Mansoura Journal of Computers and Information Sciences

ETL Semantic Model for Big Data Aggregation, Integration, and Representation

Abeer Saber
Faculty of computers and information systems , C.S dep.
Kafr El-Sheikh University, Kafr El-Sheikh 33511, Egypt
abeer_saber@fci.kfs.edu.eg

Aya M. Al-Zoghby
Faculty of computers and information systems , C.S dep.
Mansoura University Mansoura 35516, Egypt
aya_el_zoghby@mans.edu.eg

Samir Elmougy
Faculty of computers and information systems , C.S dep.
Mansoura University, Mansoura 35516, Egypt
mougy@mans.edu.eg

ABSTRACT
Semantic web introduces new benefits for many research topics on big-data. It semantically maintains a large amount of data and provides meaningful meaning of unstructured data contents. Big data refers to large scale. It is used to describe a massive collection of datasets in different formats. The semantic and structural heterogeneity are the biggest problems that still face the aggregating, integrating, and storing big data. In this paper, we solved both of the problems of columns redundancy that are produced from the semantic heterogeneity and the problem of structural heterogeneity through developing and implementing a new ETL model based on semantic and ontology technologies. Geospatial data is used as a case study because its integration is complex and usually suffers from the variety of resources and the representation of the produced big data. The results of using this model showed that it solves the problem of heterogeneity in several data sources and it improves the data integration and representation.

General Terms
Semantic web, big data, ETL models, linked data.

Keywords
Semantic web, big-data, ontology, structural heterogeneity, semantic heterogeneity

1. INTRODUCTION
Semantic Web (SW) is a mesh of data representing meanings through connectivity and using logical rules to share information across applications depending on a single and main building block called ontology [1] [2]. Ontology is the backbone of SW which describes its metadata schemas, provides concepts vocabulary, describes the entity types in the world, and describes their connections [3]. It also provides a shared understanding of data services and processes, and conceptualizes scope knowledge and modeling that can be used to show the data semantics, and thus plays a role in integrating semantic data.

Big data is an evolving term that describes any huge amount of different data collection in different formats that has the potential to be mined for information [2] [4]. In order to debate about research issues for big data improvement, it is crucial to better understand the characteristics that define it such as volume, variety, velocity, value, and veracity [٥]. The users can not be able to access most of the big data as the existed traditional big data techniques could not process efficiently [٦].

Extract-Transform-Load (ETL) process is one of the most common approaches to integrating data that is composed of three main components: 1) Extract: data are extracted from varied data resources in which data is usually available in a flat file format such as CSV, XLS, and TXT or is available through a RESTful client. 2) Transform: some of the typical transformation activities are applied such as data normalization, duplication detection, integrity violation check, some regular expressions for filtering data, sorting, and grouping data 3) Load: the transformed data is stored in the data warehouse.

The variety of data involves different formats such as structured, semi-structured, and unstructured data in which the last two formats are more than “85%” that causes the problems of structural and semantic heterogeneity among data sources. Structural heterogeneity of several data stored in the data model cannot be directly transformed to each other. Semantic heterogeneities of data mean that description of the terms is inconsistent with each other and the reflecting the link between datasets is incapable. Hence, the schema mapping and linking record are the major problems facing big data integration and representation [٧]. Big data faces many challenges such as 1) acquisition and recording of data, 2) extracting and cleaning information, 3) integrating, aggregating, and representing data, 4) queries analysis and processing [٨].

Currently, there is a moving from the era of "data on the internet" to an era of "web of data (linked data)". Linked Data (LD) tries to create the Web into a global database. LD indicates to a group of best practices to publish and interlink data on the Web [9]. To create LD, it is necessary to have data available on the Web in a standard, reachable, and manageable format. In addition, the relationships between data are needed.

LD is depending on some of SW technologies and Hypertext Transfer Protocol (HTTP) to publish structured data on the Web and to connect data between different data sources, effectively allowing data in one data source to be

linked to data in another data source [10]. SW is designing principles for sharing machine-readable interlinked data on the Web. This links between different datasets making understandable not only to humans but also to machines.

