BudayaKB: Extraction of Cultural Heritage Entities from Heterogeneous Formats

Hadi Syah Putra
Universitas Indonesia
Depok, Indonesia
hadi.putra@bukalapak.com

Rahmad Mahendra
Universitas Indonesia
Depok, Indonesia
rahmad.mahendra@cs.ui.ac.id

Fariz Darari
Universitas Indonesia
Depok, Indonesia
fariz@cs.ui.ac.id

ABSTRACT

Cultural heritage is the root of national identities, and contributes to tourism, economics, industry, and business. Digital preservation of cultural heritage is therefore crucial, particularly in a form that is easily processable by machines. Available cultural heritage information on Web sources (e.g., Wikipedia) is presented in multiple formats, such as free-form text, lists, and tables. Such formats, however, lack structures and links to other information sources. The provision of cultural heritage information as a knowledge base, that is both structured and linked, would pave new ways for consuming such information. In this paper, we propose an approach to extract entities of cultural heritage from diverse formats (i.e., text, lists, and tables), and to construct a knowledge base of cultural heritage entities, called BudayaKB, using RDF data model that provides an integrated, format-independent view. Our extraction approach follows on the observation that cultural heritage entities are often written down either with common noun descriptors (e.g., `Jiwa temple`) or hypernym-hyponym sentence patterns (e.g., ...Acehnese `traditional weapons` such as `Rencong`...). We evaluate our approach to Indonesian Wikipedia, and achieve a precision of 84% for extracting Indonesian cultural heritage entities. The extracted entities are then imported and linked to the Wikidata KB, allowing greater interoperability of cultural heritage information. BudayaKB is openly available at https://budayakb.cs.ui.ac.id/.

CCS CONCEPTS

• **Information systems → Data extraction and integration**; **Resource Description Framework (RDF)**; **Information extraction**.

KEYWORDS

Cultural heritage, Entity extraction, Knowledge base

ACM Reference Format:
Hadi Syah Putra, Rahmad Mahendra, and Fariz Darari. 2019. BudayaKB: Extraction of Cultural Heritage Entities from Heterogeneous Formats. In *9th International Conference on Web Intelligence, Mining and Semantics (WIMS2019), June 26–28, 2019, Seoul, Republic of Korea.* ACM, Seoul, Republic of Korea, 9 pages. https://doi.org/10.1145/3326467.3326487

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
WIMS2019, June 26–28, 2019, Seoul, Republic of Korea
© 2019 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-6190-3/19/06…$15.00
https://doi.org/10.1145/3326467.3326487

1 INTRODUCTION

Cultural heritage is described by UNESCO as, "the legacy of physical artifacts and intangible attributes of a group or society that are inherited from past generations, maintained in the present, and bestowed for the benefit of future generations" [25]. Cultural heritage reflects the identity of past social life [1], and can be classified into two broad categories: tangible heritage and intangible heritage. While tangible heritage can be seen and touched, such as temples, traditional weapons, and foods, intangible heritage is more about traditions, folktales, and performing arts, that are inherited from generation to generation in a community.

Indonesia is a country rich in cultural heritage. There are more than 300 ethnic groups in Indonesia, each of which has its own distinct culture not only in the aspects of languages, but also in traditional songs, foods, dances, and so forth [2]. Preserving such cultural heritage is therefore important, where one of the key preservation efforts is by leveraging digital approaches [1]. Current Web sources attempting to record information about Indonesian cultural heritage are, however, limited. Sources such as Wikipedia and government/tourism Web sites often provide cultural heritage information in the form of (unstructured) text, hindering the automatic processing of the information. On the other hand, Web knowledge bases (KBs) like Wikidata[1] and DBpedia,[2] though storing information about cultural heritage in a structured manner, suffer from data incompleteness. This opens up the question as to, *how to bridge the gap between information about cultural heritage in text-oriented Web sources such as Wikipedia, to structured Web sources such as Wikidata?*

One important challenge in answering the question is the *diversity of formats* in presenting (cultural heritage) information. Wikipedia, one of the largest encyclopediae, stores information in three different formats: free-form text, tables, and lists. Devising an approach that only considers a specific format would suffer from the limited recall in the amount of extracted knowledge. Another challenge is that, how can the extracted knowledge of cultural heritage entities be *aligned with external KBs* (e.g., Wikidata)? This challenge is not only about providing links to existing entities in the external KBs, but also adding new entities to the external KBs when necessary (that is, when they do not yet exist in there).

To address those challenges, we develop the BudayaKB[3] approach, which *extracts* knowledge about cultural heritage entities available in heterogeneous formats. We build an extraction workflow that utilizes the observation that cultural heritage entities are presented either with common noun descriptors (e.g., `Jiwa temple`)

[1] https://www.wikidata.org/
[2] https://dbpedia.org
[3] The Indonesian word *budaya* means *culture* in English.

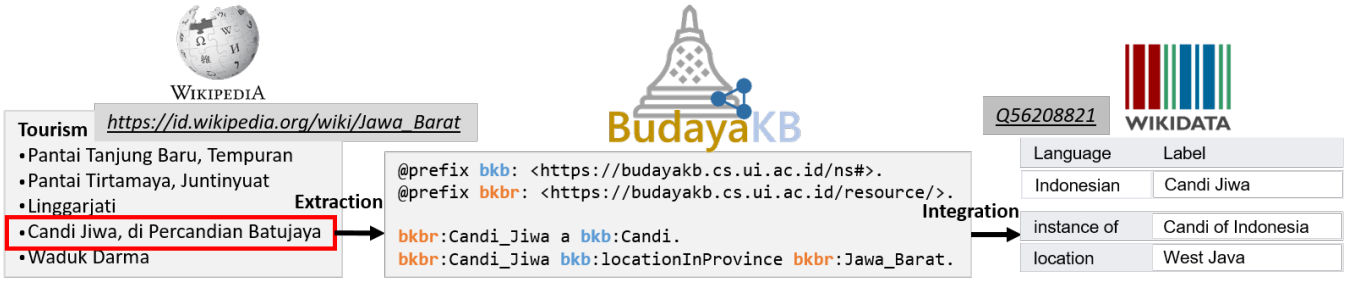

Figure 1: Extracting Jiwa temple (in Indonesian, Candi Jiwa) from the Indonesian Wikipedia page about West Java (in Indonesian, Jawa Barat), and integrating it into Wikidata

or hypernym-hyponym patterns (e.g., ...Acehnese `traditional weapons` such as `Rencong`...). We also *integrate* the extracted cultural heritage entities with the Wikidata KB, one of the central, open KBs on the Web,[4] enabling greater interoperability across data sources. Our BudayaKB approach is implemented and evaluated over Indonesian Wikipedia. Figure 1 shows the big picture of our approach in extracting and integrating cultural heritage information. It illustrates how `Jiwa temple` (ID: `Candi Jiwa`) can be extracted from a list in the Wikipedia page of the West Java province (ID: Jawa Barat) and then integrated into Wikidata.

