Model WordNet Bahasa Indonesia berbasis *Linked Data*

Hendrik[1], Andhik Budi Cahyono[2]

***Abstract*—WordNet is an online lexical database. In Computer Science domain, it plays important role in solving semantic interoperability issues. It also helps in many researches related to Natural Language Processing topic. Because of the importance of WordNet, there are many works to develop WordNet into several languages, e.g., Japanese, Arabic, and Indonesian. However, those are still not sufficient to address semantic interoperability issues. Therefore, there are several attempts to form WordNet into machine understandable format, i.e, Resource Description Framework (RDF) model. Still, there is no effort to form WordNet Bahasa Indonesia into RDF format. This paper presents the process of forming WordNet Bahasa Indonesia into Linked Data form. This process involves several phases, which are identifying data sources, data extraction, data transformation, data loading into relational database, and mapping database model into RDF model. The latest is done by using D2RQ framework, resulting the WordNet Bahasa Indonesia as Linked Data format. This data set is linked to WordNet-RDF of Princetown University.**

***Intisari*— WordNet merupakan sebuah basis data leksikal yang bersifat daring. Di bidang ilmu komputer, WordNet berperan penting dalam membantu menyelesaikan permasalahan interoperabilitas sistem dari sisi semantik. WordNet juga banyak membantu berbagai penelitian di bidang *Natural Language Processing* (NLP). Beberapa contoh penggunaan WordNet adalah untuk mendukung aplikasi tanya-jawab secara otomatis, analisis sentimen, dan klasifikasi teks. Sedemikian pentingnya peran WordNet sehingga banyak upaya peneliti untuk mengembangkan WordNet ke berbagai bahasa selain Bahasa Inggris, seperti Bahasa Jepang, Bahasa Arab, dan Bahasa Indonesia. Meskipun sudah sedemikian banyak penelitian terkait WordNet, dirasa belum memadai untuk menjawab permasalahan interoperabilitas dari sisi semantik. Oleh karena itu, mulai ada upaya untuk menjadikan WordNet ke dalam model yang *machine understandable*, misalnya ke dalam model *Resource Description Framework* (RDF), tetapi belum ada upaya menjadikan WordNet Bahasa Indonesia dalam format *machine understandable*. Makalah ini memaparkan pengembangan model WordNet Bahasa Indonesia berbasis *Linked Data*. Tahapan pengembangan meliputi identifikasi sumber data, ekstraksi data, transformasi data, pemuatan data ke dalam basis data relasional, serta pemetaan basis data relasional ke model RDF. Proses pemetaan menggunakan *framework* D2RQ dan menghasilkan WordNet Bahasa Indonesia berbasis *Linked Data*. *Data set* ini ditautkan dengan WordNet-RDF dari Princetown University.**

***Kata Kunci*— WordNet, *Linked Data*, NLP, Linguistik, RDF.**

I. PENDAHULUAN

Menurut data yang dirilis laman We Are Social (https://www.techinasia.com/indonesia-web-mobile-statistics-we-are-social), pada awal tahun 2016 terdapat sekitar 88,1 juta pengguna internet aktif di Indonesia, meningkat 20% bila dibandingkan pada awal tahun 2015. Laporan tersebut juga menunjukkan bahwa terjadi peningkatan jumlah pengguna aktif media sosial sebesar 6% dibandingkan jumlah pengguna media sosial tahun sebelumnya, yakni sebesar 79 juta pengguna. Facebook masih menjadi *platform* media sosial yang paling aktif digunakan, sementara itu BlackBerry Messenger (BBM) menjadi aplikasi *instant messenger* teraktif yang digunakan.

Sebagaimana dipahami, media sosial telah banyak mengubah kehidupan masyarakat modern. Mulai dari cara berkomunikasi, menyebarkan pesan, hingga menjadi alat rekayasa sosial [1]. Media sosial pernah dimanfaatkan dalam menyebarkan beragam pesan dan opini seperti pada kasus Prita, perseteruan KPK dan Polri, kampanye pemilihan presiden, hingga kasus penistaan agama oleh salah seorang calon pada perhelatan pemilihan kepala daerah.

Berbagai macam opini dan pesan di media sosial ditambah dengan berbagai informasi melalui berbagai macam media daring lainnya dapat diproses menjadi suatu informasi berkualitas tinggi. Pesan maupun opini yang ada di media sosial merupakan data yang sifatnya tidak terstruktur. Umumnya berbagai data tersebut harus diolah menggunakan *Natural Language Processing* (NLP). NLP memungkinkan interaksi antara komputer dengan bahasa alami manusia.