In this work, we propose a new semantic ETL model for data integration, aggregation, and representation improvement using SW technologies. This proposed model that is based on ontology matching overcomes both the problems of columns redundancy that are produced from the semantic heterogeneity and the problem of structural heterogeneity. Through this work and its experiments, we use some different tools and packages as 1) Uniform Resource Identifier (URI) that is used as the criterion to identify the name and location of a file or resource in a uniform format [11], 2) RDF helps data merging of various schemas and allows various data to be mixed [12] [13], 3) OWL is built upon RDF for processing information on the internet [14], 4) SPARQL is a query language and protocol for RDF facilitates the distributed RDF data query; it is used in this research for data properties extraction [15], 5) Protégé "ontology editor" is used to build ontologies for knowledge-based applications [16].

Nowadays, Geospatial data becomes an important domain for many decision makers. Its integration becomes more complex especially with and usually suffers from the diversity of format representation [21]. Geospatial ontology supports the meaning of features and objects that are saved its database.

Section 2 discusses some related works. Section 3 illustrates a case study for geospatial data integration problems. Section 4 introduces the proposed workflow of ETL. In Section 5, the experimentation is presented while the conclusion is given in Section 6.

2. RELATED WORK

Most of the technical difficulties that usually appear when dealing with big data integration are because of the data formats variety including **structured and semantic**. To solve this difficulty, many existing works were dependent on SW and metadata. Semantic technologies are added recently to ETL process to solve these problems.

Jiang et al. [17] designed a semantic ETL process using ontologies to catch the semantics of domain model and resolve semantic heterogeneity. This model assumes that the data resources type is the only relational database. Huang et al. [18] automatically extracted data from different marine data resources and transformed it to unified schemas relying on the applied database to semantically integrate it. Srividya et al. [٦] and Bansal et al. [٤] produced a semantic ETL process for integrating and publishing structured data from various sources as LD by inserting semantic model and instance into transforming layer using the OWL, RDF, and SPARQL technologies. While Bergamaschi et al. [19] enhanced the ETL definitions by allowing semantic transforming of semi-automatic inter-attribute through recognizing data sources schemas and semantically grouping attribute values.

Zhang et al. [20] introduced a semantic approach for extracting, linking, and integrating geospatial data from several structured data sources. It also solves the individuals' redundancy problem which faces the data integration. The basic idea of this model is using ontologies to convert the extracted data from different sources to RDF format followed by linking similar entities in the generated RDF files using the linking algorithm. Next step is to use SPARQL queries to eliminate data redundancy and combine complementary properties for integration using integration algorithm. Du, H., et al. [21] developed technique for solving the redundancy problems between individuals in data integration using SW technologies.

Souza et al. [22] and Yunianta, Arda, et al. [23] discussed the mapping between data schemas in which the mapping process between columns name is adjusted manually. Xiong et al. [24] proposed and implemented a new model to aggregate the online educational data sources from the internet and mobile networks using some semantic techniques like ontologies and metadata to enhance the aggregation results. Gollapudi and Sunila [25] built an architecture that combined the data lake with semantic technologies for aggregating data from various sources.

Kang et al. [١] constructed a semantic big data model **which in line** with the map reduces framework to semantically store data. However, this model did not support the integration of data from existing database system. Semantic models for data aggregation, integration, and representation is still rare in research and suffers from many challenges such as semantic and structural heterogeneities. Here, we aim to solve these problems and to enhance the aggregation, integration, and representation of big data using some of SW techniques.

3. A CASE STUDY

To clarify the proposed semantic ETL process, a case study of geospatial cities data is presented. Geospatial data sources are aggregated from various data sources over the internet. Data used in this paper is from "Gaslamp media" [26], and "OST/SEC Map group" [27]. Gaslamp media provides cities information such as country name, zip-code, longitude, latitude, state, and city. OST/SEC Map group provides information such as country_fip, ST, State, LON, LAT, name, and ZPROG_DISC. This data in these sources is suffering from the semantic heterogeneity such as (longitude, LON), (latitude, LAT), and (state, ST). So, it will be more useful to overcome this semantic heterogeneity problem to make this data integrated successfully before storing it.

4. THE PROPOSED ETL MODEL

Our proposed semantic ETL model, shown in Fig. 1, aims to semantically aggregate geospatial resources from the internet for integrating and restoring it in the semantic model. Aggregating resources using a geospatial ontology for aggregation is firstly processed as presented in Fig. 2, while metadata is presented in [24]. We adopted this model to aggregate geospatial data resources using *Geospatial data aggregation ontology*, as presented in Fig. 3, and then the three ETL phases are applied. Extracting data from the aggregated geospatial resources is the first phase. This extracted data is different from each other and have different schemas. Hence, there is no semantic meaning for it, and their structures are different. Therefore, to align and link this data, some efficient SW tools are applied in the second phase.