The rest of this paper is structured as follows. Section 2 presents the background of our approach. Section 3 describes the BudayaKB approach and its implementation to extract and integrate cultural heritage entities. We report on an evaluation of our approach in Section 4 and explore use cases in Section 5. Section 6 discusses related work and Section 7 concludes this paper.

2 BACKGROUND

In this section, we discuss the background material for our BudayaKB approach which comprises cultural heritage taxonomy, content formats in Wikipedia, information extraction, and Semantic Web technologies.

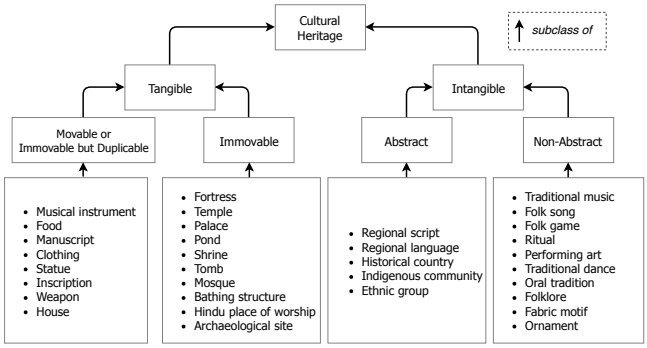

Figure 2: Cultural Heritage Taxonomy

2.1 Cultural Heritage Taxonomy

Cultural heritage can be classified into tangible and intangible heritage. This broad classification can be refined further by synthesizing the cultural heritage categorization in the e-Indonesiana portal,[5] and in the work of [10, 21], resulting in the taxonomy as displayed in Figure 2. The taxonomy breaks down tangible cultural heritage into movable or immovable but duplicable, and immovable heritage. Moreover, intangible heritage can be classified into abstract and non-abstract heritage. The leaves of the taxonomy are cultural heritage objects, characterized by their respective (direct and indirect) superclasses. For example, mosques are immovable and tangible cultural heritages, whereas languages are abstract and intangible cultural heritages.

2.2 Wikipedia Content Formats

Information content in Wikipedia is presented in heterogeneous formats. Even though the majority of the content takes the free-text form, a substantial amount of the content is also displayed in the form of lists and tables. Consequently, different treatments for processing those formats are required. In our approach, we do not look at the wiki markup,[6] which is a formatting language specifically made for editing Wikipedia pages. Instead, we take the HTML rendering of the wiki markup. This ensures the adaptability of our approach to more generic, non-wiki Web pages.

2.2.1 Free-form Text. Free-form text is the most common format used in Wikipedia pages. It is the least restrictive format compared to lists and tables. In HTML, text is a collection of paragraphs, enclosed with `<p>` tags. Text from Wikipedia pages often contains hyperlinks, presented with `<a>` tags, and also decorators like `<sup>` tags for superscript mode.

2.2.2 List. Lists are often used for organizing information. In HTML, there are two different types of lists: unordered lists with `<ul>` tags, and ordered lists with `<ol>` tags. HTML lists contain list items, enclosed with `<li>` tags. Besides presenting the main information, lists in Wikipedia pages may serve different purposes, such as navigational (e.g., table of content), meta-content (e.g., footer, languages, additional information, gallery), and reference (e.g., external links, citations). Lists in Wikipedia are also commonly accompanied with

[4] https://lod-cloud.net/

[5] http://e-indonesiana.cs.ui.ac.id/

[6] https://en.wikipedia.org/wiki/Help:Wikitext

headings, serving as the list titles. Such headings occur before the list elements (i.e., before `<ul>` and `<ol>` tags), and are defined with the `<h2>` to `<h6>` tags. The `<h1>` tag is excluded since it is always designated for Wikipedia page titles.

2.2.3 Table. Tables present information in rows and columns. In a way, a table is a complex form of a list. Information in tables is put in a systematic grid pattern. Wikipedia tables are no different to regular HTML tables, with the only exception that they are marked with the `wikitable` class attribute in `<table>` tags. An HTML table consists of several rows, headers, and cells, marked with `<tr>`, `<th>`, and `<td>` tags, respectively.

2.3 Information Extraction

Information presented in Wikipedia, be it text, lists, or tables, contains knowledge that needs to be automatically extracted. Information Extraction (IE) deals with the extraction of meaning and structures from such formats. Typical tasks of information extraction include entity recognition and relationship extraction. Text through information extraction approaches is often analyzed with the help of tools such as tokenizer, part-of-speech (POS) tagger, and named entity (NE) recognizer [23].

IE Tools. A tokenizer typically segments text into words and sentences based on delimiters such as spaces, commas, dots, and punctuation marks. A POS tagger groups words in a sentence into corresponding word classes, such as noun, verb, adjective, conjunction, and others [4]. An NE recognizer aims to identify named entities of type person, place, and organization, in a text.

2.4 Semantic Web Technologies

The Semantic Web is an extension to the Web that allows computers to search, combine, and process Web-based content in an intelligent fashion [12]. The Semantic Web is not only about putting structured data on the Web, but also making links between data items. Such features are made available by means of URIs and RDF. A URI, which stands for Uniform Resource Identifier, is a global, unambiguous identifier to a particular resource, which can also be a real-world object. For example, the Jiwa temple can be identified by the URI http://www.wikidata.org/entity/Q56208821.[7] Furthermore, property is also a valid resource that can be identified by a URI, as exemplified by http://www.wikidata.org/prop/direct/P17 for the `country` property in Wikidata.

RDF, which stands for Resource Description Framework, is a standard data model for data interchange on the Web.[8] Information in RDF is organized into S-P-O triples, where S is for subject, P for predicate, and O for object. URIs can appear in any position in RDF triples, whereas literal values (e.g., strings, integers) are only allowed in the object position. A collection of RDF triples is called an RDF graph. In this paper, we refer to RDF graphs as knowledge bases (KBs). Creating links between equivalent data items across different KBs can be facilitated by using the "same as" property with the URI http://www.w3.org/2002/07/owl#sameAs. The middle part of Figure 1 shows two RDF triples about Jiwa temple, serialized in the Turtle format.[9] The first triple describes that Jiwa temple is of type Candi (in English, Temple), whereas the second triple describes the location of the temple.

