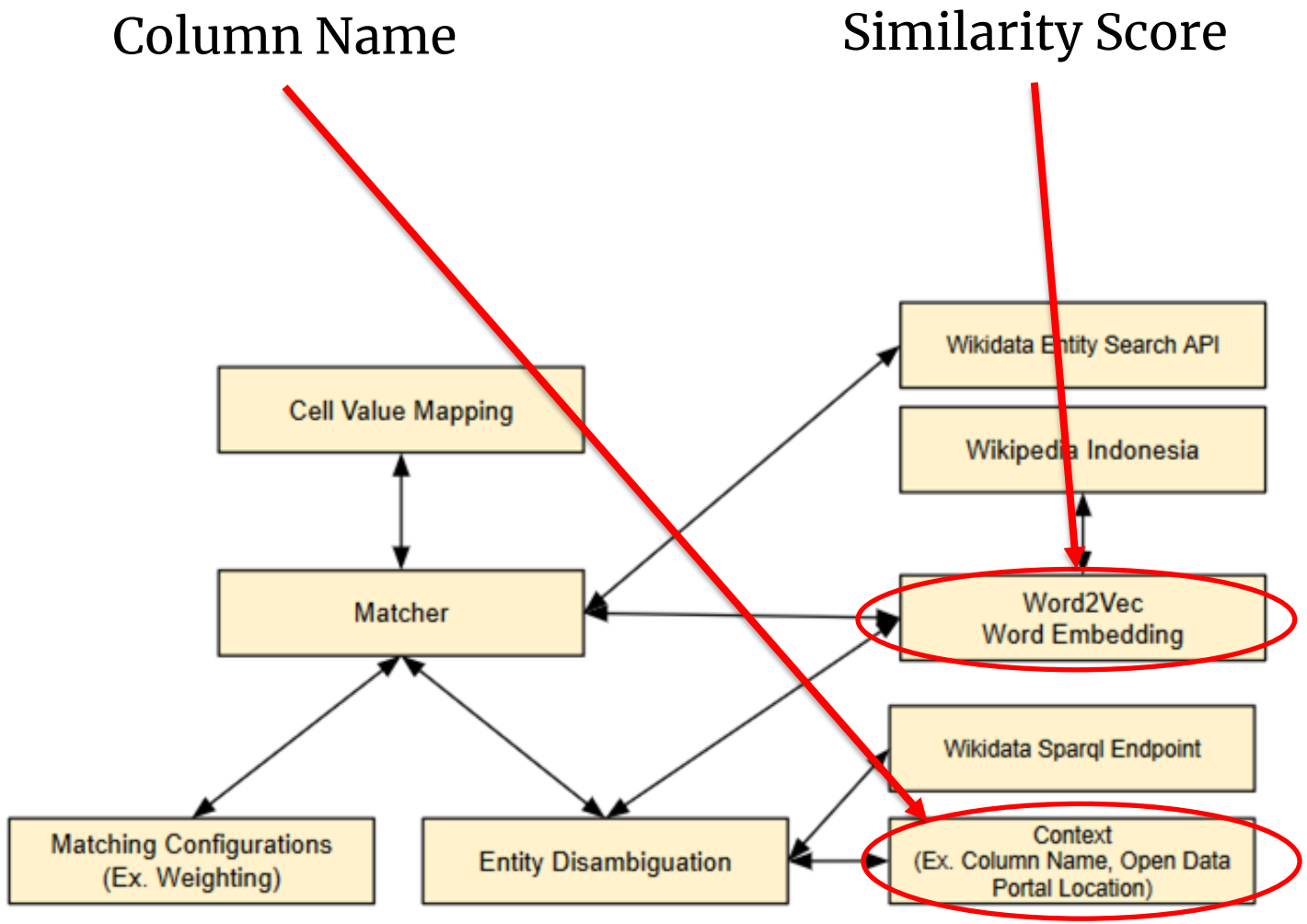
Entity Linking



Wikidata Allignment

Entity Linking

Disambiguation



Wikidata Allignment

Context in Entity Linking

Kelurahan
Kalisari
Wijaya Kusuma
Cengkareng Barat
Cipinang Cempedak
Kelapa Gading Barat
Slipi
Krukut

Kalisari (Q12488768)

urban community in Jakarta Timur, Indonesia

In more languages Configure

Language	Label
English	Kalisari
Indonesian	Kalisari
Javanese	No label defined
Sundanese	Kalisari, Pasar Rebo, Jakarta Timur

