Computers, Materials & Continua
DOI:10.32604/cmc.2021.015293
Article

Tech Science Press

Using Semantic Web Technologies to Improve the Extract Transform Load Model

Amena Mahmoud[1,*], Mahmoud Y. Shams[2], O. M. Elzeki[3] and Nancy Awadallah Awad[4]

[1]Department of Computer Science, Kafrelshiekh University, Kafrelshiekh, Egypt
[2]Department of Machine Learning, Kafrelsheikh University, Kafrelshiekh, Egypt
[3]Department of Computer Science, Mansoura University, Mansoura, Egypt
[4]Department of Computer and Information Systems, Sadat Academy for Management Sciences, Cairo, Egypt
[*]Corresponding Author: Amena Mahmoud. Email: amena_mahmoud@fci.kfs.edu.eg
Received: 14 November 2020; Accepted: 18 February 2021

Abstract: Semantic Web (SW) provides new opportunities for the study and application of big data, massive ranges of data sets in varied formats from multiple sources. Related studies focus on potential SW technologies for resolving big data problems, such as structurally and semantically heterogeneous data that result from the variety of data formats (structured, semi-structured, numeric, unstructured text data, email, video, audio, stock ticker). SW offers information semantically both for people and machines to retain the vast volume of data and provide a meaningful output of unstructured data. In the current research, we implement a new semantic Extract Transform Load (ETL) model that uses SW technologies for aggregating, integrating, and representing data as linked data. First, geospatial data resources are aggregated from the internet, and then a semantic ETL model is used to store the aggregated data in a semantic model after converting it to Resource Description Framework (RDF) format for successful integration and representation. The principal contribution of this research is the synthesis, aggregation, and semantic representation of geospatial data to solve problems. A case study of city data is used to illustrate the semantic ETL model's functionalities. The results show that the proposed model solves the structural and semantic heterogeneity problems in diverse data sources for successful data aggregation, integration, and representation.

Keywords: Semantic web; big data; ETL model; linked data; geospatial data

1 Introduction

Big Data consists of data from billions to trillions of millions of persons, all from various sources (e.g., Web, customer contact center, social media, mobile data, sales, etc.). Usually, the material is loosely structured and is frequently outdated and unavailable. Big Data is transforming science, engineering, medicine, healthcare, finance, business, and ultimately society itself. Huge volumes of data for strategic economic gain, public policy, and new insight into a wide variety of technologies are available (including healthcare, biomedicine, energy, smart cities, genomics,

transportation, etc.). Most of this knowledge, however, is inaccessible to users because we need technologies and resources to discover, transform, interpret, and visualize data to make it consumable for decision-making [1,2].

Due to the variety of data that includes different formats such as structured, semi-structured, and unstructured data, it is difficult to be processed using traditional databases and software techniques. Therefore, efficient technology and tools are needed to process data to be consumable for decision-making especially that most of them are inaccessible to users, as shown in Fig. 1 [2].

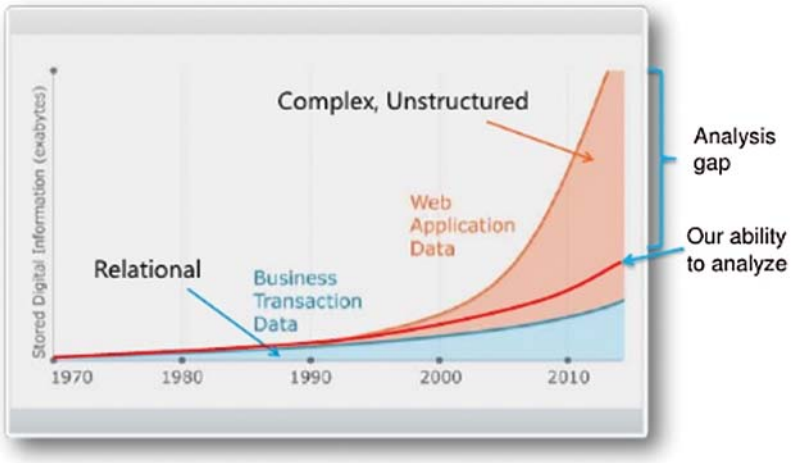

Figure 1: Processing data using traditional techniques

Nevertheless, meaningful data integration in a schema-less, and complex big data world of databases is a big open challenge. Big data challenges are not only in storing and managing this variety of data but also extracting and analyzing consistent information from it. Researchers are working on creating a common conceptual model for the integrated data [3]. The method of publishing and linking structured data on the web is called Linked Data [4]. This data is machine-readable, its meaning is explicitly defined, it is linked to other external data sets, and it can be linked to from other data sets as well [5].

Extract-Transform-Load (ETL) procedure is one of the most popular techniques in data integration. It covers the process of loading data from the source system to the data warehouse. This process consists of three consecutive stages: extracting, transforming, and loading, as shown in Fig. 2.

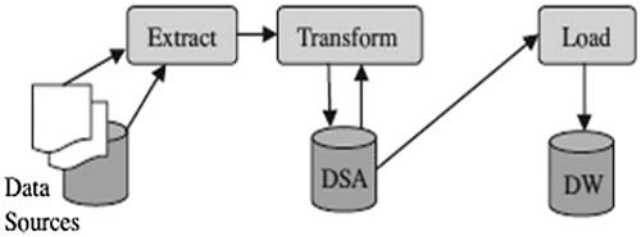

Figure 2: Traditional ETL model

To accurately exploit web data, a system needs to be capable to read the exact semantic meaning of web-published information. An acknowledged way to publish machine-readable information is to use Semantic web (SW) technologies. The purpose of SW technologies is to fix a common vocabulary and a set of interpretation constraints (inferring rules) to semantically express metadata over web information and allow doing some reasoning on it. More specifically, SW presents human knowledge through structured collections of information and sets of inference rules [6,7].

