EduBD: A Machine Understandable Approach to Integrate Information of Educational Institutions of Bangladesh

Shima Chakraborty[1], Md. Hasan Hafizur Rahman[2], Md. Hanif Seddiqui[1], Sajal Chandra Debnath[3]

[1] Dept. of Computer Science & Engineering, University of Chittagong, Chittagong - 4331, Bangladesh
[2]Department of Computer Science & Engineering, Comilla University, Comilla, Bangladesh
[3]Junior GIS & RS Analyst

ABSTRACT

Web contents related to educational institutions as well as their geographic data of a country is an emerging field of data sharing and consolidating with suitable data repositories to extract useful information to improve and maintain the area-based quality of education. The information of these institutions is available in the unstructured or semi-structured format from heterogeneous sources on the web that originate challenges to develop a smarter method in finding specific information using traditional search engines. The integration of these heterogeneous data along with geographic information is a formidable task to explore implicit knowledge of educational institutions that assist people to improve these systems and it is still getting researchers' attention. In this regard, our research demonstrates the feasibility of semantic web technologies for converting and integrating these unstructured or semi-structured information by introducing machine understandable description of resources and instances by consisting of their concepts and relations, called a repository name EduBD. This yields a magnificent knowledge graph with .75 million nodes, i.e. RDF triples to facilitate reasoning meaningful information by a graph based query language SPARQL. Furthermore, we utilized our knowledge-base EduBD with one more accessible spatial data source Geo-Bangladesh using Physical-Cyber-Social computing approach to examine spatial contents and social information of each institution and therefore achieved the semantic interoperability of our Linked Open Data (LOD) application by eliminating natural language polisemy problems.

KEYWORDS

Semantic Web, Ontology, Linked Open Data, Physical-Cyber-Social computing & Semantic Information Integration

1. INTRODUCTION

The heterogeneous web repositories of educational institutions available in unstructured and semi-structured format along with their spatial information are inadequate to develop a state-of-the-art method in finding accurate information on the web. These repositories consist of geographic data of educational institutions such as *address of these institutions* as well as their academic and administrative information. These useable contents originate on the World Wide Web *(WWW)* using traditional technologies such as, Hypertext Markup Language, Dynamic HTML and so on. Although these technologies produce more lucrative web pages related to educational institutions, they are incompatible to define the semantics in metadata to enable machines to extract the meaning of these representations. These representations use predefined

DOI : 10.5121/ijwest.2016.7101

tags to affirm the style of a sequence of characters rather than defining ubiquitous meaning in a document on the web. Let us take into account the HTML description for *"Comilla university located in Comilla district"* without their semantics instead of a sequence of characters as follows:

< p > Comilla Zilla School located in Comilla district < p >

Although human being, those who are familiar with the terms *School*, *Located*, *District* and *Comilla* can extract the meaning from this representation by implementing their intuitive learning knowledge, whereas the machines merely consider this representation as a sequence of characters.

Furthermore, eXtensible Markup Language *(XML)* is a semi-structured data representation technique having the facilities of defining user defined tags to represent domain specific information in the hierarchical fashion. The lack of uniqueness in these user defined tags of XML are hardly understandable to machines, however, these representations are easily human interpretable and understandable [1][2]. On the other hand, Database Management Systems *(DBMSs)* are used as back-end along with these conventional technologies to represent information of institutes as web applications by performing various structured query operations such as insert, delete, update, search and so on [3]. In this case, web application developers design schema structures those are restricted to define attributes and their respective constraints such as primary keys, foreign keys and null rules to develop meaningful associations to create a chain of data on the web. As a result, a closed world assumption is applied to relational database model that loss big semantic in the process of data modeling [4].