The main purpose of the second phase is to transform the extracted data to RDF format for linking. It consists of five processes: the first process is data preparation that contains some typical transformation activities to prepare the data including normalizing data, removing duplicates, integrity violation checks, filtering, sorting, grouping etc. Next, data attributes are aligned according to the constructed ontology then data is transformed to RDF format using the proposed *Trans–Data-to–RDF algorithm* based on geospatial ontology shown in Fig 4. However, this algorithm transforms CSV files only. Thereby, the structured and semi-structured data files like XML, EXCEL, JSON, and databases are first transformed to CSV format using some online tools [28] [29] before applying the algorithm. The generated RDF files data files are integrated using integration algorithm in [20]. Finally, the integrated data is stored in the data warehouse using the methodology reported in [1].

The proposed *Trans–Data-to–RDF algorithm* depends on alignment API [30] [31] and Jena ontology API [32] to transform the geospatial CSV data to RDF format. First, it extracts the file name and considers it as the class name. Then, it extracts the column name and every row's data to consider them as data properties and individuals respectively. The next step is measuring the similarities between the data properties name in both of ontology generated from CSV data and inputs ontology to solve the semantic heterogeneity problem between them with an automatic mapping between columns name. Depending on the value of the similarities measure, the algorithm converts the data properties in the source data to their correspondence in the input ontology (when the measured value is more than 0.5). Finally, it generates the RDF file with the new data properties name. Table 1 presents the differences between the existing models and our proposed model.