3 THE BUDAYAKB APPROACH

The BudayaKB approach aims to address the problem of information gap about cultural heritage between text-oriented Web sources, such as Wikipedia, and structured Web sources, such as Wikidata. In this approach, we develop two modules: the extraction module and the integration module. The extraction module performs data extraction from Wikipedia to build BudayaKB, a knowledge base about Indonesian cultural heritage, while the integration module links BudayaKB to Wikidata.

3.1 Extraction Module

The extraction module leverages information extraction approaches over text-oriented Web sources. The module comprises three steps: lexicon building, heterogeneous data formats extraction, and knowledge base (KB) generation (see Figure 3). In the lexicon building step, we build two types of lexicons, that is, the hypernym and descriptor lexicons. The lexicons contain the category names as shown in Figure 2, plus additional sources. The heterogeneous data formats extraction step is based on information presentation of Web sources, and is divided into three independent processes: table, list, and free-form text extraction. Those processes use the lexicons built in the previous step to detect cultural heritage entities which are presented either with common noun descriptors or hypernym-hyponym patterns. The cultural heritage entities resulted in the extraction step are then transformed into RDF data model in the KB generation step to finally build BudayaKB.

In order to explain our approach in a more concrete way, we will now describe a scenario of the Web sources from which we perform information extraction. The sources of extraction are Indonesian Wikipedia articles about Indonesian provinces (34 articles) and cities (515 articles). To obtain the list of provinces and cities, we scrape these two articles: *Daftar provinsi di Indonesia*[10] (EN: List of provinces in Indonesia) and *Daftar kabupaten dan kota di Indonesia*[11] (EN: List of districts and cities in Indonesia), respectively.

3.1.1 Lexicon Building. There are two types of lexicons used to perform data extraction: descriptor lexicon and hypernym lexicon.

Descriptor Lexicon. The descriptor lexicon consists of cultural heritage common noun descriptors and is used to detect cultural heritage entities by their descriptor occurrences. A cultural heritage category may have one or more different descriptors. For example, ethnic group entities (e.g., `Acehnese`) in the text may be written either with the descriptor `tribe` or `ethnic`.

To detect cultural heritage entities, we can match their descriptors by using regular expressions generated from entries in the descriptor lexicon. If there is a compound noun having its descriptor matched with any descriptors in the lexicon, the compound noun will be extracted as a cultural heritage entity and be given a category according to the matched descriptor (see Figure 4). In the

[7] Wikidata provides internal identifiers for its resources. Identifiers for entities start with "Q", whereas identifiers for properties start with "P".
[8] https://www.w3.org/RDF/
[9] https://www.w3.org/TR/turtle/
[10] https://id.wikipedia.org/wiki/Daftar_provinsi_di_Indonesia
[11] https://id.wikipedia.org/wiki/Daftar_kabupaten_dan_kota_di_Indonesia

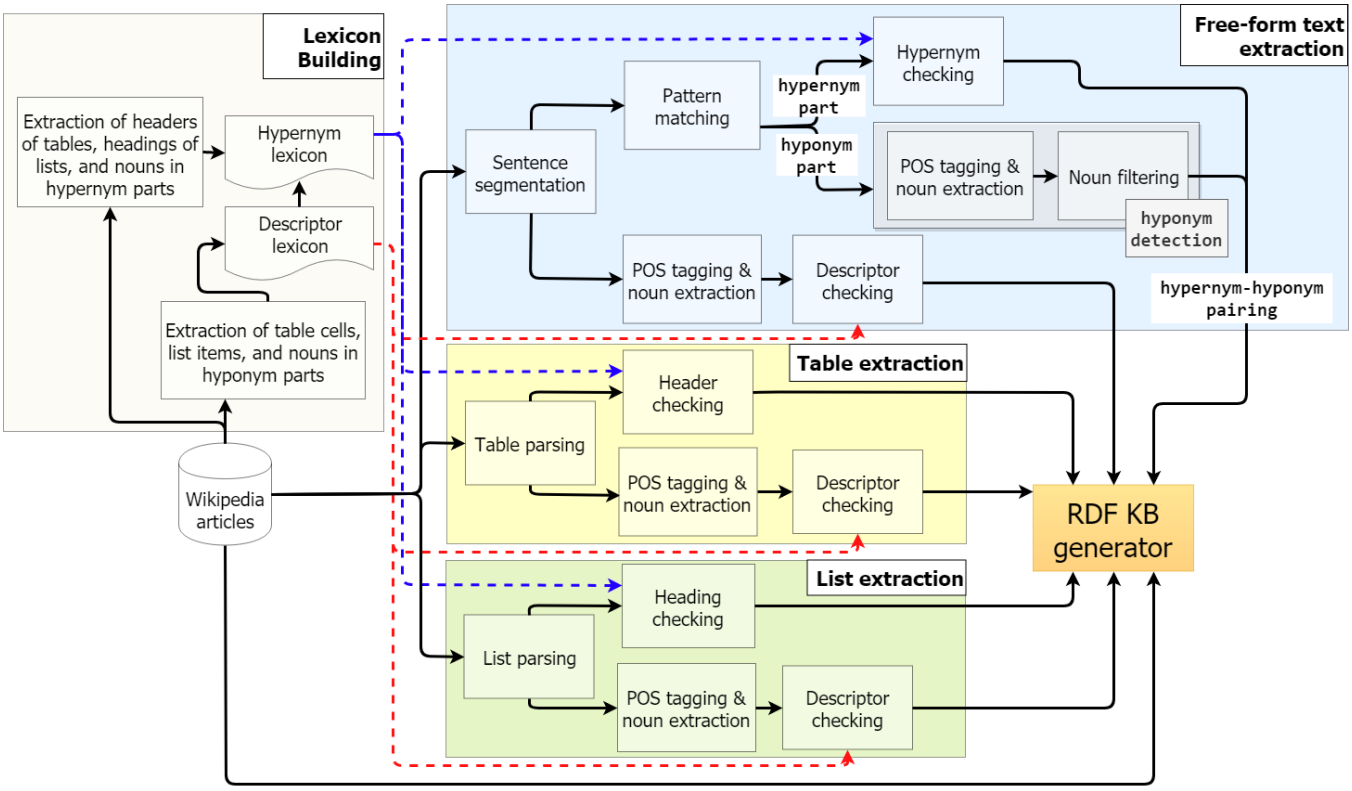

Figure 3: BudayaKB Data Extraction Flow

figure, the descriptor part of the compound noun `Acehnese tribe` matches with an entry in the descriptor lexicon. Hence, the entity will be extracted and stored as a pair (`ethnic group`, `Acehnese tribe`) where the first element is the culture heritage category and the second is the entity name.