Therefore, spatially dispersed institutes related information are more important to policy planners and executing authorities to address major problems of the education sector as well as these information also useful to mass people to evaluate these institutions in the area of interest. In this regard, numerous Geographic Information Systems *(GISs)* are used as advanced tools to reform education system by analyzing and synthesizing spatial data, visualizing these data on the map and interpreting these data to make Spatial Decision Support Systems (SDSS) [5][6][7]. These *GIS* systems use heterogeneous data sources of various formats mainly attained from diversified proprietary software [8][9]. These heterogeneous web contents from diverse GISs exhibit challenges in the data integration operation. For this reason the meaningful description of data on the web are emerging to deal with the development of interoperable geographical applications and software [10][11], geographical information retrieval [12] and automated spatial reasoning [13]. In this regard, a semantic model is used to create metadata by defining common vocabularies to integrate data across various domain and addresses semantic interoperability issues by identifying categories, concepts, relations and rules to develop the next generation web, *Semantic web* (*SW*), coined by Berners-Lee et al. [14][15][116]. In addition, the exponential growth of machine interpretable metadata repositories as background knowledge of various domains on the web, the *Physical-Cyber-Social (PCS)* computing [17] approach on the education sector of a country is a motivated, however sophisticated problem with considerable application areas such as *originate a geographical network of educational institutions, perform spatial analysis to compare the quality of education, exploring compatible implicit knowledge on the web, geographically establish e-education services, develop spatial decision support systems and so on* that access various data sources from physical and social world related to education systems. For these scope of applications, the major concerns of our research is to convert heterogeneous geographic information of educational institutions of Bangladesh into machine understandable representation using rich semantic *RDF* framework by following an ontology model and we called this knowledge-base ***EduBD*** that attains resource sharing and interoperability issues effectively. Moreover, we perform a large number of experiments using graph based query language *SPARQL* [18] to retrieve *geo-coordinates* of each institutions from our generic semantic data repository *Geo-Bangladesh*, a spatial knowledge-base related to the administrative structure of Bangladesh [19] in a faster and efficient way by using various communication technologies [20][21][22]. Then these *geo-coordinates* are used to measure distance of educational institutions from physical sensors as well as to extract location of neighboring institutions including their relevant information on the map to increase search ability by eliminating natural language polisemy problems that spectacles the performance of our application.

The rest of the paper is organized as follows. **Section II** describes some basic terminologies to understand the subsequent contents of this paper. Our present structure of educational institutions is articulated in **Section III** while **Section IV** focuses on the details data preparation procedure and **Section V** represents the approach of our research. The procedure of building a semantic knowledge repository is given in the first subsection while the following two steps demonstrate the process of producing our knowledge-base, ***EduBD*** as LOD along with the integration of data across various domains using *PCS* computing approach. **Section VI** includes the experiments and evaluation by performing a lot of experiments using *SPARQL*. Concluded thought of our work as well as some future directions are described in **Section VII**.

2. GENERAL TERMINOLOGIES

This segment of our article popularizes a few basic terminologies including their respective notion to readers those are used all over this paper. It includes the theme of ontology, RDF Model and HTTP-URIs, Linked Open Data and Interoperability, Physical-Cyber-Social computing approach, Geographic Information System, Global Positioning System and Global System for Mobile Communication to comprehend the essence of our research.

2.1. Ontology

Ontology is defined as "an explicit, formal specification of a shared conceptualization of a domain of interest" [23][24] serve as the first brick of the semantic web [25][26]. The core ontology, *S*, structure is defined in the *Eq. 1* as five tuples,

$$S = (C, \leq_C, R, \sigma, \leq_R) \tag{1}$$

In *Eq. 1,* C and R demonstrate *concepts* and *relations* of a domain those are disjoint one each others. The notation $\leq_C$ serve as partial order on C, called concept hierarchy or *taxonomy* and the notation $\leq_R$ performs on *R*, called *relation hierarchy* [27].

2.2. RDF Model and HTTP-URI

Resource Description Framework *(RDF)* represents real world information in the graph pattern as *subject-predicate-object* form, known as *triples* [28]. In this model resource, *items of interest of a domain* are described using *HTTP- URIs* to manage globally unique name to access properties and make associations with other resources by escaping duplication of data on the web without central execution [29]. The Fig. 1 illustrates the structure of a triple for the knowledge piece *"Comilla Zilla School located in Comilla"* to comprehend the essence of semantic knowledge representation.