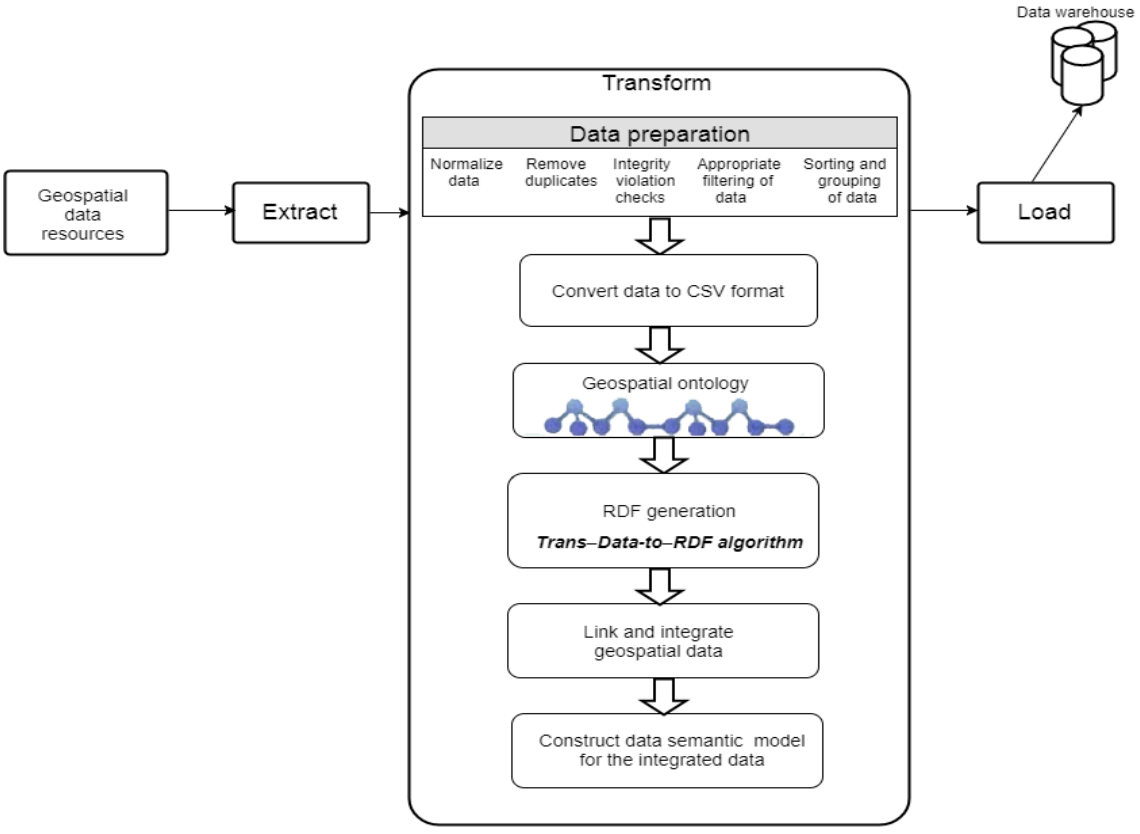

Fig 1: The proposed ETL model

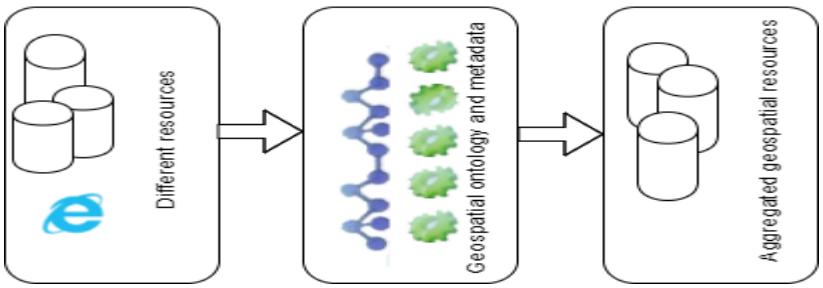

Fig 2: Geospatial data resources aggregation

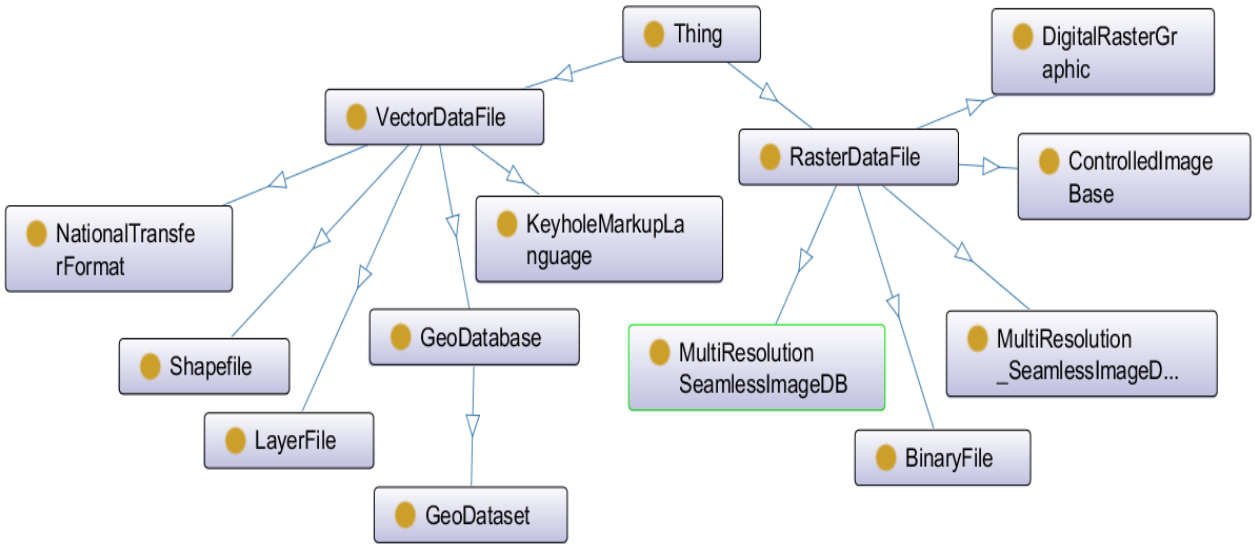

Fig 3: Geospatial data aggregation ontology

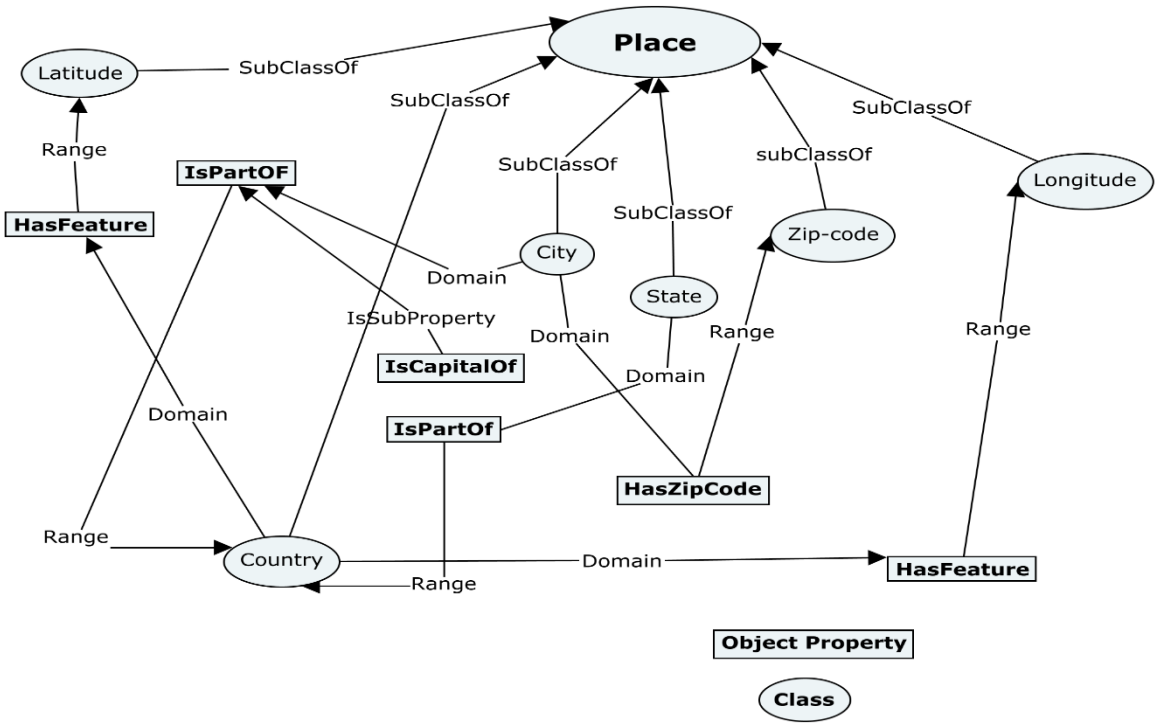

Fig 4: Geospatial ontology

	Trans–Data-to–RDF algorithm
1	**Input:** $src_{1:}$ Geospatial CSV file,
2	$src_{2:}$ Geospatial Ontology file
3	**Output:** real: alignmentmeasure_value // matching value between entities in src_1 and src_2 files,
4	RDF data file generation
5	**Variables:**
6	file: CSV_File // read the CSV data
7	string: RDF_className // class name of the generated RDF data
8	array: Column_Listing // list for storing all columns name
9	string: column_name // string store column name in the input CSV file
10	string: column_data // string store every value of the CSV data
11	object: csvModel // model to hold the RDF data which generated from the Geospatial CSV input file
12	object: geospatial_ontology // model to hold the Geospatial Ontology file
13	object: dataPropertys // object from DatatypeProperty class
14	object: csvIndividual // object for creating individuals
15	string: property // string to get the data property name from the Colomn_Listing
16	object : First_ontology // object from JENAOntology
17	object : Second_ontology // object from JENAOntology
18	object: alignmentmeasure // for creating the alignment process between First_ontology and Second_ontology
19	int : counter // initial value equal 0
20	string: *entity1 // hold the* data properties name of the First_ontology in each cell
21	string: *entity2 // hold the* data properties name of the Second_ontology in each cell
22	**Processing:**
23	**Begin:**
24	// First Stage:
25	//First Step: Read Geospatial CSV file
26	CSV_File ← src_1