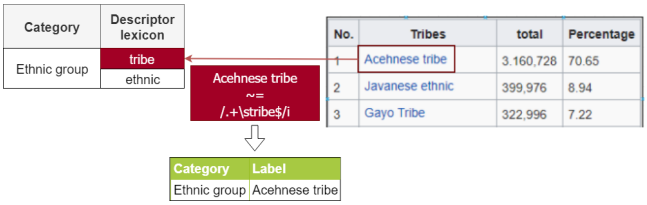

Figure 4: Example of extracting entities with descriptor lexicon

We manually enrich the descriptor lexicon by adding entries from: existing cultural heritage entities in DBpedia and Wikidata as well as in the e-Indonesiana portal, and extracted nouns from table cells, list items, and hypernym parts in sentences of Indonesian Wikipedia articles about Indonesian provinces and cities. For example, descriptors added for detecting traditional dance entities are `tari` and `tarian` (dance).[12] In total, the descriptor lexicon contains 62 descriptors.

[12]http://e-indonesiana.cs.ui.ac.id/echnh-ng/index.php/view/cat/1/0

Hypernym Lexicon. A hypernym-hyponym relation connects more general words with more specific words, or vice versa [13], e.g., Sundanese (hyponym) with language (hypernym). The earliest study about it is in Hearst's research [11]. She manually defined lexico-syntactic patterns (shown in Table 1) to discover hypernym-hyponym relations from large text English corpus. In practical settings, hypernym-hyponym relations can also take form as headers-cells in tables and headings-list items in lists.

The hypernym lexicon contains cultural heritage category-related keywords, useful for detecting cultural heritage entities by their categories. As a category and an entity have a hypernym-hyponym relation (and the other way around, an entity and a category have a hyponym-hypernym relation), a category-entity pair can be extracted by pattern matching using regular expression with sentence patterns (see Table 1) representing hypernym-hyponym relations, as exemplified in Figure 5. In those patterns, the `hypernym` tag is only allowed to consist of exactly one category-related keyword, while the `hyponym` tag may consist of one or more cultural heritage entities, which are in text separated by comma(s).

We build the hypernym lexicon by matching sentence patterns in Table 1 with free-form text in the province and city pages of Indonesian Wikipedia, and retrieving the matched hypernym parts. We also enrich the hypernym lexicon by extracting relevant headers from tables and relevant headings from lists in those pages. All descriptors in the descriptor lexicon are also included in the hypernym lexicon since they can be interpreted as types. Another

Table 1: Sentence patterns for hypernym-hyponym relations

Pattern (Indonesian)	Hearst Pattern (English)
<hypernym> seperti <hyponym> dan	<hypernym> such as <hyponym> and/or <hyponym>
<hypernym> termasuk <hyponym>	<hypernym> including <hyponym> and/or <hyponym>
<hyponym> adalah <hypernym>	such <hypernym> as <hyponym> and/or <hyponym>
<hyponym> menjadi <hypernym>	<hyponym> and/or other <hypernym>
<hyponym> merupakan <hypernym>	<hypernym> especially <hyponym> and/or <hyponym>

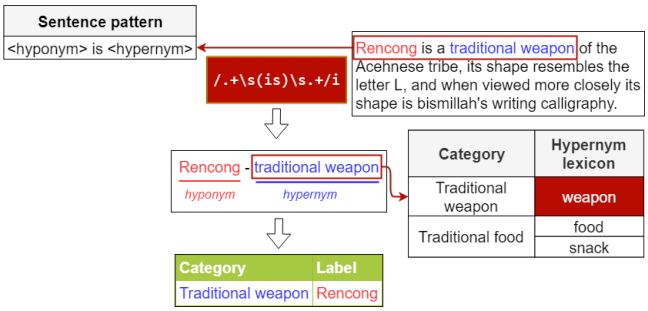

Figure 5: Example of extracting entities with hypernym lexicon

source to enrich the hypernym lexicon is category names in the taxonomy shown in Figure 2. For example, based on the extracted headings, `jajanan` (EN: snack) is added into the lexicon to detect traditional food entities. `Jajanan` appears as a heading title for a list of snacks, found in the Wikipedia page of Bali.[13] In total, there are 100 keywords in the hypernym lexicon, containing 38 more keywords than those in the descriptor lexicon. We intentionally differ the content of the hypernym lexicon to that of the descriptor lexicon based on our assumption that cultural heritage entities are seldom written along with those 38 descriptors — e.g., `alat musik` (EN: `musical instrument`).

3.1.2 Table extraction. Cultural heritage entities in a table can be obtained by detecting hypernyms in the table headers, and descriptors of compound nouns in the table cells. In HTML structure, headers and cells are children of rows, so the table parsing starts from finding all table rows. For each row, we process headers or cells inside the row. If any headers contain a hypernym that matches an entry in the hypernym lexicon, all cells below the header (that is, cells having the same column number with the header) are extracted as cultural heritage entities and categorized based on the matched hypernym. Furthermore, if any compound nouns in cells contain a descriptor that matches an entry in the descriptor lexicon, the cell is extracted as a cultural heritage entity and categorized based on the matched descriptor.

3.1.3 List extraction. In a list, cultural heritage entities may appear in the list items part. In order to determine whether a list item contains desired entities, we match its headings starting from the closest heading to `<h2>` with entries in the hypernym lexicon, as well as its compound nouns with entries in the descriptor lexicon. A list item may contain simply an entity alone, or an entity extended with additional description, separated by a certain delimiter (e.g., `Kupat Bongkok` – Kupat from Bongkok Village[14] with '–' as a delimiter). For this reason, we also perform delimiter detection in list items, retrieving only the entities we are interested in.

3.1.4 Free-form text extraction. As for free-form text extraction, we first perform sentence segmentation. Each sentence is then matched with the patterns [20] as shown in Table 1. Next, the matched sentences are divided into two parts: the hypernym part and the hyponym part. The hypernym part is matched with the hypernym lexicon to determine the entity category. The hyponym candidates in the hyponym part, with the help of POS tagging to detect nouns, are extracted. Using NER, we further filter out from those hyponym candidates, the entities having persons, organizations, and locations, as they cannot be cultural heritage entities. Finally, the remaining entities are paired with their categories.

For example, given the sentence *"Every region in South Kalimantan has foods as regional characteristics, such as Hulu Sungai Selatan with dodol ..., and others"*,[15] the `<hypernym>` part will be *"Every region in South Kalimantan has foods as regional characteristics"* and the `<hyponym>` part will be *"Hulu Sungai Selatan with dodol ..., and others"* as per the first pattern in Table 1. Performing POS tagging over the hyponym part gives us nouns, which can be entity candidates: `Hulu Sungai Selatan` and `dodol`. On top of the POS tagging results, we perform NER, showing that `Hulu Sungai Selatan` is a location entity, meaning that it cannot be a cultural heritage entity. Up to this step, the entity `dodol` still remains. We now determine the category of the extracted entity. Here, the hypernym part is tokenized and then matched with the hypernym lexicon. The only matched token is `foods`, so `dodol` will be extracted and stored as a pair (`food`, `dodol`).