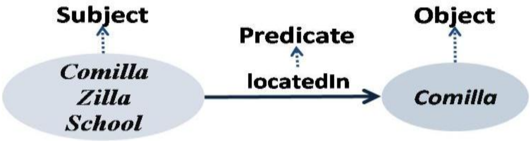


Figure 1: RDF Triple for *"Comilla Zilla School located in Comilla"*

2.3. LOD and Interoperability

The structure of a knowledge base is characterized as

$$KB = (C, R, I, i_C, i_R)$$

including two disjoint sets *C* and *R* those are interpreted in our previous subsection. The notation *I* is a set of elements called *instances* and two functions i_C and i_R are *concept instantiation* and *relation instantiation* respectively.

In our application we use the structure of ontology and ontology knowledge-base to express our domain data as Linked Data and distribute these knowledge pieces as Linked Open Data *(LOD)* to establish a universal data source on the web [29][30]. These knowledge-base develop the web as interoperable because interoperability performs the operation of information swapping between two or more parties by resolving syntactic and semantic issues properly [31][32].

2.4. PCS Computing

The Physical-Cyber-Social *(PCS)* computing paradigm incorporates information from cyber space data sources and social perceptions by utilizing computing and communication capabilities to extract intuitive knowledge that neither traditional computing nor human intelligence can explore [33][17]. This perceptive knowledge is analyzed by expanding the web with the capabilities of sensing, processing and self adapting to perform interactions between physical and virtual world through exclusive addressing scheme [34]. In this connection, a close look of *PCS* computing paradigm is illustrated in the Figure 2 to comprehend the significance of our research.

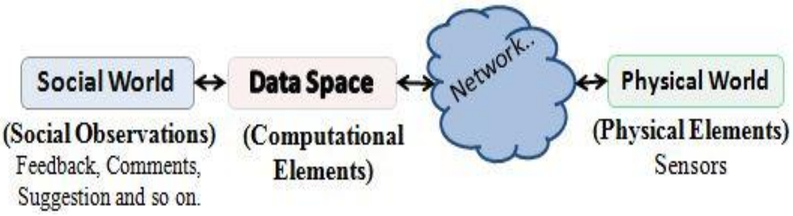


Figure 2: The Physical-Cyber-Social *(PCS)* computing approach

2.5. GIS, GP S and GSM

The Geographic Information Systems *(GISs)* process *geo-referenced* data by combining hardware and software for processing, analyzing and displaying in the form of maps, reports and charts. These data are spatial, *points, lines or areas known as locational data* and attribute *(non locational)* data, *features of points, lines or areas* [35].

Furthermore, the Global Positioning System *(GPS)* is a satellite-based uninterrupted, 3-D positioning and navigation system whereas Global System for Mobile Communication *(GSM)* use three GSM base-stations to locate a point in the form of *{latitude, longitude}* on the earth [36] [37].

3. PRESENT STRUCTURE OF OUR PROBLEM DOMAIN

There are a large number of educational institutions in Bangladesh without collaboration of one institute with others. As a result it is hard to integrate information for making decisions to improve the quality of these educational institutions. These educational institutions include schools, colleges, madrasahs, universities, research centers, libraries; learning organizations and so on are spatially dispersed in the range of latitude: *$20^0$34N to $26^0$38N* and longitude: *$88^0$01E to $92^0$41E* (*147570* square kilometers) to produce educated and skilled manpower. This area of Bangladesh is coordinated by a standard administrative hierarchy that originates a top-down approach from division to villages to perform smooth administrative functions. In this framework, the number of divisions, districts, upazilas and administrative thanas, pourashava or municipalities, wards, union councils, mouza or mohollas and villages are *7, 64, 500,509, 265,* *2407, 4451, 67100 and 87968* respectively [19][38]. The logical relationship between different levels of entities in this framework is as follows:

country ⟵ *division* ⟵ *district* ⟵ *upazilas / thana*
⟵ *pourashava* ⟵ *ward / union* ⟵ *mouza / mohollas* ⟵ *village*