Kalisari (Q10947624)

village in Batang Regency, Indonesia

In more languages Configure

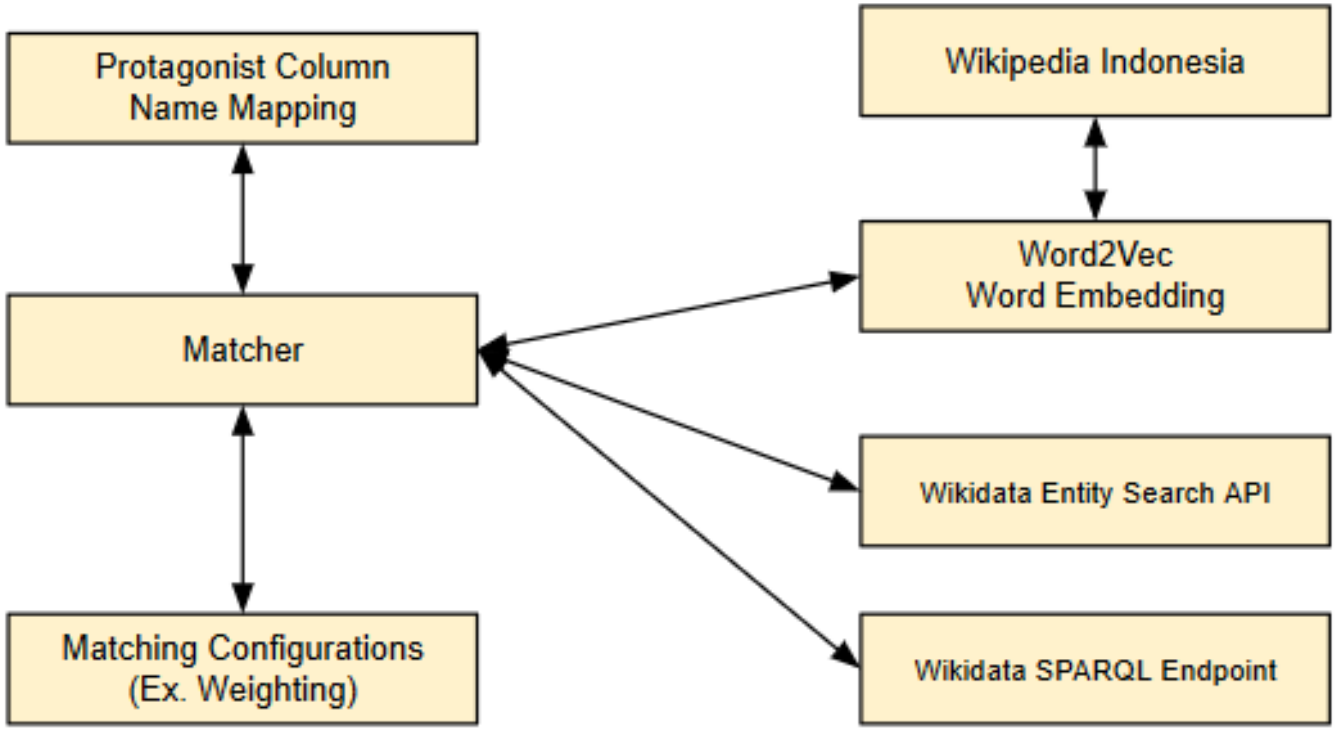
Language	Label
English	Kalisari
Indonesian	Kalisari
Javanese	Kalisari
Sundanese	Kalisari, Reban, Batang

Source: (https://wikidata.org)

```
SELECT ?item ?itemLabel
WHERE
{
    wd:X wdt:P31 ?item .
    SERVICE wikibase:label { bd:serviceParam wikibase:language "id" }
}
```