27	//Second Step: Convert data in CSV into RDF format

```
28  RDF class name= src1 name
29  // create the data properties from columns name
30      For all column_names in src1 {
31          column_name ← column value
32          DatatypeProperty dataPropertys =
              csvModel.createDatatypeProperty(column_name);
33          Coloumn_Listing.add(column_name); }
34      // set every row data as a new individual
35          while (! End-of-file(src1))
36          { column_data ← column value
37          Individual csvIndividual = cvsClass.createIndividual ();
38          String property = Coloumn_Listing.get (counter++);
39          csvIndividual.addProperty(csvModel.getDatatypeProperty(property), column_data); }
40  OntModel geospatial_ontology ← read src2
41  // Using "Alignment API" to calculate similarities
42  JENAOntology First_ontology = new JENAOntologyFactory().newOntology(csvModel, true);
    JENAOntology Second_ontology = new JENAOntologyFactory().newOntology(geospatial_ontology, true);
43  // Aligning data properties between two ontologies
44  AlignmentProcess alignmentmeasure = new SMOANameAlignment();
45  alignmentmeasure.init (First_ontology, Second_ontology); // takes the source and target ontology to
    alignment
46  alignmentmeasure.align (First_ontology.getdataproperties(), Second_ontology.Countryclass.
    getdataproperties());
47  For all cells c in alignmentmeasure
48  {
49  alignmentmeasure_value = cell.getStrength(); // get measre value
    entity1= cell.getObject1().toString(); // get data property name of First_ontology
50  entity2= cell.getObject2().toString(); // get data property name of Second_ontology
51      If (alignmentmeasure_value >0.5)
52      {
            entity1.value ← entity2.value;
53      }
54  }
55  // Second Stage
56  Save First_ontology as RDF format in xml file
    END
```