As for the extraction using the descriptor lexicon, we perform POS tagging on the whole sentences to detect compound nouns. We then match the compound nouns with the descriptor lexicon to determine the entity category. Those that are matched are then extracted as cultural heritage entities.

3.1.5 KB Generation. The output of the table, list, and free-form text extraction processes is of the form (`category`, `entity`). This becomes the RDF triple of the form (`entity`, `rdf:type`, `category`), where both the entity and category have been transformed into

[13] https://id.wikipedia.org/wiki/Bali#Jajanan

[14] https://id.wikipedia.org/wiki/Kabupaten_Tegal

[15] https://id.wikipedia.org/wiki/Kalimantan_Selatan

their respective URIs within the BudayaKB namespaces.[16] Additionally, based on Table 2 we add location information by inferring from the Wikipedia article where the cultural heritage entity extraction is performed. This location information is also transformed into RDF triples. All the transformations are done based on the BudayaKB ontology.[17] In the ontology, classes are defined based on the taxonomy as shown in Figure 2.

An example of generated triples based on this KB generation procedure can be seen in the middle part of Figure 1. Both triples describe the resource `bkbr:Candi_Jiwa`, which represents the Jiwa temple object in the real world. The first triple states that Jiwa temple is a temple, whereas the second states that Jiwa temple is located in West Java (represented by the resource `bkbr:Jawa_Barat`).

3.2 Integration Module

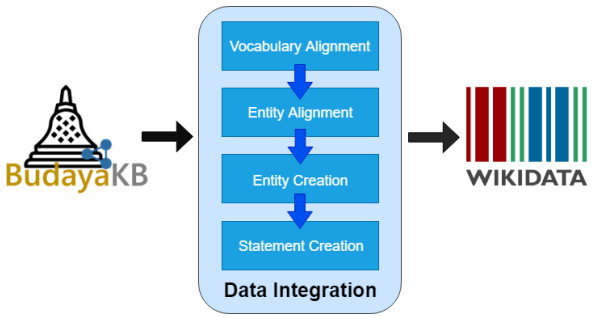

Figure 6: Data integration workflow from BudayaKB to Wikidata

The data integration process from BudayaKB to Wikidata (as shown in Figure 6) consists of four steps: vocabulary alignment, entity alignment, entity creation, and statement creation. The entity creation step is optional as there might already exist in Wikidata the cultural heritage entities we want to add. In this case, we simply enrich the statements of the existing entities. The properties of cultural heritage entities that we add (or enrich) to Wikidata are type, Indonesian label, country (that is always set to Indonesia), and location (province and city).

3.2.1 Vocabulary Alignment. In this step, we align the BudayaKB ontology and the Wikidata ontology. The alignment is done over classes (e.g., the musical instrument `bkb:AlatMusikTradisional` is mapped to `Q34379` in Wikidata) and properties (e.g., `rdf:type` is mapped to `P31` and location properties in Table 2 are mapped to `P276` in Wikidata).

3.2.2 Entity Alignment. Wikidata uses an identifier scheme that starts with "Q", followed by a certain unique number. It differs with BudayaKB which uses a human-readable identifier scheme based on the entity label. Consequently, resources in BudayaKB need to be mapped to the appropriate resources in Wikidata. The alignment is done semi-automatically using the MediaWiki API `wbsearchentities`[18]. In general, it accepts input parameters, such as the label of an item we want to search and the language of the label. The resources mapped are provinces, cities, and cultural heritage entities.

3.2.3 Entity and Statement Creation. The creation of entities and statements is done using the QuickStatement2 tool.[19] It accepts two types of command to import: V1 and CSV. The command type we use is CSV. Below are some command examples:

- Create a new entity (e.g., Suwawa script, or in Indonesian, Aksara Suwawa) and add statements to the entity.

```
qid,Lid,P31,S143,P17,S143,P276,S143
,Aksara Suwawa,Q8192,Q155214,Q252,Q155214,Q5067,Q155214
```

- Create statements to enrich existing entities (e.g., Calung).

```
qid,P31,S143,P17,S143,P276,S143
Q4201135,Q34379,Q155214,Q252,Q155214,Q3724,Q155214
```

The first row of the command is a header describing how the columns are interpreted. The columns are **qid**: item ID (if it is empty, a new item is created), **Lid**: item label in Indonesian, **P31**: property that specifies the item category, **S143**: source property that specifies the reference of the statement created, **P17**: country associated with the item, and **P276**: location of the item (province and city). The second row contains statements about an item. In the second example above, the statements created can be interpreted as follows:

- Calung (`Q4201135`) is an instance of (`P31`) a musical instrument (`Q34379`);
- Calung (`Q4201135`) is from the country (`P17`) of Indonesia (`Q252`); and
- Calung (`Q4201135`) is located in (`P276`) West Java (`Q3724`).

Moreover, for each created statement, we add reference information using the property "imported from Wikimedia project" (`S143`) with the value "Indonesian Wikipedia" (`Q155214`).

4 EVALUATION

We conduct an experimental evaluation to measure the quality of the extraction results using the BudayaKB approach and the integration results using MediaWiki API.

4.1 Extraction results

Particularly, we evaluate:

(1) the results of direct extraction in the form of pairs of a cultural heritage entity and its category; and
(2) the location information of cultural heritage entities, inferred from the title of the extracted articles.

We use precision and recall as the evaluation metrics. Precision states the number of true and successfully obtained entities compared to the total of successfully obtained entities. Recall states the number of true and successfully obtained entities compared to the total of true entities.