This administrative area consists of educational institutions where national level authorities formulate various policies, service delivery mechanisms, allocation and utilization of resources whereas lower level institutions execute these policies effectively. In Bangladesh the education systems are General Education System *(GES)*, Madrasah Education System *(MES)*, Technical-Vocational Education System *(TVES)* and Professional Education System *(PES)*. These education stages further divided into five levels are primary, junior, secondary, higher secondary and tertiary level respectively. These levels consist of instances of educational institutions those are located at the lower level of the administrative framework. The Figure 3 depicts the interrelation between educational institutions and the administrative framework of Bangladesh.

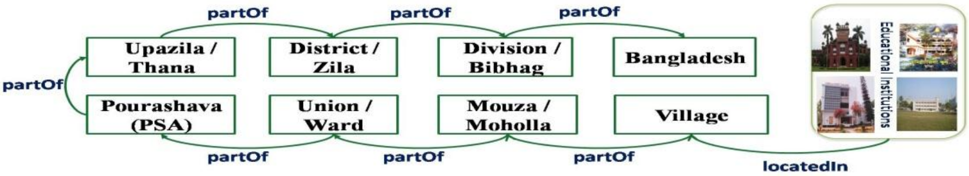


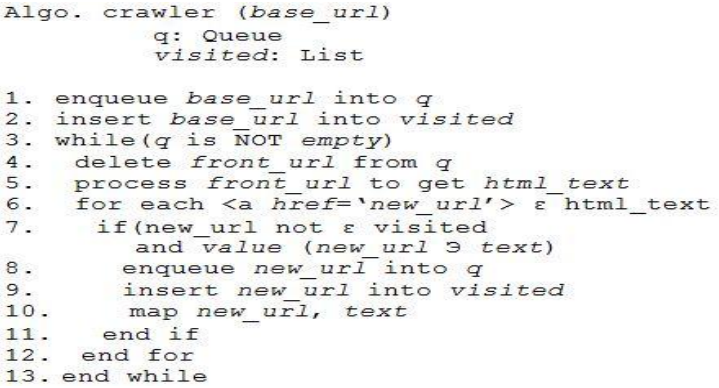
Figure 3: The logical relationship between educational institutions and administrative framework of Bangladesh

4. DATASETS FOR EDUCATIONAL INSTITUTIONS

The educational institute related data of Bangladesh is not usable from a single data provider's organization. It is a challenging task to accumulate and validate useful data on the web from accessible data sources. We collect and verify data for our application in both manual and automatic process. In this regard, we introduce a crawler to enrich our data collection along with the operation of cleaning HTML texts and parse these contents into suitable format those are illustrated in the following sub-sections.

4.1. Crawling URLs

Our crawler extracts information based on Uniform Resource Locator *(URL)*. The algorithm of the crawler is demonstrated in Figure 4 that uses Breadth-First-Search *(BFS)* to find necessary information efficiently.



```
Algo. crawler (base_url)
          q: Queue
          visited: List

enqueue base_url into q
insert base_url into visited
while(q is NOT empty)
  delete front_url from q
  process front_url to get html_text
  for each <a href='new_url'> ε html_text
    if(new_url not ε visited
       and value (new_url ∋ text)
      enqueue new_url into q
      insert new_url into visited
       map new_url, text
    end if
  end for
end while
```

Figure 4: Pseudo code of how our crawler works.

In the algorithm crawler of Figure 4, the queue, *q* and a list, *visited* are initialized with base url are demonstrated in line 1 and 2. Removing each url available from the front of a queue (as line 4), the text inside the url is processed by JSoup [40], demonstrated at line 5, to retrieve further new urls. For each (as line 6) retrieved new url (as line 7), insert the url at the rare of the queue, *q* and insert it into the list, *visited* which is demonstrated at line 8 and 9 respectively. The process is continued until the queue, *q* is empty (as line 3).