Table 1. Comparison between existing semantic models and the proposed model

Paper	Aggregation	Integration	Representation	Individuals Redundancy	Columns Redundancy	Linking
Souza et al. (2006)	×	×	×	×	✓ (manually on the ontology)	×
Du, H., et al. (2011)	×	×	×	✓	×	×
Zhang et al. (2013)	×	✓	×	✓	×	✓
Xiong et al.(2014)	✓	×	×	×	×	×
Gollapudi and Sunila (2014)	✓	×	×	×	×	×
Kang et al. (2014)	×	×	✓	×	×	×
Bansal et al. (2014)	×	✓	×	×	×	✓
Yunianta, Arda, et al. (2017)	×	✓	×	×	✓ (manual)	✓
Our proposed model	✓	✓	✓	×	✓ (Trans–Data-to–RDF) algorithm	✓

5. Experimentations and Results

We solved the **semantic heterogeneity** problem of data in sources [26] and [27] by measuring the similarities between their attributes and the applied geospatial ontology. Table 2 shows the result of the first stage of applying our proposed algorithm. In **Gaslamp media** [dataset1] [26], all the similarity values are equal to 1 because the names of the ontology attributes are similar to the names of geospatial dataset attributes. But in **OST/SEC Map group** [dataset2] [27], the similarity value between the *lon* attribute and the *longitude* in constructed ontology is 0.825, which indicates that lon and longitude are the same data. Also, between lat and latitude is 0.84. Although the similarity value between zprog_disc and zip_code is 0.05 which also indicates that both are different.

Fig. 6 and Fig. 7 shows the second stage of applying the proposed algorithm in which CSV input data is converted to RDF format to solve **structural heterogeneity** problem for semantically integrating. According to the similarities result of the first stage, the attributes names in the generated RDF file have been determined through the measured values of similarity. For example, in the dataset [27], (lon, lat) attributes are transformed to (longitude, latitude) because the similarity values are more than 0.5. Also, the (zprog_disc, st) are not transformed to (zip_code, state) because the measuring values are less than 0.5.

Table 2. The results of similarity between" Gaslamp media" dataset, OST/SEC Map group dataset attributes and ontology

Ontology Attributes	"Gaslamp media" **[Dataset1]** Attributes	OST/SEC Map group **[Dataset2]** Attributes	Similarity Value between **[Dataset1]** and ontology	Similarity Value between **[Dataset2]** and ontology
State	State	State	1	1
State	×	St	×	0.09
Country	Country	×	1	×
City	City	Name	1	×
Longitude	Longitude	Lon	1	0.825
Latitude	Latitude	Lat	1	0.84
Zip_code	Zip_code	Zprog_disc	1	0.05

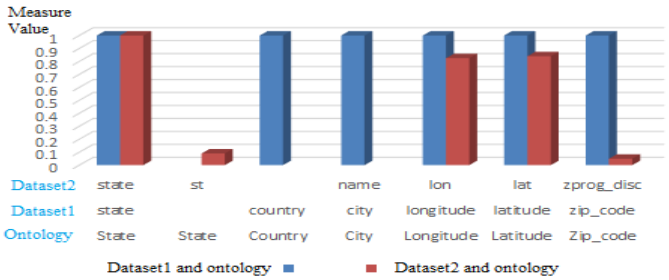

Fig 5: Similarity values between" Gaslamp media" dataset, OST/SEC Map group dataset attributes and ontology