By using the SW formats, web resources can be enriched with annotations and other markups capturing the semantic metadata of resources. The first motivator of SW is data integration, which is a significant bottleneck in many IT applications. Current solutions to this problem are mostly ad hoc each time, a specific mapping is made between the data models (schemas) of the data sources involved. In addition to that, if the data sources' semantics were described in a machine-interpretable way, the mappings could be constructed at least semi-automatically. The second motivator is more intelligent support for end-users. If the computer programs can infer consequences of information on the web, they can give better support in finding information, selecting information sources, personalizing information, combining information from different sources, and so on.

Unlike the documentation of semantics, the approaches to the complex description of ETL problems are presented in the field of graphic modeling, however their scope of application is essentially limited, and the resulting benefits from the application are not high.

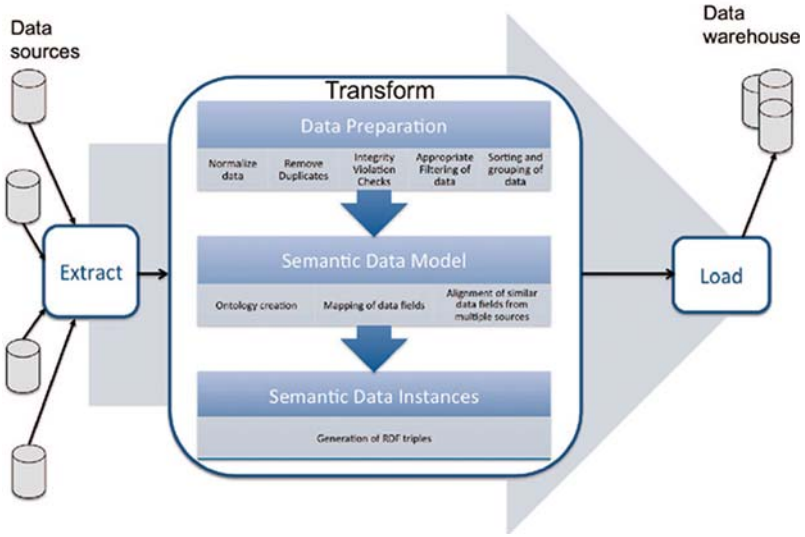

Figure 3: Semantic ETL model

Currently, we are moving from the era of "data on the web" to the era of "web of data (linked data)." Linked Data (LD) is introduced as a step in transforming the web into a global database. The term LD refers to a group of best practices for publishing and interlinking data on the web [8,9]. Creating LD requires having data available on the web in a standard, reachable, and manageable format. Besides, the relationships among data are required [10], as shown in Fig. 3. LD depends on some SW technologies and Hypertext Transfer Protocol (HTTP) to publish

structured data on the web and to connect data from different data sources to allow data in one data source to be linked to data in another data source effectively [11,12]. SW contains design principles for sharing machine-readable interlinked data on the web. These links for different datasets make them clearly understood not only for humans but for machines as well.

LD facilitates data integration and navigation through complex data owing to the standards to which it adheres. Guidelines allow easy upgrades and extensions to data models. Besides, representation under a set of global principles also increases data quality. Moreover, the database of semantic graphs representing LD creates semantic links between varied sources and disparate formats [13,14].

2 Related Work

Most of the technical difficulties that typically appear when dealing with big data integration results from a variety of data formats, including structured and semantic. Many existing studies depend on SW and metadata. Semantic technologies have been added recently to the ETL process to alleviate these problems.

Srividya et al. [15] designed a semantic ETL process using ontologies to capture the semantics of a domain model and resolve semantic heterogeneity. This model assumes that the data resources' type is the only relational database. Huang et al. [1] automatically extracted data from different marine data resources and transformed them into unified schemas relying on an applied database to integrate it semantically. Sonia et al. [16] and Lihong et al. [17] produced a semantic ETL process for integrating and publishing structured data from various sources as LD by inserting a semantic model and instances into a transforming layer using the OWL, RDF, and SPARQL technologies. Mahmoud et al. [18] enhanced the ETL definitions by allowing semantic transforming of semi-automatic, inter-attributes through the identification of data source schemes and semantic grouping of attribute values.

Mei et al. [19] introduced a semantic approach for extracting, linking, and integrating geospatial data from several structured data sources. It also solves the individuals' redundancy problem facing data integration. The basic idea of this model is to use ontologies to convert extracted data from different sources to RDF format followed by linking similar entities in the generated RDF files using the linking algorithm. The next step is to use SPARQL queries to eliminate data redundancy and combine complementary properties for integration using an integration algorithm. Isabel et al. [20] developed a technique for solving the redundancy problems between individuals in data integration using SW technologies.

Boury et al. [21] and Saradha et al. [22] discussed the mapping between data schemas in which the mapping process between column names is adjusted manually. Jadhao et al. [23] proposed and implemented a new model to aggregate online educational data sources from the internet and mobile networks using such semantic techniques as ontologies and metadata to enhance the aggregation results. Ying et al. [24] built a combined data lake using semantic technologies within architecture for aggregating data from numerous sources.

Kang et al. developed a semantic big data model that reduces the context for semantically storing data in line with a map. However, the inclusion of data from existing database structures has not been facilitated by this model. In science, semantic models for data aggregation, convergence, and representation are still uncommon and face many obstacles, such as semantic and structural heterogeneity. Here, we suggest using some semantic strategies to resolve these issues and improve the aggregation, integration, and representation of big data [25,26].

3 A Case Study

A case study of city data is used to explain the new workflow features. Internet contains numerous data services, such as MapCruzin group [27], Data.gov [28], United States Census [29], OST/SEK Map group [30], USCitiesList.org [31], and Gaslamp media [32]. Data are stored in these resources in different formats such as *shape_file*, *comma-separated values (CSV)*, and *DBF date file* data. The *OST/SEC GIs map group* data resource provides data such as *country_fip*, *ST*, *LON*, *LAT*, STATE, *name*, and *PROG_DISC*, while *data.gov* provides *country*, *countryfips*, *longitude*, *latitude*, *PopPlLat*, *PopPlLong*, *state*, and *state_fip*. *United States Census* provides *countryFP*, *name*, *Aland*, and *Awater*. Besides, *country*, *name*, *longitude*, *latitude*, *land area*, *water area*, *zip_codes*, and *area code* are provided in *USCitiesList.org*. The last data resource from *Gaslamp media* contains *zip_code*, *longitude*, *latitude*, *city*, and *state*.