Pada bidang NLP, ada beberapa tugas utama yang sering dilakukan, misalnya *parsing*, *sentiment analysis*, *topic segmentation*, *word segmentation*, *question and answering*, serta *relationship extraction*. Guna melakukan beberapa hal tersebut, umumnya peneliti di bidang NLP memerlukan suatu kamus atau daftar kata. Daftar kata tersebut dapat dibuat sendiri maupun menggunakan yang sudah jadi. Untuk Bahasa Inggris, terdapat suatu kamus leksikal yang banyak digunakan oleh peneliti NLP, yakni WordNet atau biasa dikenal sebagai Pricetown WordNet (PWN), yang dapat diakses di https://wordnet.princeton.edu/.

WordNet merupakan sebuah basis data leksikal yang bersifat daring. Pengembangannya didasarkan pada teori psikolinguistik memori leksikal manusia. Pada WordNet, kata kerja, kata benda, kata sifat, dan kata keterangan dikelompokkan menjadi kumpulan sinonim kognitif (*synset*), guna merepresentatsikan konsep-konsep yang berbeda [2]. WordNet banyak dimanfaatkan oleh bidang ilmu lainnya yang terkait dengan linguistik. Di bidang ilmu komputer, WordNet berperan penting dalam membantu menyelesaikan permasalahan interoperabilitas sistem dari sisi semantik. WordNet juga banyak membantu berbagai penelitian di bidang NLP.

Contoh penggunaan WordNet adalah untuk mendukung aplikasi tanya-jawab secara otomatis [3] - [5]. Contoh lain, WordNet digunakan dalam aktivitas analisis sentimen yang

[1,2] *Staf pengajar, Jurusan Teknik Informatika Fakultas Teknologi Industri Universitas Islam Indonesia, Jl. Kaliurang KM 14,5 Sleman 55285 INDONESIA (tlp: 0274- 895 287 ext. 122; fax: 0274-895007; e-mail: hendrik@uii.ac.id, andhikbudi@uii.ac.id)*

biasa digunakan di media sosial seperti Twitter [6], [7]. Selain itu, WordNet juga digunakan dalam proses klasifikasi teks [8], [9]. Masih banyak contoh lain pemanfaatan WordNet di berbagai bidang.

Sedemikian pentingnya peran WordNet sehingga banyak upaya peneliti untuk mengembangkan WordNet ke berbagai bahasa, seperti Bahasa Arab dan Bahasa Jepang [10], [11]. Untuk WordNet juga sudah pernah dikembangkan untuk Bahasa Indonesia [12]. Upaya tersebut kemudian memotivasi upaya untuk mengembangkan WordNet Bahasa Indonesia yang dikombinasikan dengan Bahasa Malaysia maupun Bahasa Melayu (kombinasi Bahasa Malaysia dan Bahasa Indonesia) [13]. Selain dengan Bahasa Malaysia, WordNet bahasa Indonesia juga dihubungkan dengan berbagai bahasa negara Asia lainnya seperti Thailand, Jepang, Laos, Mongolia, dan sebagainya, melalui suatu wadah Asian WordNet [14].

Meskipun sudah sedemikian banyak penelitian terkait WordNet, namun dirasa belum memadai untuk menjawab permasalahan interoperabilitas dari sisi semantik. Oleh karena itu, kemudian mulai ada upaya untuk menjadikan WordNet ke dalam model yang lebih *machine understandable*, misalnya ke dalam model *Resource Description Framework* (RDF) [15]. WordNet berbasis RDF ini menjadi salah satu *data set* hasil upaya untuk mempublikasikan dan menautkan berbagai sumber daya lingustik sebagaimana yang dilakukan oleh *Open Linguistics Working Group* (OWLG) dalam bentuk *Lingustics Linked Open Data* (LLOD) *Cloud* [16]. Meski demikian, belum ada upaya melakukan hal yang sama untuk Wordnet Bahasa Indonesia.

Berdasarkan kondisi tersebut, dicoba dilakukan upaya menjadikan WordNet Bahasa Indonesia ke dalam bentuk *machine understandable format* dalam model RDF dan ditautkan dengan *data set* lain menggunakan prinsip *Linked Data* (LD). Diharapkan adanya WordNet Bahasa Indonesia dalam bentuk *machine understandable format* dapat membantu meningkatkan kualitas penelitian berbasis NLP dengan konteks ke-Indonesia-an. Hal ini juga dapat memberikan pengaruh positif bagi perkembangan penelitian di bidang *Big Data* yang saat ini sedang banyak dilakukan.

Guna memudahkan dalam menjelaskan upaya pengembangan model WordNet Bahasa Indonesia dalam bentuk *machine understandable format*, makalah ini diorganisasikan menjadi beberapa bagian. Bagian pertama adalah pendahuluan, menjelaskan latar belakang dilakukannya upaya ini serta tujuan yang ingin dicapai. Pada bagian kedua, dijelaskan beberapa konsep dasar terkait RDF dan LD. Selanjutnya, tahapan pelaksanan dijelaskan pada bagian ketiga, yakni metodologi. Pada bagian keempat, dijelaskan mengenai hasil pembuatan model serta pembahasan terhadap model tersebut. Akhirnya, pada bagian kelima, makalah ini ditutup dengan kesimpulan terkait hasil yang diperoleh serta saran atau rencana pengembangan penelitian selanjutnya.