<owl:NamedIndividual rdf:about="individual 1">

<rdf: type rdf: resource="Country"/>

<zip_code> 501.0 </zip_code>

<latitude> 40.0922326 </latitude>

<longitude> -27.637078 </longitude>

<city> Holtsville </city>

<state> NY </state >

<country> Suffolk </country >

</owl:NamedIndividual>

Fig 6: Converting CSV data in "Gaslamp media" dataset to RDF format by Trans–Data-to–RDF algorithm

<owl:NamedIndividual rdf:about="individual 1">

<rdf: type rdf: resource="Country"/>

<name> East calais </name>

<st> VT </st>

<state> Vermont </state>

<zprog_disc> 50 </zprog_disc >

<longitude> - 72.4303 </longitude>

<latitude> 44.3664 </latitude>

</owl:NamedIndividual>

Fig 7: Converting CSV data in "OST/SEC Map group" dataset to RDF format by Trans–Data-to–RDF algorithm

If the structured data is in RDF format, then we use SPARQL for extracting entities in the dataset as shown in Fig. 8.

```
SELECT DISTINCT ?properties
WHERE {
s:Country    rdfs:subClassOf*  ?superclass.
?properties  rdfs:domain       ?Country.
?properties  rdf:type          ?propertyType
}
```

properties
longitude
latitude
country
state
city
zip_code

Fig 8: Attributes extraction from RDF files using SPRQL

6. CONCLUSION

The proposed model in this paper solves the structural and semantic heterogeneity among several resources of data and improves the data storing in the data warehouse by proposing an algorithm based on ontology matching. This allows the user to aggregate, link, integrate, and store the data from several geospatial data sources in "structured and semi-structured" formats using semantic web technologies. First, aggregating the geospatial data resources semantically and then integrating it semantically. Finally, data are represented and restored semantically. A case study of Geospatial data is used to illustrate our work because this data integration is still a complex task due to its diversity of resources format and representation. So, it usually suffers from the semantic **heterogeneity**. The experimental results showed that the proposed model with its used algorithm overcame the

problems of columns redundancy that are produced from the semantic heterogeneity and the problem of structural heterogeneity.

We plan to work on improving our algorithm to solving the redundancy problem between individuals for improving the integration process.

7. REFERENCES

[1] Kang, Li, Li Yi, and L. Dong.: Research on Construction Methods of Big Data Semantic Model. Proceedings of the World Congress on Engineering (WCE 2014). Vol. 1, London, UK (2014).

[2] Ahmed, Zeeshan, and Detlef Gerhard. Web to Semantic Web & Role of Ontology. arXiv preprint arXiv: 1008.1331 (2010).

[3] Jain, Vishal, and Mayank Singh.: Ontology-based information retrieval in semantic web: A survey. International Journal of Information Technology and Computer Science (IJITCS) 5(10), 62 (2013).

[4] Bansal, Srividya K.: Towards a semantic extract-transform-load (ETL) framework for big data integration. In: Big Data (BigData Congress), 2014 IEEE International Congress on. IEEE, pp. 522-529. Anchorage (2014).

[5] Bizer, C., Boncz, P., Brodie, M. L., & Erling, O.: The meaningful use of big data: four perspectives--four challenges. ACM SIGMOD Record, 40(4), 56-60 (2012).

[6] Srividya K. Bansal, Sebastian Kagemann.: Integrating Big Data: A Semantic Extract-Transform-Load Framework. Computer 48(3): 42-50 (2014).

[7] Arputhamary, B., & Arockiam, L.: A Review on Big Data Integration. Int J Comput Appl, 21-26 (2014).

[8] BERTINO, Elisa.: Big Data – Opportunities and Challenges. IEEE 37th Annual Computer Software and Applications Conference, pp. 479-480. Kyoto, Japan (2013).

[9] Linked open data Homepage, https://ontotext.com/knowledgehub/fundamentals/linked-data-linked-open-data/, last accessed 2017/10/15.

[10] Bizer, C., Heath, T., Idehen, K., & Berners-Lee, T. Linked data on the web (LDOW2008). In Proceedings of the 17th international conference on World Wide Web, pp. 1265-1266. Beijing, China. ACM (2008).