[16] The namespaces are: `https://budayakb.cs.ui.ac.id/resource/` with `bkbr` prefix, and `https://budayakb.cs.ui.ac.id/ns#` with `bkb` prefix.
[17] https://budayakb.cs.ui.ac.id/ns
[18] https://www.wikidata.org/w/api.php?action=help&modules=wbsearchentities
[19] https://tools.wmflabs.org/quickstatements

Table 2: Properties for locations (i.e., province and city) of cultural heritage entities distinguished by their categories

Category	Province property	City property
Regional script (`bkb:AksaraDaerah`)	`bkb:developedInProvince`	`bkb:developedInCity`
Ethnic group (`bkb:SukuBangsa`)	`bkb:populationPlaceInProvince`	`bkb:populationPlaceInCity`
Local language (`bkb:BahasaDaerah`)	`bkb:spokenInProvince`	`bkb:spokenInCity`
Immovable heritage (`bkb:WarisanBudayaTidakBergerak`)	`bkb:locationInProvince`	`bkb:locationInCity`
Movable heritage (`bkb:WarisanBudayaBergerak`)	`bkb:originFromProvince`	`bkb:originFromCity`
Intangible heritage (`bkb:WarisanBudayaTakBenda`)	`bkb:originFromProvince`	`bkb:originFromCity`

We manually build a gold standard of about 1700 category-entity pairs, compiled from 34 Wikipedia articles about Indonesian provinces. Comparing the gold standard with the extracted data using the BudayaKB approach gives a precision of 84% and a recall of 62%. Some explanations as to why false positives occur are as follows: (1) list headings may contain two or more category-related keywords (e.g., *"Musical instruments and traditional dances"*), giving an ambiguity; and (2) context understanding is not supported (e.g., the heading *"Language"* can also mean the list of words — instead of languages — used in a region). The recall result suggests that the sentence patterns as shown in Table 1 can still be completed further, for instance, to handle the sentence *"The traditional Acehnese house is called Rumoh Aceh"*.

We also evaluate the precision of the location extraction, inferred from the title of the extracted articles. Of the 155 entities we sample randomly, there are 12 wrong entities, giving a 92% precision.

4.2 Integration results

As described in Section 3.2.2, resources such as provinces, cities, and cultural heritage entities are mapped to items in Wikidata using the MediaWiki API. After being checked manually, all province and city entities are mapped to the correct items in Wikidata. As for the cultural heritage entities, 400 are mapped to the wrong items in Wikidata out of 1411 items retrieved using the API. We enrich the statements of the 1011 items that are mapped correctly. From 400 wrongly mapped entities, 224 entities are found in Wikidata and the other 176 are not. The rest are imported to Wikidata as new entities, along with 1796 entities that could not be retrieved using the API. In total, we create 1972 new cultural heritage entities in Wikidata. The full integration results grouped by cultural heritage categories are shown in Table 3.

Table 3: Integration results grouped by cultural heritage categories

Category	Mapped correctly	Mapped incorrectly		Not mapped	New entities	Total
		Exist in Wikidata	Do not exist in Wikidata			
Archaeological site	1	0	0	59	59	60
Bathing structure	0	0	0	1	1	1
Clothing	1	1	3	23	26	28
Ethnic group	105	101	0	39	39	245
Fabric motif	7	2	2	13	15	24
Folk game	0	1	0	3	3	4
Folk song	1	3	16	65	81	85
Folklore	5	5	1	20	21	31
Food	362	28	81	724	805	1195
Fortress	18	1	1	52	53	72
Hindu place of worship	12	1	0	24	24	37
Historical country	72	29	5	44	49	150
House	2	0	0	14	14	16
Indigenous community	0	0	0	2	2	2
Inscription	21	2	0	12	12	35
Manuscript	6	7	0	21	21	34
Mosque	59	3	27	101	128	190
Musical instrument	8	0	0	7	7	15
Oral tradition	1	0	6	9	15	16
Ornament	0	0	0	1	1	1
Palace	14	1	0	28	28	43
Performing art	17	1	1	17	18	36
Pond	1	0	1	5	6	7
Regional language	110	19	0	34	34	163
Regional script	2	0	0	6	6	8
Ritual	33	12	3	52	55	100
Shrine	1	0	0	0	0	1
Statue	0	1	0	15	15	16
Temple	67	0	1	25	26	93
Tomb	5	2	0	217	217	224
Traditional dance	65	3	9	120	129	197
Traditional music	6	1	2	31	33	40
Weapon	9	0	17	12	29	38
Total	1011	224	176	1796	1972	3207

5 USE CASES

From the extraction and integration of cultural heritage entities, we explore various use cases to show the benefits of our BudayaKB approach. All the use cases we give here rely on SPARQL [9], a query language for RDF, that can be used to retrieve knowledge from BudayaKB (and other RDF KBs). Queries in SPARQL are based on patterns that resemble RDF triples, except that the patterns may now contain variables. SPARQL query evaluation is done by matching such patterns over RDF triples.

Use Case 1: Basic Query Answering. We show how (simple) questions about cultural heritage can now be answered in just one place, unlike the previous situation where cultural heritage entities are scattered into different pages with heterogeneous formats. For example, the query below asks for "which province is Rawon (a culinary dish) originally from".

```
SELECT * WHERE {
  bkbr:Rawon bkb:originFromProvince ?prov
}
```

The query can be evaluated over our BudayaKB SPARQL server,[20] giving the answer `bkbr:Jawa_Timur` (EN: East Java).

Use Case 2: Cultural Heritage Analytics. Here, we showcase how BudayaKB can be analyzed to provide insights about the statistics of cultural heritage entities from different Indonesian provinces. The query below ranks Indonesian provinces based on the number of temples (ID: `Candi`) they have, in descending order.

[20] https://budayakb.cs.ui.ac.id/dataset.html?tab=query&ds=/budaya

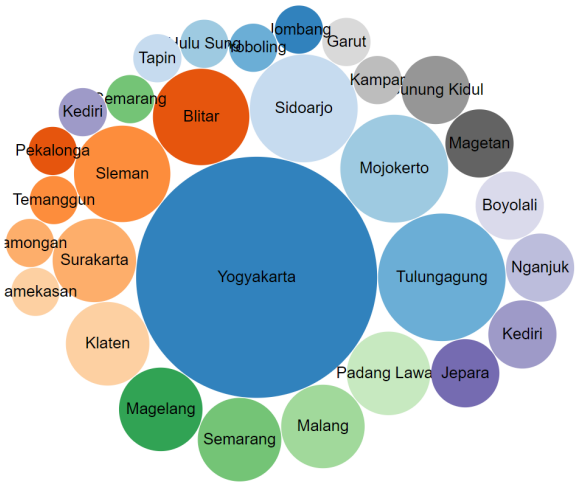

Figure 7: A bubble chart visualization for "number of temples grouped by Indonesian cities/regencies"

```
SELECT ?prov (COUNT(?temple) AS ?numTemple) WHERE {
  ?temple a bkb:Candi .
  ?temple bkb:locationInProvince ?prov .
} GROUP BY ?prov ORDER BY DESC(?numTemple)
```

The query result gives the insight that Yogyakarta is the province with the most temples (25), followed by East Java (23), and Central Java (15).

Use Case 3: Visualization. In this use case, we create a visualization of cultural heritage entities in Wikidata, taking into account all the BudayaKB entities that have been imported and linked to Wikidata. The visualization in Figure 7 displays a bubble chart from the query "number of temples grouped by Indonesian cities/regencies". This visualization is created using the Wikidata query service.[21] Other visualization styles include bar charts, graphs, and maps.