II. *RESOURCE DESCRIPTION FRAMEWORK* DAN *LINKED DATA*

A. Resource Description Framework (RDF)

RDF merupakan model data yang memungkinkan untuk mendeskripsikan berbagai sumber daya di *web*. RDF bertujuan untuk memudahkan interoperabilitas antar aplikasi yang saling bertukar informasi yang *machine understandable*, sehingga memudahkan pemrosesan otomatis di Web [17].

RDF memungkinkan dilakukannya dekomposisi terhadap suatu pengetahuan menjadi bagian-bagian yang lebih kecil yang disebut sebagai *statement*. Suatu *statement* terdiri atas tiga komponen, yaitu subjek, predikat, dan objek. Berdasarkan karakteristik ini, suatu *statement* juga disebut sebagai *triple*. Tabel I memuat beberapa contoh *statement* yang menghasilkan sebuah pengetahuan.

TABEL I
DAFTAR BEBERAPA *STATEMENT* RDF

Subjek	Predikat	Objek
Universitas_islam_indonesia	merupakan	universitas
Universitas_islam_indonesia	berdiri_pada	“1945”
Universitas_islam_indonesia	berlokasi_di	Yogyakarta
Universitas_islam_indonesia	memiliki_fakultas	“8”

Pada model RDF, penamaan subjek dan predikat dituliskan menggunakan *Uniform Resource Identifier* (URI), sedangkan objek, karena memungkinkan berupa *resource* maupun literal, penamaannya berbeda. Jika berupa *resource*, maka penamaannya juga dituliskan dengan URI. Adapun objek sebagai *literal* (*string* atau numerik) dituliskan tanpa URI. Pada umumnya objek berupa *literal* dituliskan menggunakan tanda kutip sebagaimana ditunjukkan pada Tabel I, dengan nilai objek “1945” dan “8”.

Pada umumnya, RDF dinotasikan dengan menggunakan *graph*, dengan subjek dan objek sebagai *node*, dan predikat sebagai *edge*. Adapun untuk kepentingan kemudahan akses oleh mesin, RDF biasa diserialisasikan dengan beberapa format. Beberapa format populer adalah RDF/XML, *Notation*-3, dan *N-Triple*.

B. Linked Data (LD)

LD merupakan suatu metode untuk mempublikasikan dan menghubungkan data terstruktur di *web* [18]. Konsep LD telah dirumuskan oleh Tim-Berners Lee (TBL), penemu Web, sejak tahun 2006. Meski demikian, LD baru mulai populer sejak diperkenalkan pada tahun 2009 di konferensi TED [19]. LD merupakan realisasi dari visi TBL terkait teknologi *web* yang dikembangkan semenjak 90-an. TBL mengharapkan, melalui LD, Web tidak hanya memuat dokumen yang saling terhubung, namun juga data. Dengan demikian, Web dapat dilihat sebagai basis data global yang memuat berbagai macam *data set* yang saling terhubung. Pada akhirnya, LD diharapkan dapat meminimalkan pulau-pulau data yang seringkali muncul di berbagai aplikasi sistem informasi.

LD memuat beberapa panduan *best-practice* sebagai berikut.

1. Memberi pengenal unik berupa URI terhadap *things* (segala sesuatu yang ada di dunia nyata).
2. Menggunakan HTTP URI, sehingga *things* tersebut dapat diakses lebih lanjut.
3. Menggunakan standar (misal: RDF dan SPARQL) agar setiap orang dapat mengakses informasi yang bermanfaat terkait *things* tersebut.

4. Menambahkan tautan pada URI untuk menghubungkan *things* tersebut dengan informasi terkait lainnya pada *data set* yang lain.

Secara spesifik, upaya untuk mempublikasikan dan menautkan berbagai sumber daya linguistik dengan prinsip LD juga dilakukan oleh OWLG. OWLG merupakan suatu jaringan yang bertujuan mempublikasikan berbagai sumber daya linguistik di bawah lisensi terbuka. OWLG telah menghasilkan suatu (LLOD) *Cloud* yang memuat sekitar 130an *data set* linguistik yang terbagi menjadi enam jenis sumber daya, yaitu *Corpora*, *Terminologies*, *Thesauri and Knowledge Bases*, *Lexicons and Dictionaries*, *Linguistic Resource Metadata*, *Linguistic Data Category*, dan *Typological Databases*. Contoh *data set* populer di bidang linguistik yang telah dimodelkan menjadi model LD adalah W3C WordNet dan *Manually Annotated Sub-Corpus* (MASC).

Ada beberapa keuntungan pemodelan sumberdaya linguistik sebagai LD dibandingkan dengan model formal yang sudah ada, yaitu sebagai berikut [20].