Tab. 1 represents the semantic heterogeneity in these sources. However, these data will be more useful if it is represented and stored in a semantic model after integrating it semantically and then removing data duplications. Some data in these resources are the same but are referred to use different names such as (*city, name*), (*Aland, land area*), (*Awater, water area*), (*country_fip, countryFP, countryfips*), (*LON, longitude*), and (*LAT, latitude*). This incompatibility causes many problems in data integration and hence the generic geospatial ontology is applied to transform this data into RDF format for easily integrating using Jena and SPARQL query. The following step is to represent these data and to store them semantically in the semantic big data model.

Table 1: Semantic heterogeneity from diverse databases

Data resource	Attribute name					
	Country	City	Longitude	Latitude	State	Water area
Gaslamp media	Country	City	Longitude	Latitude	State	
Data.gov	Country	Name	Longitude	Latitude	State	
OST/SEC group		Name	LON	LAT	State	
US Census	CountryFP	Name			StateFP	Awater
USCities org	Country	Name	Longitude	Latitude	State	Water_area

4 Proposed Approach

The first approach proposed collecting geospatial data services by the geospatial ontology seen in Fig. 4. The suggested semantic model of ETL, shown in Fig. 5, aims to aggregate various geospatial data services from the network semantically and combine the derived resource data semantically to store it as a geospatial semantic big data model. Next, metadata is combined over the internet using geospatial data resources from various resources, as shown in Fig. 6. Then, the three steps of the ETL are performed.

The first phase is extracting data from the aggregated geospatial resources. These extracted data are different from each other and have different schemas. Hence, they have no semantic meaning, and their structures are different. We use SW technologies in the second phase to align and link this data.

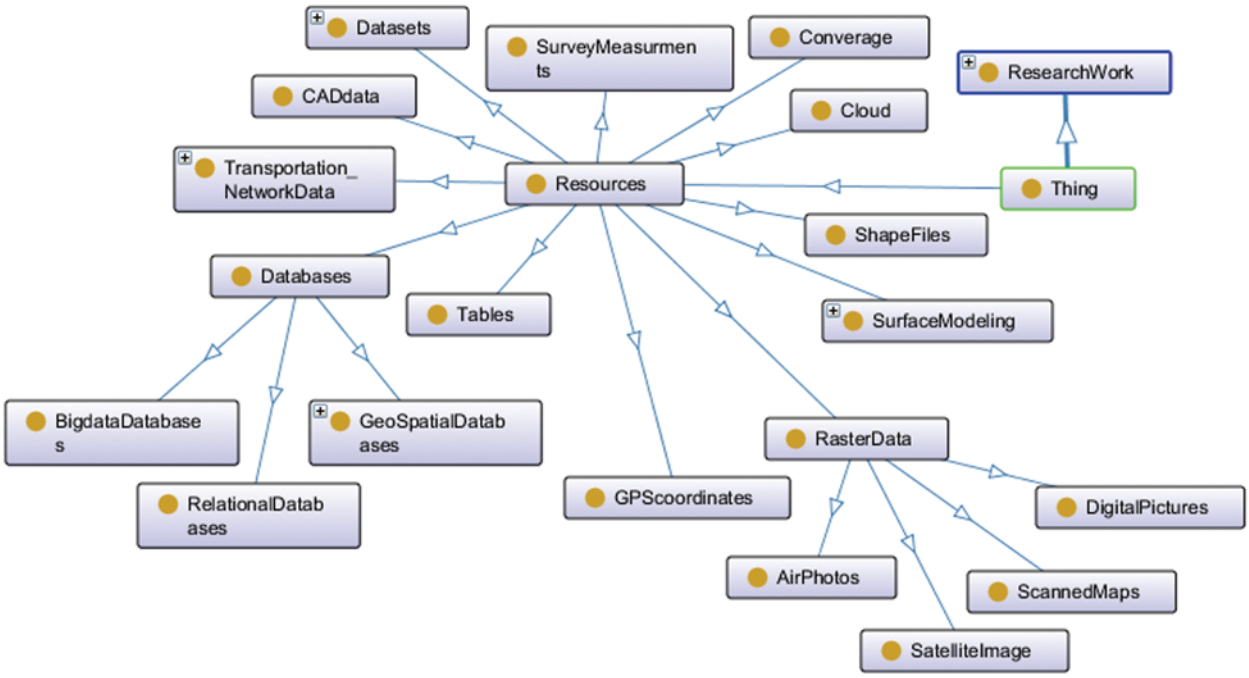

Figure 4: The geospatial ontology used for aggregation of heterogeneous services

The principal purpose of the second phase is to prepare the extracted data and transform it into the RDF format for linking. This consists of five procedures. The first procedure is data preparation which contains some typical transformation activities for preparing the data. This includes such activities as normalizing data, removing duplicates, checking for integrity violations, filtering, sorting, grouping, and dividing data according to format. Additionally, it transforms data to RDF format (structured, semi-structured, or unstructured). The RDF generation, shown in Fig. 7 for structured and semi-structured data, is based on the standardized geospatial ontology shown in Fig. 8. The derived data are then converted to RDF format.