6 RELATED WORK

NER. Named-entity recognition (NER) has always been at the forefront of natural language processing (NLP) developments since it was first highlighted in [8]. Traditionally, the goal of the NER task involves identifying the names of people, organizations, and locations. In [14], those names are redefined further into fine-grained types (e.g., actor as the specific role of person). Modern NER tasks may include the identification of counting entities, that is, numbers identifying the cardinality of an entity's property like "how many provinces South Korea has" [17]. Typical NER approaches are based on handcrafted rules, lexicons, machine learning, or neural networks [26]. Named-entity linking (NEL) is closely related to NER, which maps entity mentions in a text to their knowledge base (KB) correspondences. Most notable challenges in NEL are name variations (e.g., Donald Trump vs. Mr. Trump) and ambiguity (e.g., Dresden in Germany vs. Dresden in Ohio, US) [22]. Popular NEL systems include DBpedia Spotlight [16], Babelfy [18], and AIDA [13]. Our BudayaKB approach complements the existing systems by focusing on the domain of cultural heritage and linking the extracted entities to the Wikidata KB.

Extracting Entity Pairs Given a Relationship Type. In this task, we are given one or more relationship types, and our goal is to find all occurrences of those relationships in a corpus in order to obtain entity pairs. Most work in this area, such as [3, 15, 24], has been done on open document collections like the Web, where one cannot assume that entities are already marked. In a widely cited paper, Hearst [11] showed that the lexico-syntactic pattern "Y such as X" can be used to mine large text corpora for word pairs X:Y in which X is a hyponym of Y. The hypernym-hyponym relation relates generic terms or classes to their specific instances, such as cultural heritage categories to cultural heritage entities. We adopt this technique in our approach, and provide a workflow and an implementation in extracting cultural heritage entities in Wikipedia.

KB Construction. Several approaches exist for constructing KBs from tables and free-form text. Dimou et al. [6] introduced RML, a mapping language from heterogeneous data sources to RDF. RML enables the creation of RDF mapping definitions from data sources in tabular formats (e.g., relational DB, CSV), XML, and JSON, and provides a high degree of mapping customization (e.g., URI templating, joins). Nonetheless, their work does not consider mapping from textual sources and therefore is not suitable for the entity extraction task we are dealing with. In [19], Muñoz et al. proposed an approach to extract RDF triples from Wikipedia tables using DBpedia as a reference dataset. Over a Wikipedia table, their approach extrapolated existing relationships between DBpedia entities found in a part of the table, to the rest of the table. Their work assumed that table cells contain wiki-links, which could be directly mapped to DBpedia entities and did not look at string literals. Our work differs to theirs, in that we concentrate more on entity extraction (from string literals) as opposed to relation extraction. Exner and Nugues [7] developed a system to extract DBpedia RDF triples from unstructured text. The main components of the system include a semantic parser, a coreference solver, and a DBpedia entity linker. An experimental evaluation of the system over 200 randomly sampled sentences from English Wikipedia articles reported an F1-score of 66.3%. However, unlike our BudayaKB approach, they did not examine the extraction of RDF triples from (Wikipedia) tables and lists.

7 CONCLUSIONS AND FUTURE WORK

The aim of this paper is to propose an approach to extract cultural heritage entities from diverse formats (i.e., text, lists, and tables) in order to construct a knowledge base called BudayaKB, which provides contents about cultural heritage entities in a structured form. Our extraction approach follows on the observation that cultural heritage entities are often written down either with common noun descriptors or categories, which could be related to the hypernym-hyponym relationship. Our approach relies on lexicons that are built semi-automatically, consisting of common noun descriptors and category-related keywords for cultural heritage entities. To evaluate our BudayaKB approach, we compare the extraction results with a gold standard, achieving 84% precision and 62% recall.

[21] http://bit.ly/bubbleTemple

We consider several future directions. Our approach is not able to tackle entity resolution problem — one entity may have different labels (e.g., Minangnese language may be written Minangnese or Minangkabaunese language). It is also not able to tackle entity disambiguation problem — different entities may have the same label (e.g., Javenese and Balinese gamelan may be written gamelan). Both problems can be dealt with by extracting attributes of the entities and then carrying out similarity calculations among them. Adding quality metadata (e.g., completeness metadata [5]) on top of the extracted entities in BudayaKB can also be an interesting future direction, as this lets users perceive the quality state of BudayaKB.

ACKNOWLEDGMENTS

This work was partially supported by Wikimedia Indonesia under the Indonesian Content Creation Program and by Bukalapak.

REFERENCES

[1] Nurul Fajrin Ariyani, Alief Yoga Priyanto, Sarwosri Sarwosri, and Riyanarto Sarno. 2017. Pemodelan Granularitas Temporal untuk Mencari Relasi Antar Objek Warisan Budaya Indonesia dengan Menggunakan Ontologi. *JUTI: Jurnal Ilmiah Teknologi Informasi* 15 (01 2017), 72. https://doi.org/10.12962/j24068535.v15i1.a637

[2] Risda Asfina and Ririn Ovilia. 2016. Be Proud of Indonesian Cultural Heritage Richness and Be Alert of Its Preservation Efforts in the Global World. *Humanus: Jurnal Ilmiah Ilmu-Ilmu Humaniora* 15, 2 (10 2016), 195–206. https://doi.org/10.24036/jh.v15i2.6428

[3] Michele Banko, Michael J. Cafarella, Stephen Soderland, Matt Broadhead, and Oren Etzioni. 2007. Open Information Extraction from the Web. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence (IJCAI'07)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2670–2676. http://dl.acm.org/citation.cfm?id=1625275.1625705

[4] Douglas R. Cutting, Julian Kupiec, Jan O. Pedersen, and Penelope Sibun. 1992. A Practical Part-of-Speech Tagger. In *3rd Applied Natural Language Processing Conference, ANLP 1992, Trento, Italy, March 31 - April 3, 1992*. 133–140. http://aclweb.org/anthology/A/A92/A92-1018.pdf

[5] Fariz Darari, Werner Nutt, Giuseppe Pirrò, and Simon Razniewski. 2018. Completeness Management for RDF Data Sources. *TWEB* 12, 3 (2018), 18:1–18:53. https://doi.org/10.1145/3196248