1. Meningkatkan interoperabilitas untuk berbagai macam aplikasi.
2. Memudahkan integrasi pengetahuan dari berbagai sumber daya yang ada.
3. Memudahkan akses dan distribusi sumber daya linguistik.
4. Membantu pengembangan infrastruktur NLP, komputasi leksikografi, atau korpus bahasa.

C. WordNet dalam Berbagai Bahasa

Upaya menjadikan WordNet dalam berbagai bahasa sudah banyak dilakukan pada berbagai penelitian sebelumnya. Pada dasarnya, ada dua pendekatan untuk membangun WordNet, yaitu *expand* dan *merge* [13]. Pendekatan yang pertama dilakukan dengan cara menerjemahkan *synset* PWN ke dalam suatu bahasa tertentu. Melalui pendekatan ini, struktur semantik PWN juga ditransfer secara langsung. Pendekatan kedua dilakukan dengan cara membuat sendiri WordNet independen dalam bahasa tertentu untuk kemudian disesuaikan dengan *synset* pada PWN. Pendekatan kedua memungkinkan pengembangan WordNet yang lebih memperhatikan keunikan semantik pada bahasa target. Meskipun demikian, pendekatan pertama lebih sederhana dan mudah dilakukan daripada pendekatan yang kedua [12].

Arabic WordNet (AWN) menggunakan pendekatan *expand* untuk pengembangannya. Selain itu, AWN juga dihubungkan dengan *Suggested Upper Merged Ontology* (SUMO). SUMO digunakan untuk memaksimalkan konsistensi semantik tautan-tautan hiponim. Pada AWN, suatu *synset* dari PWN dipetakan ke suatu istilah umum SUMO atau suatu istilah yang ekuivalen dengan *synset* tersebut. Istilah tersebut kemudian dipetakan ke suatu istilah atau kata dalam Bahasa Arab [10]. *Data set* hasil pengembangan dapat diakses menggunakan aplikasi berbasis Java.

Pendekatan yang sama dilakukan pada *Japanese* WordNet (WN-ja). Meski demikian, WN-ja ditautkan tidak hanya dengan SUMO, tetapi juga dengan GoiTaikei, sebuah leksikon semantik bahasa Jepang, dan Open ClipArt *Library* [11]. Untuk mengakses *data set* WN-ja, pengguna dapat mengunduh data dalam bentuk basis data sqlite3 maupun XML. Data yang sudah diunduh selanjutnya dapat dipasang di mesin pengguna.

Sementara itu untuk WordNet Bahasa Indonesia (WN-ind), meskipun menggunakan pendekatan *expand*, tetapi berbeda dari pendekatan yang dilakukan WN-ja maupun AWN. Alih-alih menggunakan suatu ontologi untuk menjadi perantara pemetaan konsep dengan PWN, WN-ind menggunakan Kamus Besar Bahasa Indonesia (KBBI). Pada WN-ind ini, setiap *synset* pada PWN dipetakan dengan suatu definisi pada KBBI [12]. Proyek WN-ind hanya menyediakan antarmuka berbasis *web* untuk mengakses datanya.

Selain pengembangan WordNet untuk sebuah bahasa tertentu, beberapa upaya lain dilakukan dengan menggabungkan WordNet beberapa bahasa. Contohnya adalah upaya menggabungkan WordNet Bahasa Malaysia (WN-msa) dan WordNet Bahasa Indonesia (WN-ind) menjadi sebuah WordNet Bahasa (WN-msa) [13]. WordNet Bahasa Indonesia juga sudah dikolaborasikan dengan delapan bahasa negara lainnya, yakni Jepang, Vietnam, Laos, Thailand, Sri Lanka, Nepal, Myanmar, dan Mongolia [14]. Kolaborasi sembilan WordNet bahasa ini disebut sebagai *Asian* WordNet.

Proyek WordNet Bahasa menyediakan antarmuka berbasis *web* sebagaimana WN-ind serta menyediakan data mentah yang dapat diunduh sebagai berkas teks dengan format *tab-delimited*. Sementara itu, untuk *Asian* WordNet, akses hanya dapat dilakukan untuk WordNet berbahasa Thai.

Adapun upaya mengembangkan suatu WordNet dalam model RDF telah dimulai sejak 2006 [15]. Proyek ini merupakan upaya W3C menyediakan standar konversi WordNet ke dalam model RDF/OWL. Upaya ini kemudian diimplementasikan pula pada [11]. Selain mengonversi WN-ja ke dalam model RDF, juga dilakukan upaya menggabungkannya dengan DBpedia *Japanese* dan dipublikasikan sebagai suatu *Linked Open Data* dengan lisensi CC-BY pada Data Hub.

III. METODOLOGI

Berikut tahapan-tahapan proses pengembangan model WordNet Bahasa Indonesia berbasis LD.