The RDF generation algorithm used for this transformation is as follows:

Trans–Data-to–RDF algorithm

Input: src_1: Geospatial CSV file,
src_2: Geospatial Ontology file
Output: real: *alignmentmeasure*_value // matching value between entities in $src_1$and src_2 files,
RDF data file generation
Variables:
file: CSV_File // read the CSV data
string: RDF_className // class name of the generated RDF data
array: Column_Listing // list for storing all columns name
string: column_name // string store column name in the input CSV file
string: column_data // string store every value of the CSV data

(Continued)

```
object: csvModel // model to hold the RDF data which generated from the Geospatial CSV input file
object: geospatial_ontology // model to hold the Geospatial Ontology file
object: dataPropertys // object from DatatypeProperty class
object: csvIndividual // object for creating individuals
string: property // string to get the data property name from the Colomn_Listing
object: First_ontology // object from JENAOntology
object: Second_ontology // object from JENAOntology
object: alignmentmeasure // for creating the alignment process between First_ontology and Second_ontology
int: counter // initial value equal 0
string: entity1 // hold the data properties name of the First_ontology in each cell
string: entity2 // hold the data properties name of the Second_ontology in each cell
Processing:
Begin:
// First Stage:
    //First Step: Read Geospatial CSV file
    CSV_File ← src1
    //Second Step: Convert data in CSV into RDF format
    RDF class name = src1 name
    // create the data properties from columns name
        For all column_names in src1 {
                column_name ← column value
                DatatypeProperty dataPropertys =
                 csvModel.createDatatypeProperty(column_name);
                Coloumn_Listing.add(column_name); }
        // set every row data as a new individual
            while (! End-of-file(src1))
                { column_data ← column value
                Individual csvIndividual = cvsClass.createIndividual ();
                String property = Coloumn_Listing.get (counter++);
                csvIndividual.addProperty(csvModel.getDatatypeProperty(property),
                 column_data); }
  OntModel geospatial_ontology ← read src2
  // Using "Alignment API" to calculate similarities
JENAOntology First_ontology = new JENAOntologyFactory().newOntology(csvModel, true);

JENAOntology Second_ontology = new JENAOntologyFactory().newOntology(geospatial_ontology, true);
// Aligning data properties between two ontologies
AlignmentProcess alignmentmeasure = new SMOANameAlignment( );
alignmentmeasure.init (First_ontology, Second_ontology); // takes the source and target
ontology to alignment
alignmentmeasure.align (First_ontology.getdataproperties(), Second_ontology.Countryclass.
```

(Continued)

```
52  getdataproperties());
53 For all cells c in alignmentmeasure
54 {
55 alignmentmeasure_value = cell.getStrength( ); // get measre value
56 entity1 = cell.getObject1( ).toString( ); // get data property name of First_ontology
57 entity2 = cell.getObject2( ).toString( ); // get data property name of Second_ontology
58              If (alignmentmeasure_value > 0.5)
59              {
60                    entity1.value ← entity2.value;
61               }
62 }
63 // Second Stage
64 Save First_ontology as RDF format in an XML file
65
66
67
68
   END
```

