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
• 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
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
• 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
OD2WD: Open Data to Wikidata

- Online at: [http://od2wd.id](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
OD2WD Architecture

OD2WD System

Preprocessing
Preprocessing data to prepare the data for conversion
- Data Cleaning
- N/A Values Removal

Implicit Metadata Extraction
Implicit metadata extraction for triple extraction
- Datatype Detection
- Protagonist Detection

Linking
Entity and Class Linking to Wikidata
- Entity Linking
- Class Linking

Mapping
Mapping predicate column name to Wikidata Property
- Property Mapping

Publication
Triples serialization and publication
- QuickStatements Creation

Input
- Open Data Indonesia
- Open Data

Output
- Wikidata
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 $k$, $1 \leq k \leq m$, $C_k = \text{Prot}(T)$, the protagonist of $T$.

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

```
Subj rdf:type Cls;
    Pred_1 Obj_1; \ldots; Pred_{k-1} Obj_{k-1}; Pred_{k+1} Obj_{k+1}; \ldots; Pred_m Obj_m.
```

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

- $\text{Subj} = \text{LinkRes}(c_k)$, the Wikidata entity corresponding to $c_k$ according to AP2,
- $\text{Obj}_j = \text{LinkRes}(c_j)$, the Wikidata entity corresponding to $c_j$ according to AP2;
- $\text{Pred}_j = \text{MapRes}(C_j)$, the Wikidata property corresponding to column header $C_j$ according to AP1;
- $\text{Cls} = \text{ClassRes}(C_k)$, 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.
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 $T$ in 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]+` 
| | `^([0-2][0-9]|(3)[0-1])([/,-])(0)[0-9])|((1)[0-2])$` 
| Time | `^[+-]?([1-8]?\d(\d+)\d+)\d(\d+)?90(\d)+,\d+],[-+]?(180(\d)+)?$` 
| Globe Coordinate | `^[+-]?([1-8]?\d(\d+)\d+)\d(\d+)?90(\d)+,\d+],[-+]?(180(\d)+)?$` 
| URL | `^[a-zA-Z0-9\-\@]+\.[a-zA-Z0-9\-\@]+$` 
| Literal String | `[\.,!?$#'\%\&\*\-\(|\]`
Mapping/Linking: Disambiguation Challenge

City
- Depok
- Jakarta
- Bandung
- Semarang
- Aceh
- Medan
- Bogor

WIKIDATA
Sumber: (https://wikidata.org)
Mapping/Linking: Disambiguation Challenge

Ciity
- Depok
- Jakarta
- Bandung
- Semarang
- Aceh
- Medan
- Bogor

Sumber: (https://wikidata.org)
Wikidata Allignment Mapping
Wikidata Allignment

Mapping

Disambiguation

Similarity Score

Data Type

Property Loading

Elastic Search

Word2Vec Word Embedding

Matching Configurations (Ex. Weighting)

Wikipedia Indonesia

Column Name Mapping

Matcher
Wikidata Allignment

Entity Linking
Wikidata Alignment

Entity Linking
### Wikidata Alignment

Context in Entity Linking

<table>
<thead>
<tr>
<th>Kelurahan</th>
<th>Kulisari</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wijaya Kusuma</td>
<td></td>
</tr>
<tr>
<td>Cengkareng Barat</td>
<td></td>
</tr>
<tr>
<td>Cipinang Cempedak</td>
<td></td>
</tr>
<tr>
<td>Kelapa Gading Barat</td>
<td></td>
</tr>
<tr>
<td>Slipi</td>
<td></td>
</tr>
<tr>
<td>Krukut</td>
<td></td>
</tr>
</tbody>
</table>

**Source:** (https://wikidata.org)

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

Class Linking
 Wikidata Alignment

Class Linking

Disambiguation

Class Filtering

Similarity Score

Protagonist Column Name Mapping

Matcher

Matching Configurations (Ex. Weighting)

Wikipedia Indonesia

Word2Vec Word Embedding

Wikidata Entity Search API

Wikidata SPARQL Endpoint
Alignment Patterns

AP1: applied to non-protagonist column headers

```plaintext
_: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 .
```

AP2: applied to cell values

```plaintext
_:link1 od2wd-prop:type skos:closeMatch ;
  od2wd-prop:from "EntityName" ;
  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

```plaintext
_:link1 od2wd-prop:type skos:closeMatch ;
  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 .
```
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](https://youtu.be/oOjJdOQ8dwM)

Thank You