A. Identifikasi Sumber Data

Pada tahap ini, dilakukan identifikasi sumber data yang menjadi rujukan utama dalam proses transformasi. Proses ini menghasilkan tiga sumber data yang menjadi kandidat sebagai sumber data penelitian ini, yakni WordNet Bahasa (http://wn-msa.sourceforge.net/), Wordnet Bahasa Indonesia (http://bahasa.cs.ui.ac.id/iwn/), dan *Asian* WordNet (http://www.asianwordnet.org/). Dari ketiganya, WordNet Bahasa dipilih sebagai sumber utama karena kemudahan dalam memperoleh data mentah. WordNet Bahasa dapat diunduh melalui URL https://sourceforge.net/p/wn-msa/tab/HEAD/tree/trunk/. WordNet Bahasa Indonesia dan *Asian* WordNet tidak menyediakan data mentah yang mudah diunduh.

B. Ekstraksi Data

Setelah data mentah diperoleh, tahapan selanjutnya adalah melakukan ekstraksi data. Paket sumber data WordNet Bahasa memuat tiga berkas data, yakni dua berkas terkait WordNet Bahasa Indonesia dan satu berkas terkait WordNet Bahasa Malaysia. Ketiga berkas tersebut berformat *tab-delimited*. Ekstraksi data dilakukan menggunakan perangkat lunak pengolah angka (*spreadsheet software*). Pada makalah ini, hanya digunakan dua berkas WordNet Bahasa Indonesia. Satu berkas memuat data definisi dari suatu *synset*, sedangkan berkas lain memuat *lemma*-nya.

Pada tahap ini dilakukan proses pembersihan data. Setelah dilakukan pemeriksaan terhadap berkas WordNet Bahasa Indonesia yang memuat daftar *lemma*, terdapat sejumlah 119.904 baris data. Meski demikian, ditemukan sebanyak 12.668 baris data yang memiliki nilai *lemma* 0. Data yang baik seharusnya memuat nilai *lemma* berupa *string*/karakter. Dengan demikian, data semacam ini perlu dibersihkan sebelum dilakukan proses selanjutnya.

Adapun berkas WordNet Bahasa Indonesia yang berisi definisi memuat sebanyak 12.668 baris data, dengan 12.663 baris data memiliki *synset-id* (pengidentifikasi suatu *synset*) yang unik. Sedangkan lima sisanya meskipun tidak unik, tetapi memiliki definisi yang berbeda. Hal tersebut tidak dianggap sebagai *dirty data*, sehingga tidak diperlukan proses pembersihan data untuk berkas ini.

C. Transformasi Data

Setelah dilakukan proses ekstraksi dan pembersihan data, langkah selanjutnya adalah melakukan transformasi data. Transformasi data dilakukan dengan melakukan format ulang pola *synset-id* dari data sumber menjadi pola *synset* yang sesuai dengan pola *synset* yang ada pada WordNet-RDF.

Pola *synset-id* pada data sumber memiliki format sebagai berikut.

```
[8 digit angka]-[n|v|a|r|s|p]
```

dengan [n|v|a|r|s|p] merupakan *part-of-speech*. Adapun pola *synset-id* pada PWN-RDF menggunakan pola

```
[9 digit angka]-[n|v|a|r|s|p]
```

dengan digit pertama pada pola tersebut mewakili kode angka untuk *part-of-speech* sebagaimana ditunjukkan pada Tabel II.

TABEL II
DAFTAR KODE *PART-OF-SPEECH* VERSI WORDNET

Part-of-speech	**Kode huruf**	**Kode angka**
Noun (kata benda)	n	1
Verb (kata kerja)	v	2
Adjective (kata sifat)	a	3
Adverb (kata keterangan)	r	4
Adjective satellite	s	3
Phrase (frasa)	p	4

Pada proses transformasi data ini, dilakukan penambahan satu digit angka *synset-id* dari setiap data dari data sumber agar sesuai dengan format PWN-RDF. Sebagai contoh, suatu *synset-id* pada data sumber adalah 00001740-a. Merujuk pada Tabel II, maka hasil transformasi *synset-id* tersebut menjadi 300001740-a. Proses ini dilakukan agar setelah WordNet Bahasa Indonesia berada dalam format RDF, data tersebut dapat ditautkan dengan PWN-RDF.

D. Pemuatan Data ke Basis Data Relasional

Setelah data sumber bersih dan sudah dalam kondisi sesuai dengan format yang diinginkan, selanjutnya dilakukan pemuatan data-data dalam format teks ke dalam tabel-tabel basis data relasional. Pada kasus ini, digunakan DBMS MariaDB. Adapun struktur tabel yang digunakan untuk pemuatan data ini ditunjukkan pada Tabel III dan Tabel IV. Selanjutnya, data dimuat ke dalam basis data relasional menggunakan perintah *LOAD DATA INFILE*, sebagaimana ditunjukkan pada Gbr. 1. Pada Gbr. 1, terlihat terdapat dua berkas yang dimuat ke dalam basis data, yakni data terkait *lemma* dan data terkait definisi.