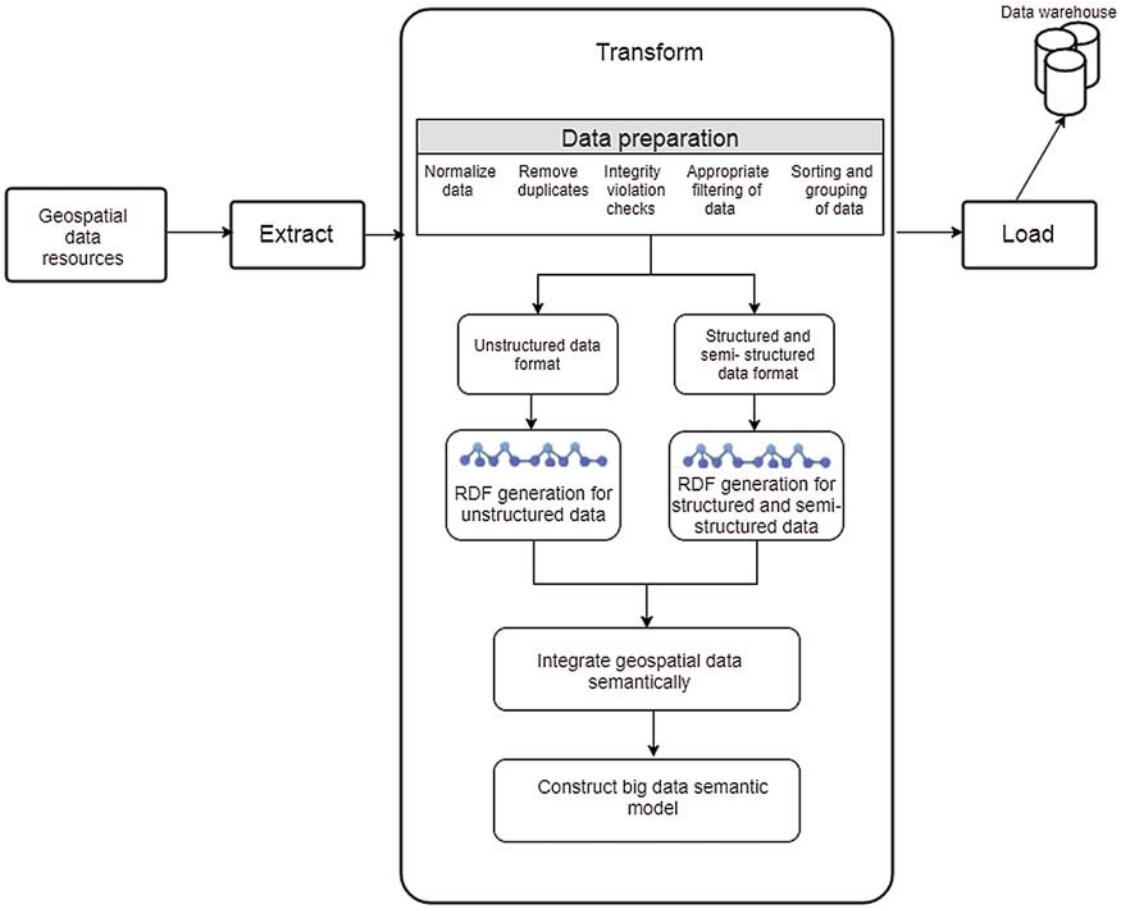

Figure 5: Semantic ETL model

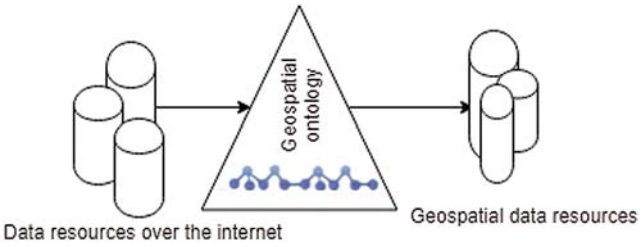

Figure 6: Geospatial data resources aggregation

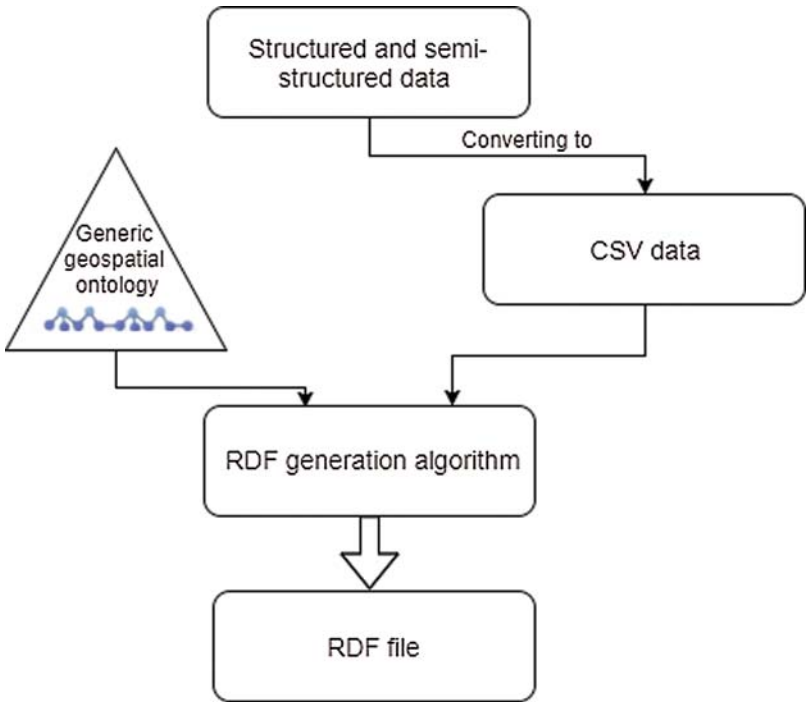

Figure 7: RDF file generation from data files of structured and semi-structured data

Using the API 4.0 [33,34] alignment, this algorithm transforms the CSV data file into an RDF file by the default geospatial ontology used. Thus, structured, and semi-structured data files (such as XML, EXCEL, and JSON) are translated into CSV data files before the RDF generation algorithm is applied. Structuring analysis is used to remove the noisy components and generate metadata information, followed by a data mining operation consisting of two procedures, linguistic and semantic analysis as shown in Fig. 9.

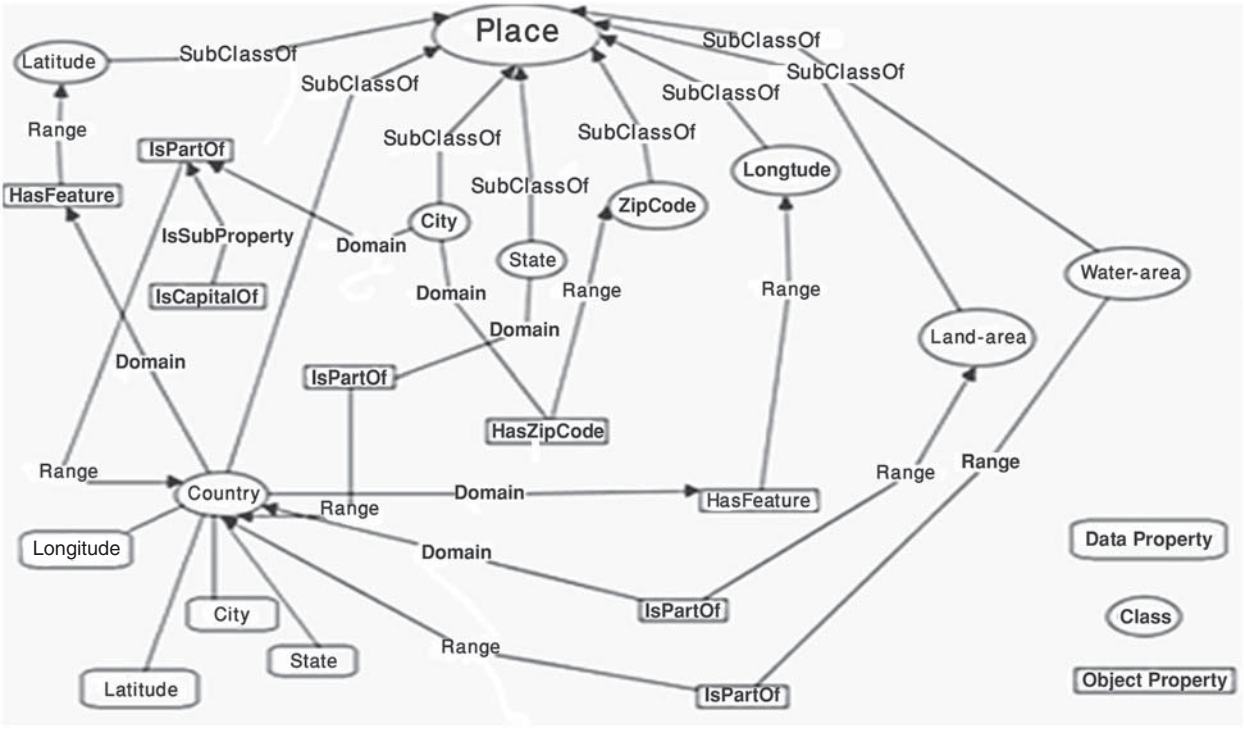

Figure 8: Generic geospatial ontology

The linguistic analysis method involves two steps. Firstly, phrase splitting which includes a speech tagger section, morphological examination, a JAPE transducer, and a root gazetteer. The JAPE transducer implements specific laws centered on regular expressions over the annotated corpus. It is the responsibility of the onto root gazetteer to take domain ontology as an input to construct an annotated corpus with the geospatial entities. The second step is the semantic analysis that is used to catch the hidden relationships between the annotated entities in the textual details. The output of the linguistic analysis is used as the input to the system of semantic analysis, which uses fundamental semantic rules adapted from [34] to extract the relationships from unstructured textual data.