Wikidata Allignment

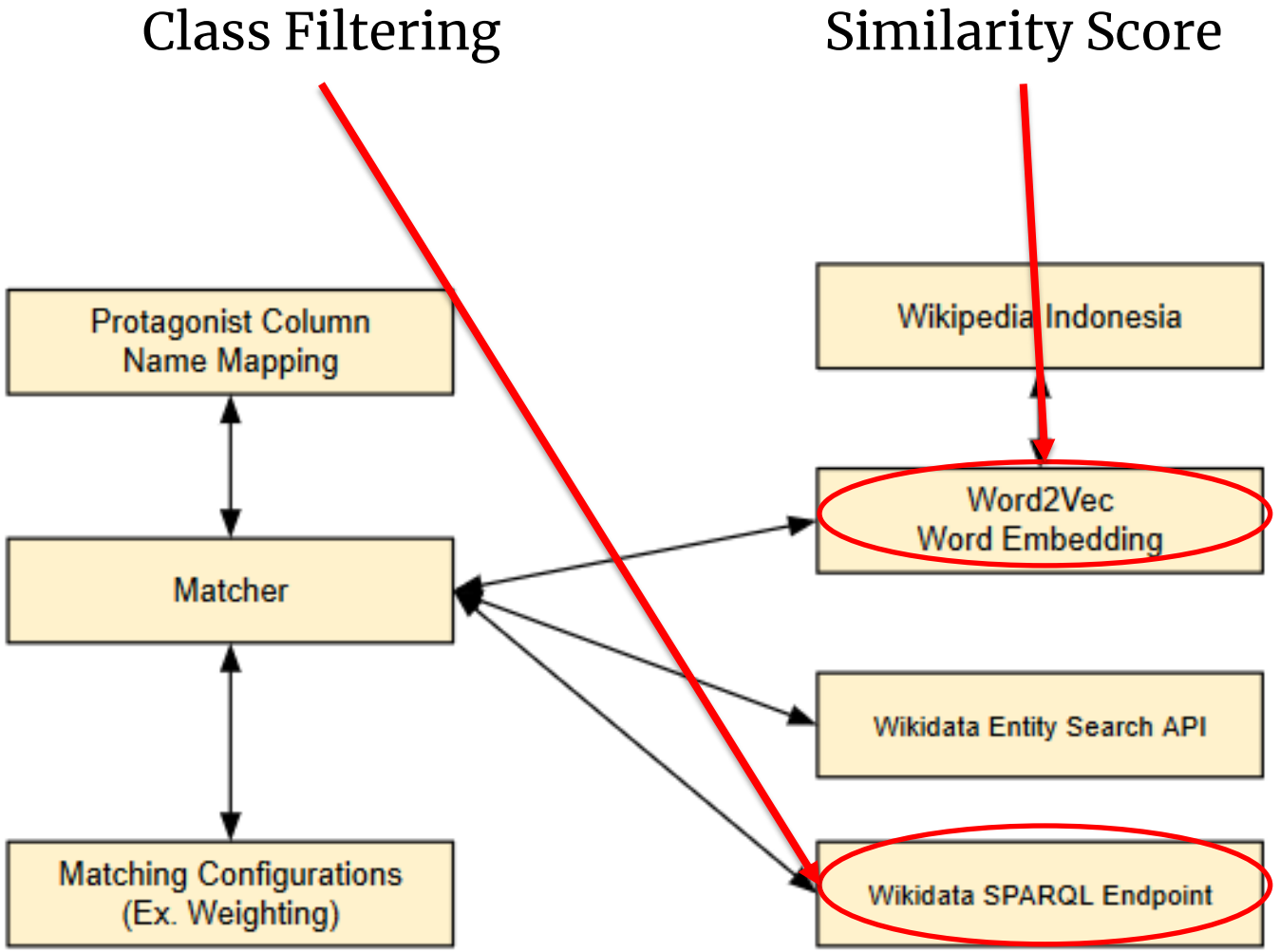
Class Linking



Wikidata Allignment

Class Linking

Disambiguation



Alignment Patterns

AP1: applied to non-protagonist column headers

```
_:link1 od2wd-prop:type skos:broadMatch ;
        od2wd-prop:from "ColName" ;
        od2wd-prop:to   wdt:Y ;
        od2wd-prop:confidence "Num"^^xsd:decimal ;
        od2wd-prop:generated_from od2wd:od2wdapi ;
        od2wd-prop:when "Time"^^xsd:dateTime .
```

AP1: applied to protagonist column headers

```
_:link1 od2wd-prop:type skos:closeMatch ;
        od2wd-prop:from "ColName" ;
        od2wd-prop:to   wd:Y ;
        od2wd-prop:confidence "Num"^^xsd:decimal ;
        od2wd-prop:generated_from od2wd:od2wdapi ;
        od2wd-prop:when "Time"^^xsd:dateTime .
```