TABEL III
STRUKTUR TABEL WORDNETINDO_LEMMA

No.	**Nama Kolom**	**Tipe data**
1	id	Int(11)
2	sysnsetid	Char(11)
3	language	Char(3)
3	lemma	Varchar(100)

TABEL IV
STRUKTUR TABEL WORDNETINDO_DEF

No.	**Nama Kolom**	**Tipe data**
1	id	Int(11)
2	sysnsetid	Char(11)
3	definition	Varchar(100)

```
LOAD DATA LOCAL INFILE 'wnid-def_formatted.txt'
INTO TABLE wordnetindo_def
LINES TERMINATED BY '\r' (synsetid, definition);

LOAD DATA LOCAL INFILE 'wnid-lemma_formatted.txt'
INTO TABLE wordnetindo_lemma
LINES TERMINATED BY '\r' (synsetid, lemma,
language);
```

Gbr. 1 Perintah pemuatan data.

E. Pemetaan Data dari Model Relasional ke dalam Model RDF

Langkah yang dilakukan berikutnya adalah memetakan data yang berada pada model relasional menjadi data model RDF. Pada tahap ini, digunakan perangkat lunak D2RQ (http://d2rq.org). D2RQ merupakan suatu *platform* yang memungkinkan pengguna mengakses data dalam model RDF secara virtual dari model basis data relasional tanpa perlu menduplikasinya dalam suatu penyimpanan RDF.

Pada tahap ini, digunakan *generate-mapping tool*. *Tool* ini digunakan untuk menghasilkan sebuah berkas *mapping* dengan cara menganalisis skema dari sebuah basis data. Pada berkas ini, setiap tabel dipetakan menjadi sebuah kelas RDFS berbasis nama tabel. Selain itu, setiap kolom dari tabel tersebut menjadi properti sesuai nama kolomnya. Berkas ini dapat digunakan apa adanya atau dilakukan modifikasi.

```
# Table wordnetindo_lemma
map:wordnetindo a d2rq:ClassMap;
        d2rq:dataStorage map:database;
        d2rq:uriPattern "wnid/@@wordnetindo_lemma.synsetid@@";
        d2rq:class vocab:wnid_lemma;
        d2rq:classDefinitionLabel "wordnetindo";
        .
map:wnid_label a d2rq:PropertyBridge;
        d2rq:belongsToClassMap map:wordnetindo;
        d2rq:property rdfs:label;
        d2rq:pattern "wnid #@@wordnetindo_lemma.synsetid@@";
        .
map:wnid_synsetid a d2rq:PropertyBridge;
        d2rq:belongsToClassMap map:wordnetindo;
        d2rq:property vocab:synsetid;
        d2rq:propertyDefinitionLabel "wordnetindo_lemma synsetid";
        d2rq:column "wordnetindo_lemma.synsetid";
        .
map:wnid_language a d2rq:PropertyBridge;
        d2rq:belongsToClassMap map:wordnetindo;
        d2rq:property vocab:language;
        d2rq:propertyDefinitionLabel "wordnetindo_lemma lang";
        d2rq:column "wordnetindo_lemma.lang";
        .
```

Gbr. 2 Sebagian isi berkas *mapping*.

Dalam makalah ini, dilakukan beberapa modifikasi terkait berkas *mapping* yang dihasilkan tersebut. Contoh sebagian isi dari berkas konfigurasi ditunjukkan pada Gbr. 2. Beberapa perubahan yang dilakukan adalah sebagai berikut.

1. Penamaan kelas yang semula secara *default* berlabel ”wordnetindo_lemma”, sesuai nama tabel pada basis data relasional, menjadi ”wordnetindo”.
2. Kelas wordnetindo_def yang memuat definisi dari suatu *lemma* tidak ditampilkan.
3. Kolom ”definition” pada tabel wordnetindo_def di-*mapping* menjadi properti ”definition” dari kelas wordnetindo. Hal ini dilakukan dengan menggunakan subproperti ”join” dari properti PropertyBridge.
4. Penamaan *instance* dari kelas ”wordnetindo” yang secara *default* memiliki label dengan format ”wordnetindo_lemma#”<id>, dengan ”id” merupakan *primary key* dari tabel wordnetindo_lemma, menjadi label dengan format ”wnid#”<*sysnset-id*>. Hal ini dilakukan untuk menyesuaikan dengan pola penamaan pada PWN-RDF.
5. Penambahan properti ”owl:sameAs” pada kelas wordnetindo untuk menerapkan prinsip LD ke-4. Nilai dari properti ini memiliki pola label ”http://wordnet-rdf.princetown. edu/wn31/”<*synset-id*>.

IV. HASIL DAN PEMBAHASAN

Untuk menguji hasil proses yang dilakukan, digunakan *tool* d2r-server yang merupakan bagian dari *framework* D2RQ. *Tool* ini berfungsi untuk mempublikasikan basis data relasional sebagai RDF. Melalui *tool* ini, pengguna dapat melakukan navigasi terhadap isi basis data dalam format RDF dan melakukan *query* dengan menggunakan bahasa SPARQL. *Data set* ini dapat diakses melalui alamat URL: http://hendrikworks.net:2020.