The final procedure is gathering the geospatial data for the semantic model. The data linking algorithm is used in this procedure to link RDF data files semantically before merging them to address the problem of semantic heterogeneity. Next, the linkage and integration algorithm are used to compare and avoid the redundancy of entities before the integration process, accompanied by merging all data from RDF files into a single file. Finally, the semantic model for storing integrated geospatial data semantically using the technique is built-in. The generated semantic model is stored in the data warehouse in the last phase of the semantic ETL model.

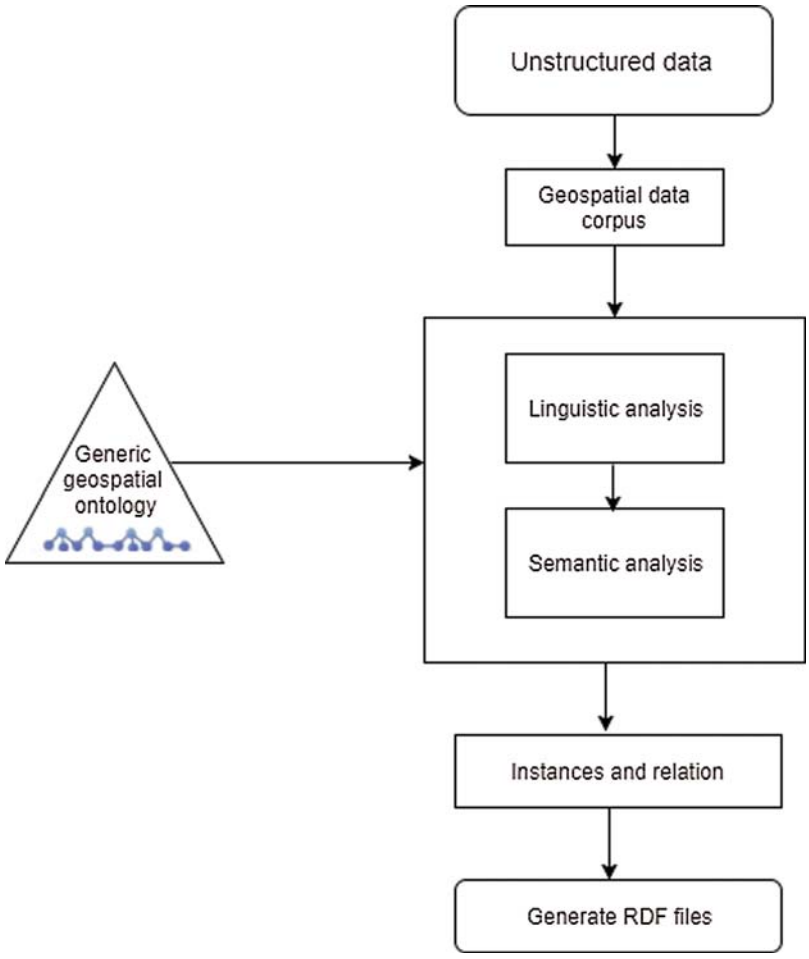

Figure 9: RDF file generation from unstructured data files, adapted from [34]

5 Discussion of Experiment

5.1 Experimental Setup

Specific SW technologies are used to implement the proposed approach, as follows:

1. Uniform Resource Identifier: Defining and finding properties such as the default web pages, offering a baseline to represent the characters used in most languages of the world, and to classify resources [34].
2. RDF: Internet data-sharing model, which defines the metadata of websites and ensures interoperability between applications. This facilitates the data merging of various schemes and allows the mixing, exposure, and sharing of structured and semi-structured data across different applications [35].
3. SPARQL: The RDF query language and protocol used to query, retrieve, and process RDF-format data [36].
4. OWL: An SW language built on top of RDF. Written in XML, it represents things, classes of things, and links between items of knowledge of things [37].

5. Alignment API: Offers abstractions for ontology, alignment, and correspondence network notes, as well as coercive building blocks such as matches, evaluators, renderers, and parsers.
6. XML (eXtensible Markup Format): An extensible format that enables users to construct their document identifiers. Provides the syntax of the material structure inside documents [38].
7. Program Eclipse.
8. Protégé "ontology editor": This is an open-source editor for the construction of ontology domain models and knowledge-based applications [39].

5.2 Results

The semantic heterogeneity problem that arose in data integration was solved using the proposed RDF generation algorithm based on the alignment API and a generic geospatial ontology to describe the similarities between ontology and RDF data properties, as seen in Tabs. 2–5.