AP2: applied to cell values

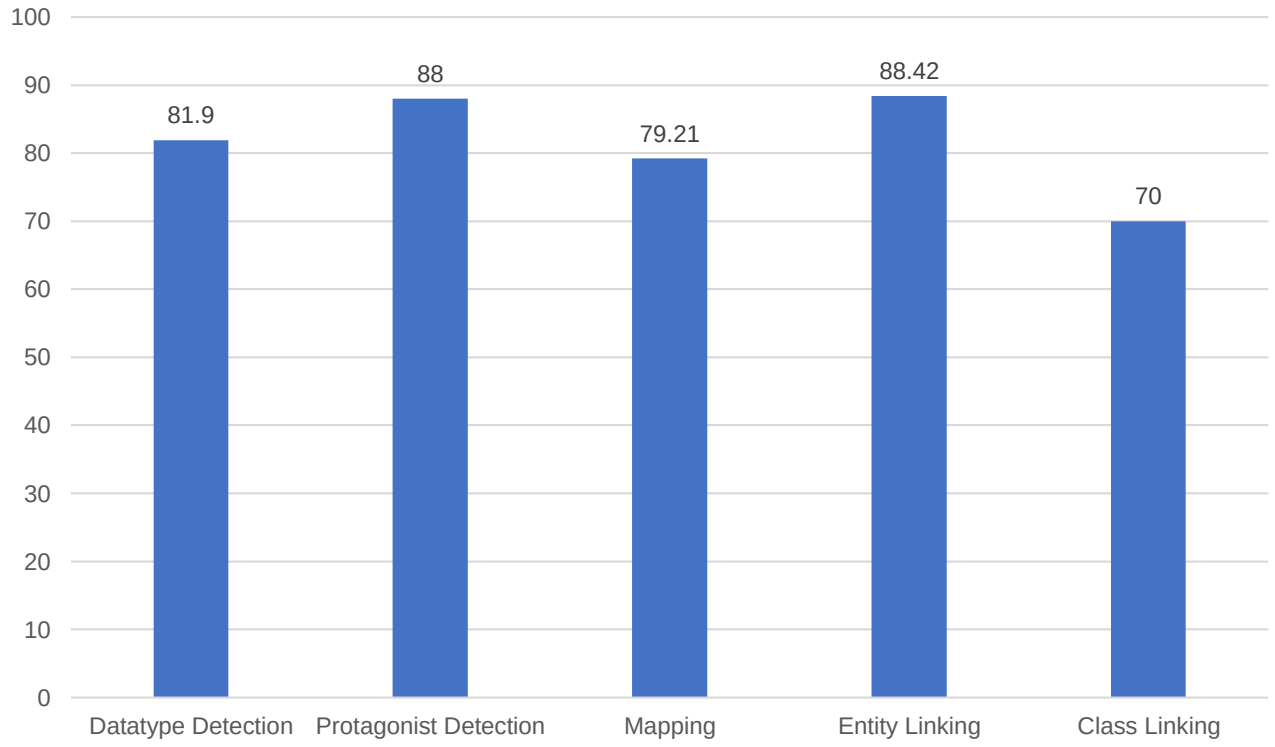
```
_:link1 od2wd-prop:type skos:closeMatch ;
        od2wd-prop:from "EntityName" ;
        od2wd-prop:to   wd:Y ;
        od2wd-prop:confidence "Num"^^xsd:decimal ;
        od2wd-prop:generated_from od2wd:od2wdapi ;
        od2wd-prop:when "Time"^^xsd:dateTime .
```

Conversion Accuracy

Performance measurement on 50 CSV documents from Indonesia's open data portal (compared against human judgement)

20256 new statements has been added to Wikidata

Below is a chart describing the accuracy of each conversion phase. Inaccuracy causes: value irregularity, nested structure (minority), inadequate corpus coverage for embedding



Future Work

Prototypical tool for converting tabular CSVs to RDF graphs and republish them to Wikidata.

Improvement on conversion accuracy by incorporating more context information

Handling more types of tables: horizontal listings, enumeration, matrix, etc.

Study better encoding of the patterns and their applicability and usage in other open data portals

Acknowledgement

2019 PITTA B research grant
“Analysis and Enrichment of Wikidata Knowledge Graph" from Universitas Indonesia

Wikimedia Indonesia project “Peningkatan Konten Wikidata."

Students at Universitas Indonesia as human evaluators

Raisha Abdillah from Wikimedia Indonesia for final quality checks prior to deploying data to Wikidata

Video demo: https://youtu.be/oOjJdOQ8dwM

Thank You