[6] Anastasia Dimou, Miel Vander Sande, Pieter Colpaert, Ruben Verborgh, Erik Mannens, and Rik Van de Walle. 2014. RML: A Generic Language for Integrated RDF Mappings of Heterogeneous Data. In *Proceedings of the Workshop on Linked Data on the Web co-located with the 23rd International World Wide Web Conference (WWW 2014), Seoul, Korea, April 8, 2014. (CEUR Workshop Proceedings)*, Christian Bizer, Tom Heath, Sören Auer, and Tim Berners-Lee (Eds.), Vol. 1184. CEUR-WS.org. http://ceur-ws.org/Vol-1184/ldow2014_paper_01.pdf

[7] Peter Exner and Pierre Nugues. 2012. Entity Extraction: From Unstructured Text to DBpedia RDF triples. In *Proceedings of the Web of Linked Entities Workshop, Boston, USA, November 11, 2012 (CEUR Workshop Proceedings)*, Giuseppe Rizzo, Pablo N. Mendes, Eric Charton, Sebastian Hellmann, and Aditya Kalyanpur (Eds.), Vol. 906. CEUR-WS.org, 58–69. http://ceur-ws.org/Vol-906/paper7.pdf

[8] Ralph Grishman and Beth Sundheim. 1996. Message Understanding Conference-6: A Brief History. In *16th International Conference on Computational Linguistics, Proceedings of the Conference, COLING 1996, Center for Sprogteknologi, Copenhagen, Denmark, August 5-9, 1996*. 466–471. http://aclweb.org/anthology/C96-1079

[9] Steve Harris and Andy Seaborne (Eds.). 21 March 2013. *SPARQL 1.1 Query Language*. W3C Recommendation. http://www.w3.org/TR/sparql11-query/

[10] Khafiizh Hastuti, Erwin Yudi Hidayat, and Elkaf Rahmawan. 2013. Purwarupa Tangible Cultural Heritage Documentation Berbasis Database Multimedia. *Techno.COM — Jurnal Teknologi Informasi* 12, 4 (11 2013), 188–197. https://doi.org/10.33633/tc.v12i4.800

[11] Marti A. Hearst. 1992. Automatic Acquisition of Hyponyms from Large Text Corpora. In *Proceedings of the 14th Conference on Computational Linguistics - Volume 2 (COLING '92)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 539–545. https://doi.org/10.3115/992133.992154

[12] Pascal Hitzler, Markus Krötzsch, and Sebastian Rudolph. 2010. *Foundations of Semantic Web Technologies*. Chapman and Hall/CRC Press. http://www.semantic-web-book.org/

[13] Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. 2011. Robust Disambiguation of Named Entities in Text. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, EMNLP 2011, 27-31 July 2011, Edinburgh, UK, A meeting of SIGDAT*. 782–792. http://www.aclweb.org/anthology/D11-1072

[14] Xiao Ling and Daniel S. Weld. 2012. Fine-grained Entity Recognition. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence (AAAI'12)*. AAAI Press, 94–100. http://dl.acm.org/citation.cfm?id=2900728.2900742

[15] Rahmad Mahendra, Lilian Wanzare, Bernardo Magnini, Raffaella Bernardi, and Alberto Lavelli. 2011. Acquiring Relational Patterns from Wikipedia: A Case Study. In *Proceedings of LTC 2011*.

[16] Pablo N. Mendes, Max Jakob, Andrés García-Silva, and Christian Bizer. 2011. DBpedia spotlight: shedding light on the web of documents. In *Proceedings the 7th International Conference on Semantic Systems, I-SEMANTICS 2011, Graz, Austria, September 7-9, 2011*. 1–8. https://doi.org/10.1145/2063518.2063519

[17] Paramita Mirza, Simon Razniewski, Fariz Darari, and Gerhard Weikum. 2018. Enriching Knowledge Bases with Counting Quantifiers. In *The Semantic Web - ISWC 2018 - 17th International Semantic Web Conference, Monterey, CA, USA, October 8-12, 2018, Proceedings*. 179–197. https://doi.org/10.1007/978-3-030-00671-6_11

[18] Andrea Moro, Alessandro Raganato, and Roberto Navigli. 2014. Entity Linking meets Word Sense Disambiguation: a Unified Approach. *TACL* 2 (2014), 231–244. https://tacl2013.cs.columbia.edu/ojs/index.php/tacl/article/view/291

[19] Emir Muñoz, Aidan Hogan, and Alessandra Mileo. 2014. Using linked data to mine RDF from wikipedia's tables. In *Seventh ACM International Conference on Web Search and Data Mining, WSDM 2014, New York, NY, USA, February 24-28, 2014*, Ben Carterette, Fernando Diaz, Carlos Castillo, and Donald Metzler (Eds.). ACM, 533–542. https://doi.org/10.1145/2556195.2556266

[20] Made Nindyatama Nityasya, Rahmad Mahendra, and Mirna Adriani. 2018. Hypernym-Hyponym Relation Extraction from Indonesian Wikipedia Text. In *2018 International Conference on Asian Language Processing, IALP 2018, Bandung, Indonesia, November 15-17, 2018*, Minghui Dong, Moch Arif Bijaksana, Herry Sujaini, Ade Romadhony, Fariska Z. Ruskanda, Elvira Nurfadhilah, and Lyla Ruslana Aini (Eds.). IEEE, 285–289. https://doi.org/10.1109/IALP.2018.8629216

[21] Alfredo M. Ronchi. 2009. *eCulture — Cultural content in the digital age*. Springer-Verlag Berlin Heidelberg. https://doi.org/10.1007/978-3-540-75276-9

[22] Henry Rosales-Méndez, Barbara Poblete, and Aidan Hogan. 2018. What Should Entity Linking link?. In *Proceedings of the 12th Alberto Mendelzon International Workshop on Foundations of Data Management, Cali, Colombia, May 21-25, 2018. (CEUR Workshop Proceedings)*, Dan Olteanu and Barbara Poblete (Eds.), Vol. 2100. CEUR-WS.org. http://ceur-ws.org/Vol-2100/paper10.pdf

[23] Sunita Sarawagi. 2008. Information Extraction. *Found. Trends databases* 1, 3 (March 2008), 261–377. https://doi.org/10.1561/1900000003

[24] Peter D. Turney. 2006. Expressing Implicit Semantic Relations without Supervision. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Sydney, Australia, 313–320. https://doi.org/10.3115/1220175.1220215

[25] UNESCO. 2019. Tangible Cultural Heritage. http://www.unesco.org/new/en/cairo/culture/tangible-cultural-heritage/. Accessed: 2019-01-25.

[26] Vikas Yadav and Steven Bethard. 2018. A Survey on Recent Advances in Named Entity Recognition from Deep Learning models. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, Emily M. Bender, Leon Derczynski, and Pierre Isabelle (Eds.). Association for Computational Linguistics, 2145–2158. https://aclanthology.info/papers/C18-1182/c18-1182