Table 2: Matching attributes between ontology and data in source

Ontology attributes	Data attributes	Similarity value
State	State	1
Country	Country	1
City	City	1
Longitude	Longitude	1
Latitude	Latitude	1
Zip_code	Zip_code	1

Table 3: Matching attributes between ontology and data in source

Ontology attributes	Data attributes	Similarity value
State	State	1
State	St	0.09
City	Name	×
Longitude	Lon	0.825
Latitude	Lat	0.84
Zip_code	Zprog_disc	0.05

Table 4: Matching attributes between ontology and data in source

Ontology attributes	Data attributes	Similarity value
State	State	1
City	Name	×
Country	Country	1
Zip_code	Zip_codes	0.98
Longitude	Longitude	1
Latitude	Latitude	1
Land_area	Land_area	1
Water_area	Water_area	1

Table 5: Matching attributes between ontology and data in source

Ontology attributes	Data attributes	Similarity value
State	Statefp	0.95
Country	Countryfp	0.96
City	Name	×
Land_area	Aland	0.74
Water_area	Awater	0.78

The CSV data are used and translated into the RDF data file using the proposed RDF generation algorithm. If the used data is unstructured, the SPARQL query extracts the attributes from its RDF file, and then the proposed linking algorithm is applied to match the attributes and is translated into RDF format to validate the algorithm. Fig. 10 shows a case to illustrate this procedure.

```
the Ontology atrributes is:zip_code, latitude, longitude, city, state, country,
land_area, water_area,

 the RDF data attributes is:zprog_disc, lat, lon, st, state, name,

the matched data :
lat,lon,st,state,name,

 new data in source2 :
zprog_disc,
```

Figure 10: Example of matched attributes

The next stage is to align the attributes extracted in from both the constructed ontology and the data, as in Fig. 10. Since the attributes "name, city," "lon, longitude," and "lat, latitude" correspond to the same details, they have matched attributes. This suggests that the issues of textual and structural variability have been solved and that the data services are combined and semantically processed.

Table 6: Comparison between existing semantic models and the proposed model

Research	Aggregation	Integration	Representation	Individual redundancy	Column redundancy	Linking
Souza et al. (2006)	×	×	×	×	✓ (Manually on the ontology)	×
Du et al. (2011)	×	×	×	✓	×	×
Zhang et al. (2013)	×	✓	×	✓	×	✓
Xiong et al. (2014)	✓	×	×	×	×	×
Gollapudi & Sunila (2014)	✓	×	×	×	×	×
Kang et al. [1]	×	×	✓	×	×	×
Bansal et al. (2014)	×	✓	×	×	×	✓
Yunianta et al. (2017)	×	✓	×	×	✓ (Manual)	✓
Proposed model	✓	✓	✓	×	✓ (Trans–Data-to–RDF) algorithm	✓

The emphasis for large-scale data studies is primarily on quantity, speed, and variety. SW technology was not introduced for such data. The pace and volume of SW developments remain major challenges. The semantic heterogeneity issue created by the variety of big data is overcome in the proposed model. Machine performance shows which components of the input systems have been implemented effectively. Tab. 6 lists the differences between the existing models and the proposed model.

Confidence in the relationship between characteristics of the alignment supplier is enhanced by the magnitude of the greater interest (measurement meaning: Float between 0.0 and 1.0). Various communications systems for the API are described in [39].

6 Conclusion

This study presents a new ETL semantic model that allows for combining, associating, incorporating, and characterizing geospatial data using semantic technology from numerous geospatial resources on the internet. Geospatial data services are first aggregated semantically, and then the three steps of the ETL are combined, viewed, and processed as LD. Besides, we addressed problems of systemic and semantic heterogeneity before the integration cycle. SW technology solves the big data variety problem, but not the quantity problem.

Funding Statement: The authors received no specific funding for this study.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] L. Kang, Li Yi and L. Dong, "Research on construction methods of big data semantic model," in *Proc. of the World Congress on Engineering*, London, UK, WCE, Vol. I, 2014.

[2] B. Srividya, "Towards a semantic extract-transform-load (ETL) framework for big data integration Big Data (BigData Congress)," in *IEEE Int. Congress on IEEE*, United States, pp. 522–529, 2014.

[3] B. Christian, B. Peter, B. Michael and E. Orri, "The meaningful use of big data: Four perspectives-four challenges," *ACM SIGMOD Record*, vol. 40, no. 4, pp. 56–60, 2012.

[4] E. Zineb, M. José-Norberto, V. Alejandro and Z. Esteban, "BPMN-based conceptual modeling of ETL processes," in *14th Int. Conf. on Data Warehousing and Knowledge Discovery, DaWaK 2012, Proceedings*, September 3–6, Vienna, Austria, pp. 1–14, 2012.

[5] B. Arputhamary and L. Arockiam, "A review on big data integration," *International Journal of Computers and Applications*, vol. 5, pp. 21–26, 2014.

[6] V. Gour, S. S. Sarangdevot, G. S. Tanwar and A. Sharma, "Improve performance of extract, transform and load (ETL) in data warehouse," *International Journal on Computer Science and Engineering*, vol. 2, no. 3, pp. 786–789, 2010.

[7] B. Elisa, "Big data—opportunities and challenges," in *IEEE 37th Annual Computer Software and Applications Conf.*, Kyoto, Japan, pp. 479–480, 2013.

[8] T. Krishnaprasad and S. Amit, "Semantics-empowered approaches to big data processing for physical-cyber-social applications," in *Semantics for Big Data: Papers from the AAAI Symp.—AAAI Technical Report FS-13-04*, Arlington, Virginia, USA, pp. 68–75, 2013.

[9] A. Mahmoud, T. Hamza and M. Z. Rashad, "Prediction of chemical toxicity for drug design using AIRS algorithms and hybrid classifiers," *International Journal of Applied Mathematics & Information Sciences*, vol. 14, no. 2, Article 38, 2020.

[10] W. Hongyan and Y. Atsuko, "Semantic web technologies for the big data in life sciences," *BioScience Trends*, vol. 8, no. 4, pp. 192–201, 2014.