4.2. Cleaning HTML and Parse Contents

Traditional web pages on the web consist of various uninformative noisy blocks such as navigation panels, privacy notices, advertisements, copyright and so on. The information contained in these noisy blocks can deteriorate the efficiency of information extraction. Therefore, we remove these noisy blocks to retrieve a clean text of educational institutions and public feedback comments effectively. After crawling and cleaning of HTML texts, find suitable information blocks, we parse these blocks to identify name and address of educational institutes, academic and infrastructure facilities, feedback about these institutions from users using JSoup and convert these information into XML successfully. The feedback consists of *from*, *subject*, *body*, *date*, *time* and so on. In some cases of this conversion i.e., starting tags without closing tags, we rectify these errors manually. The process of data extraction and unification is exhibits in the Figure 5 as a block diagram.

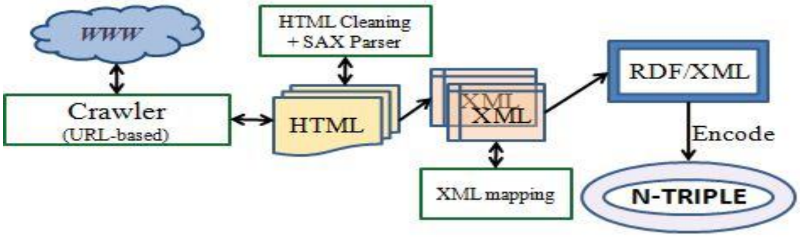
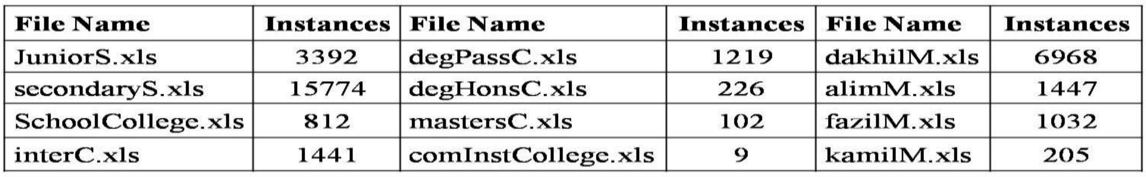


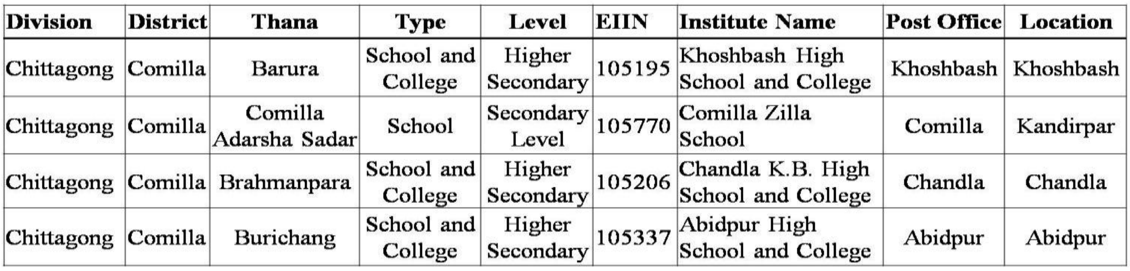
Figure 5: The block diagram for information extraction and integration

In addition, two ministries of Bangladesh *(MoE and MoPME)* are responsible to operate education services except medical education all over the country. We collect feasible information from these ministries along with Bangladesh Bureau of Educational Information and Statistics *(BANBEIS)* that operates under the *MoE*. The list of data files including their number of instances is portrayed in the Figure 6, while the arrangement of data in these files is illustrated in the Fig. 7.