Tampilan semua *resource/instance* dari kelas wordnetindo dapat dilihat pada Gbr. 3. Meski demikian, pada setiap halaman, seperti pada Gbr. 3, hanya menampilkan sebanyak 100 data saja sesuai dengan konfigurasi *default*. Selanjutnya, jika diinginkan melihat detail dari suatu *resource*, dapat diklik pada salah satu *resource* yang ada. Misalnya dipilih *resource* dengan URI: http://hendrikworks.net:2020/data/resource/wnid/100004475-n, yang direpresentasikan dengan label wnid#100004475-n, akan ditampilkan detail dalam format HTML sebagaimana ditunjukkan pada Gbr. 4.

Gbr. 3 *Instance* kelas wordnetindo.

Gbr. 4 Detail informasi suatu *instance*.

Adapun jika diinginkan melihat detail data dalam format RDF, dapat diunduh dengan menekan ikon RDF yang terletak pada sudut kanan atas hasil tembak layar Gbr. 4. Selain itu, karena data set ini sudah dalam format LD, maka pengguna dapat memperoleh informasi detail terkait suatu *sysnset-id*

dalam versi PWN. Hal itu dapat dilakukan dengan memilih objek dari properti owl:sameAs pada Gbr. 4.

Framework D2RQ juga menyediakan akses ke *data set* dengan menggunakan *query* SPARQL. Hal ini dapat dilakukan menggunakan *tool* SNORQL, suatu AJAX-*based* SPARQL untuk mengakses data dalam format RDF dengan *query* SPARQL, sebagaimana dapat dilihat contoh suatu *query* SPAQRL pada Gbr. 5. Pada Gbr. 5 tersebut, ditunjukkan *query* SPAQRL untuk menentukan semua *lemma* atau kata dalam Bahasa Indonesia yang sesuai dengan *sysnet-id* "100004475-n". Hasil dari *query* ini disajikan dalam bentuk HTML. Selain disajikan dalam format HTML, keluaran *query* juga dapat disajikan dalam format lainnya, yakni JSON, XML dan XML+XSLT.

Gbr. 5 Akses SPARQL *query* dengan *tool* SNORQL.

Pengembangan model WordNet Bahasa Indonesia berbasis LD ini memberikan beberapa keuntungan. sebagai berikut.

1. Memungkinkan untuk mengakses *data set* dalam format yang *machine understandable*, sehingga hasil yang diperoleh dapat diproses secara otomatis oleh mesin. Hal ini dimungkinkan dengan penggunaan *platform* D2RQ yang memungkinkan keluaran dalam format JSON maupun XML. Pada penelitian-penelitian sebelumnya, umumnya pengguna hanya dapat mengakses *data set* melalui antarmuka berbasis *web* dengan format HTML. Kondisi ini akan menyulitkan pengguna jika menginginkan untuk memproses lebih lanjut hasil yang diperoleh.
2. Memungkinkan untuk dapat melakukan *query* secara langsung ke *data set* menggunakan SPARQL *query*. Hal ini dimungkinkan dengan adanya fasilitas SPARQL *endpoint*. Pengguna dapat menggunakan berbagai pustaka (*library*) untuk mengakses *data set* dalam format RDF tersebut sesuai dengan bahasa pemrograman yang digunakan. Pada penelitian yang sudah ada, pengguna hanya diberikan antarmuka HTML untuk melakukan pencarian suatu kata. Selain itu format serta atribut keluaran yang dihasilkan juga sudah *rigid* ditentukan oleh pengembang. Sementara dengan model ini, pengguna dapat secara fleksibel melakukan *query* terhadap properti yang diinginkan dengan merujuk pada struktur yang dapat dirambah terlebih dulu.

V. KESIMPULAN DAN SARAN

Dalam makalah ini telah berhasil dibuat sebuah model WordNet Bahasa Indonesia berbasis LD dengan menggunakan *platform* D2RQ. Hasil pengujian dengan melakukan perambahan data menggunakan *server* D2R dapat berjalan dengan baik. Demikian pula dengan pengujian *query* terhadap data dengan menggunakan SPARQL *query*, juga menghasilkan data sesuai yang diharapkan.

Makalah ini terlah berhasil menyediakan *data set* WordNet Bahasa Indonesia dalam format *machine understandable*. Menggunakan format ini, pengguna dapat melakukan otomatisasi proses daripada dengan menggunakan WordNet model tradisional.

Meskipun demikian, dalam makalah ini belum dilakukan pengujian pemanfaatan *data set* untuk berbagai kasus aplikasi penelitian di bidang NLP. Di masa mendatang, penelitian dapat difokuskan pada pemanfaatan *data set* ini untuk beberapa kasus pemanfaatan WordNet dengan membandingkan model tradisional dengan model berbasis RDF/LD.