[11] Z. Ahmed and G. Detlef, "Web to semantic web & role of ontology," arXiv preprint arXiv, pp. 1008–1031, 2010.
[12] J. Vishal and S. Mayank, "Ontology-based information retrieval in semantic web: A survey," *International Journal of Information Technology and Computer Science*, vol. 5, no. 10, pp. 62–69, 2013.
[13] D. M. Beniamino, E. Antonio, N. Stefania and M. S. Augusto, "A semantic model for business process patterns to support cloud deployment," *Computer Science-Research and Development*, vol. 32, no. 3, pp. 257–267, 2017.
[14] N. A. Awad and A. Mahmoud, "Improving reconstructed image quality via hybrid compression techniques," *Computers, Materials & Continua*, vol. 66, no. 3, pp. 3151–3160, 2021.
[15] B. Srividya and K. Sebastian, "Integrating big data: A semantic extract-transform-load framework," *Computer*, vol. 48, no. 3, pp. 42–50, 2014.
[16] B. Sonia, G. Francesco, O. Mirko, S. Claudio and V. Maurizio, "A semantic approach to ETL technologies," *Data & Knowledge Engineering*, vol. 70, no. 8, pp. 717–731, 2011.
[17] J. Lihong, C. Hongming and X. Boyi, "A domain ontology approach in the ETL process of data warehousing, e-Business Engineering (ICEBE)," in *2010 IEEE 7th Int. Conf.*, Shanghai, pp. 30–35, 2010.
[18] A. Mahmoud, T. Hamza and M. Z. Rashad, "An approach for extracting chemical data from molecular representations," *International Journal of Advanced Computer Research*, vol. 10, no. 46, pp. 27–33, 2020.
[19] D. Mei, H. Yan-ling, D. Ming-Hua and Z. C. Zhang, "Application of ontology-based automatic ETL in marine data integration, " in *IEEE Electrical & Electronics Engineering, Symp. on*. Malaysia: Kuala Lumpur, pp. 11–13, 2012.
[20] C. Isabel, G. Venkat and M. S. Iman, "Semantic extraction of geographic data from web tables for big data integration," in *Proc. of the 7th Workshop on Geographic Information Retrieval*, Orlando, FL, USA: ACM, pp. 19–26, 2013.
[21] B. Boury and C. Anne, "Managing semantic big data for intelligence," in *Proc. of the Int. Conf. Semantic Technologies for the Intelligence, Defence and Security*, November 12–15, Fairfax, VA, pp. 4–47, 2013.
[22] A. Saradha, "Semantic integration of heterogeneous web data for tourism domain using ontology-based resource description language," *Journal of Computer Applications*, vol. 3, no. 3, pp. 1–13, 2010.
[23] H. Jadhao, D. Aghav and J. Vegiraju, "Semantic tool for analyzing unstructured data," *International Journal of Scientific & Engineering Research*, vol. 3, no. 8, pp. 1–7, 2012.
[24] Z. Ying, C. Yao-Yi, S. Pedro and K. Craig, "A semantic approach to retrieving, linking, and integrating heterogeneous geospatial data," in *Joint Proc. of the Workshop on AI Problems and Approaches for Intelligent Environments and Workshop on Semantic Cities*, Beijing, China, ACM, pp. 31–37, 2013.
[25] X. Jing, L. Yuntong and L. Wei, "Ontology-based integration and sharing of big data educational resources," in *IEEE 11th Web Information System and Application Conf.*, Tianjin, China, pp. 245–248, 2014.
[26] S. Gollapudi, "Aggregating financial services data without assumptions: A semantic data reference architecture," in *IEEE Int. Conf. on Semantic Computing*, Anaheim, CA, USA, pp. 312–315, 2015.
[27] S. Dastgheib, A. Mesbah and K. Kochut, "mOntage: Building domain ontologies from linked open data," in *Proc. 7th IEEE Int. Conf. Semantic Computing (ICSC 13)*, CA, USA, pp. 70–77, 2013.
[28] A. Calì, D. Calvanese, G. D. Giacomo and M. Lenzerini, "Data integration under integrity constraints," *Information Systems*, vol. 29, no. 2, pp. 147–163, 2004.
[29] S. Bergamaschi, F. Guerra, M. Orsini, C. Sartori and M. Vincini, "A semantic approach to ETL technologies," *Data & Knowledge Engineering*, vol. 70, no. 8, pp. 171–731, 2011.
[30] A. Halevy, A. Rajaraman and J. Ordille, "Data integration: The teenage years," in *Proc. 32nd Int. Conf. Very Large Databases*, Seoul, Korea, pp. 9–16, 2006.
[31] E. Kandogan, M. Roth, C. Kieliszewski, F. Özcan, B. Schloss *et al.,* "Data for all: A systems approach to accelerate the path from data to insight," in *Proc. IEEE 2nd Int. Congress Big Data (BigData 13)*, Santa Clara, USA, pp. 427–428, 2013.
[32] R. P. DebNathab, K. Hos, T. B. Pedersena and O. Romerob, "SETL: A programmable semantic extract-transform-load framework for semantic data warehouses," *Information Systems*, vol. 68, pp. 17–43, 2017.

[33] M. H. Seddiqui, S. Das, I. Ahmed, R. P. D. Nath and M. Aono, "Augmentation of ontology instance matching by automatic weight generation," in *Information and Communication Technologies (WICT), World Congress on IEEE*, Mumbai, India, pp. 1390–1395, 2011.

[34] A. Simitsis, P. Vassiliadis, M. Terrovitis and S. Skiadopoulos, "Graph-based modeling of ETL activities with multi-level transformations and updates," in *Proc. of DaWaK*, Copenhagen, Denmark, pp. 43–52, 2005.

[35] J. Euzenat, "An API for ontology alignment," in *International Semantic Web Conf.*, Berlin, Heidelberg: Springer, 698–712, 2004.

[36] J. Trujillo and S. Lujàn-Mora, "A UML based approach for modeling ETL processes in data warehouses," in *Proc. of ER2003*, Chicago, IL, USA, pp. 307–320, 2003.

[37] C. Thomsen and T. Bach Pedersen, "Pygrametl: A powerful programming framework for extract-transform-load programmers," in *Proceedings of the ACM 12th International Workshop on Data Warehousing and OLAP*, Aalborg, Denmark, ACM, pp. 49–56, 2009.

[38] V. Nebot and R. Berlanga, "Building data warehouses with semantic web data," *Decision Support Systems*, vol. 52, no. 4, pp. 853–868, 2012.

[39] M. Thenmozhi and K. Vivekanandan, "An ontological approach to handle multidimensional schema evolution for data warehouse," *International Journal of Database Management Systems*, vol. 6, no. 4, pp. 33–52, 2014.