[11] RDF Homepage, https://www.w3.org/RDF/ , last accessed 2017/10/15.

[12] RDF Homepage, http://www.webopedia.com/TERM/R/RDF.html, last accessed 2017/10/1.

[13] OWL Homepage, https://www.w3.org/2001/sw/wiki/OWL, last accessed 2017/10/21

[14] Wu, Hongyan, and Atsuko Yamaguchi.: Semantic Web technologies for the big data in life sciences. Bioscience trends 8(4), 192-201 (2014).

[15] SPARQL Query Language for RDF Homepage, https://www.w3.org/TR/rdf-sparql-query/, last accessed 2017/10/20.

[16] Protégé Homepage, Http://protegewiki.stanford.edu/wiki/Main_Page, last accessed 2017/10/19.

[17] Jiang, L., Cai, H., & Xu, B.: A Domain Ontology Approach in the ETL Process of Data Warehousing, e-Business Engineering (ICEBE), 2010 IEEE 7th International Conference, pp. 30-35. Shanghai (2010).

[18] Huang, O. R., Du, Y. L., Zhang, M. H., & Zhang, C.: Application of ontology-based automatic ETL in marine data integration. IEEE Electrical & Electronics Engineering (EEESYM), Symposium on, pp. 11-13. Kuala Lumpur, Malaysia (2012).

[19] Bergamaschi, S., Guerra, F., Orsini, M., Sartori, C., & Vincini, M.: A semantic approach to etl technologies. Data & Knowledge Engineering, 70(8), 717-731 (2011).

[20] Zhang, Y., Chiang, Y. Y., Szekely, P., & Knoblock, C. A.: A semantic approach to retrieving, linking, and integrating heterogeneous geospatial data. In Joint Proceedings of the Workshop on AI Problems and Approaches for Intelligent Environments and Workshop on Semantic Cities, pp. 31–37. ACM (2013).

[21] Du, H., Jiang, W., Anand, S., Morley, J., Hart, G., Leibovici, D., & Jackson, M. An ontology based approach for geospatial data integration of authoritative and crowd sourced datasets. In Proceedings of the 25th International Cartographic Conference (2011).

[22] Souza D., Salgado A.C., Tedesco P" Towards a Context Ontology for Geospatial Data Integration. In: Meersman R., Tari Z., Herrero P. (eds) On the Move to Meaningful Internet Systems 2006: OTM 2006 Workshops. OTM 2006. Lecture Notes in Computer Science, vol 4278, 1576-1585Springer, Berlin, Heidelberg, (2006).

[23] Yunianta, A., Barukab, O. M., Yusof, N., Dengen, N., Haviluddin, H., & Othman, M. S.: Semantic data mapping technology to solve semantic data problem on heterogeneity aspect. International Journal of Advances in Intelligent Informatics, 3(3), 161-172 (2017).

[24] Xiong, Jing, Yuntong Liu, and Wei Liu.: "Ontology-Based Integration and Sharing of Big Data Educational Resources." IEEE 11th Web Information System and Application Conference (WISA), pp. 245–248. Tianjin, China (2014).

[25] Gollapudi, Sunila. Aggregating financial services data without assumptions: A semantic data reference architecture." 2015 IEEE International Conference on Semantic Computing (ICSC), pp. 312-315. Anaheim, CA, USA (2015).

[26] Gaslamp media Homepage, https://www.gaslampmedia.com, last accessed 2017/10/١٩.

[27] OST/SEC Homepage, http://www.nws.noaa.gov/, last accessed 2017/10/20.

[28] Conversion Homepage, https://conversiontools.io/conversion/convert_xml_to_csv, last accessed 2018/1/13.

[29] Conversion Homepage, http://www.dbf2002.com/dbf-converter/, last accessed 2018/1/13.

[30] David, J., Euzenat, J., & Scharffe, F. Trojahn dos Santos, C.: The Alignment API 4.0, Semantic Web-

Interoperability, Usability, Applicability, 2 (1), 3-10 (2011).

[31] Euzenat, J.: An API for ontology alignment. In International Semantic Web Conference, pp. 698–712. Springer, Berlin Heidelberg (2004).

[32] Jena ontology API Homepage, https://jena.apache.org/documentation/ontology/, last accessed 2018/1/13