UCAPAN TERIMA KASIH

Penelitian dalam makalah ini dibiayai oleh Direktorat Riset dan Pengabdian Masyarakat, Kementerian Riset, Teknologi dan Pendidikan Tinggi, sesuai dengan surat Perjanjian Pelaksanaan Penelitian nomor: 001/HB-LIT/III/2016.

REFERENSI

[1] Nurudin, "Media Sosial Baru dan Munculnya Revolusi Proses Komunikasi", *Jurnal Komunikasi*, vol. 5(2), hal. 127–142, 2013.

[2] Miller, G.A., Beckwith, R., Fellbaum, C., Gross, D. dan Miller, K. J., "Introduction to WordNet: an Online Lexical Database", *International Journal of Lexicography*, vol. 3(4), hal. 235–244, 1990.

[3] Mahendra, R., Larasati, S. D., dan Manurung, R., "Extending an Indonesian Semantic Analysis-based Question Answering System with Linguistic and World Knowledge Axioms.", *Prosiding the 22nd Pacific Asia Conference on Language, Information, and Computation*, hal. 262–271, 2008.

[4] Clark, P., Fellbaum, C., dan Hobbs, J., "Using and Extending WordNet to Support Question- Answering 2 Semantic Requirements on WordNet", *Prosiding the 4th Global WordNet Conference*, 2008.

[5] Abouenour, L., Bouzoubaa, K., dan Rosso, P., "Improving QA Using Arabic WordNet", *Prosiding the 2008 International Arab Conference on Information Technology*, 2008.

[6] Andreevskaia, A., dan Bergler, S., "Mining WordNet for Fuzzy Sentiment : Sentiment Tag Extraction from WordNet Glosses", *European Chapter of the Association for Computational Linguistics.*, vol. 6, hal. 209–216, 2006.

[7] Montejo-Ráez, A., Martínez-Cámara, E., Martín-Valdivia, M. T., dan Ureña-López, L. A., "Ranked WordNet Graph for Sentiment Polarity Classification in Twitter", *Compututer Speech & Language*, vol. 28(1), hal. 93-107, 2014.

[8] Elberrichi, Z., Rahmoun, A., Bentaalah, M. A., dan Arabia, S., "Using

WordNet for Text Categorization", *International Arab Journal of Information Technology*, vol. 5(1), hal. 16–24, 2008.

[9] Sriram, B., Fuhry, D., Demir, E., dan Demirbas, H. F. M., "Short text Classification in Twitter to Improve Information Filtering", *Prosiding the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, hal. 841-842, 2010.

[10] Elkateb, S., Black, W., Vossen, P., Farwell, D., Rodriguez, H., Pease, A., dan Alkhalifa, M., "Arabic WordNet and the Challenges of Arabic", *Prosiding Arabic NLP/MT Conference*, hal. 15–24, 2006.

[11] Koide, S., Takeda, H., Kato, F., Ohmukai, I., Bond, F., Isahara, H. dan Kuribayashi, T., "DBpedia and Wordnet in Japanese", *Semantic Web Journal*, vol. 1, hal. 4–7, 2009.

[12] Putra, D. D., Arfan, A., dan Manurung, R., "Building an Indonesian Wordnet", *Prosiding the 2nd International MALINDO Workshop.*, 2008.

[13] Hirfana, N., Noor, M., Sapuan, S., dan Bond, F., "Creating the Open Wordnet Bahasa", *Prosiding the 25th Pacific Asia Conference on Language, Information, and Computation*, hal. 255–264, 2011.

[14] Riza, H., Budiono, dan Hakim, C., "Collaborative Work on Indonesian Word et through Asian WordNet (AWN)", *Prosiding the 23rd International Conference on Computational Linguistics*, hal. 9–13, 2010.

[15] van Assem, M., Gangemi, A., dan Schreiber, G., "Conversion of WordNet to a standard RDF / OWL representation", *Prosiding the 5th International Conference on Language Resources and Evaluation*, hal. 237–242, 2006.

[16] Chiarcos, C., Cimiano, P., dan Declerck, T., "Linguistic Linked Open Data (LLOD) Introduction and Overview", *Prosiding the 2nd Workshop on Linked Data in Linguistics*, 2013.

[17] Yu, L., *A Developer's Guide to the Semantic Web Programming*. Heidelberg: Springer, 2011.

[18] Heath, T. dan Bizer, C., *Linked Data: Evolving the Web into a Global Data Space*. California: Morgan & Claypool Publishers, 2011.

[19] Hendrik dan Perdana, D. H. F., "Trip Guidance: a Linked Data Based Mobile Tourists Guide", *Advanced Science Letters*, vol. 20(1), hal. 75–79, 2014.

[20] Chiarcos, C., McCrae, J., Cimiano, P. dan Fellbaum, C., *Towards Open Data for Linguistics: Linguistic Linked Data*. Springer, Berlin, 2013.