File Name	Instances	File Name	Instances	File Name	Instances
JuniorS.xls	3392	degPassC.xls	1219	dakhilM.xls	6968
secondaryS.xls	15774	degHonsC.xls	226	alimM.xls	1447
SchoolCollege.xls	812	mastersC.xls	102	fazilM.xls	1032
interC.xls	1441	comInstCollege.xls	9	kamilM.xls	205

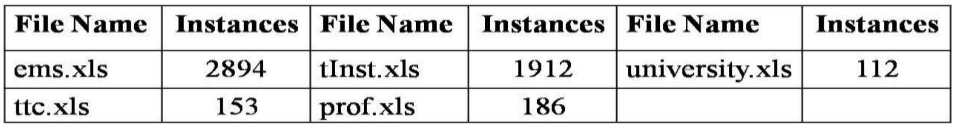
Figure 6: Data files collected from *MoE*



Division	District	Thana	Type	Level	EIIN	Institute Name	Post Office	Location
Chittagong	Comilla	Barura	School and College	Higher Secondary	105195	Khoshbash High School and College	Khoshbash	Khoshbash
Chittagong	Comilla	Comilla Adarsha Sadar	School	Secondary Level	105770	Comilla Zilla School	Comilla	Kandirpar
Chittagong	Comilla	Brahmanpara	School and College	Higher Secondary	105206	Chandla K.B. High School and College	Chandla	Chandla
Chittagong	Comilla	Burichang	School and College	Higher Secondary	105337	Abidpur High School and College	Abidpur	Abidpur

Figure 7: The snippet of the structure of records in the files

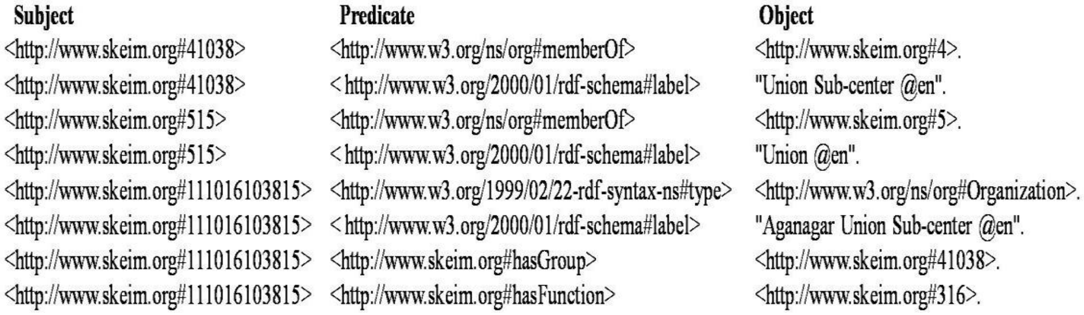
Although these data files consist of 32,627 instances of educational institutions of Bangladesh, a large number of available higher educational institutions, English Medium Schools, Teachers Training Colleges, Technical Institutes, professional institutions, Research Institutes and so on are omitted in these collected files. There are 5257 ignored instances are stored in individual files to produce as machine understandable format from *BANBEIS* and University Grant Commission *(UGC)* respectively. The list of files for these missing data is depicted in the Figure 8.



File Name	Instances	File Name	Instances	File Name	Instances
ems.xls	2894	tInst.xls	1912	university.xls	112
ttc.xls	153	prof.xls	186		

Figure 8: File lists of omitted educational institutions of Bangladesh

Moreover, the information related to medical education is accessible from Directorate of General Health Services *(DGHS)*, Ministry of Health and Family Welfare *(MoHFW)*, Directorate of Nursing Services (*DNS)* as well as from *HCN BD, a semantic database of health care network of Bangladesh* [39]. The knowledge structure of *HCN BD* is illustrated in the Figure 9.



Subject	Predicate	Object
<http://www.skeim.org#41038>	<http://www.w3.org/ns/org#memberOf>	<http://www.skeim.org#4>.
<http://www.skeim.org#41038>	< http://www.w3.org/2000/01/rdf-schema#label>	"Union Sub-center @en".
<http://www.skeim.org#515>	<http://www.w3.org/ns/org#memberOf>	<http://www.skeim.org#5>.
<http://www.skeim.org#515>	< http://www.w3.org/2000/01/rdf-schema#label>	"Union @en".
<http://www.skeim.org#111016103815>	<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>	<http://www.w3.org/ns/org#Organization>.
<http://www.skeim.org#111016103815>	< http://www.w3.org/2000/01/rdf-schema#label>	"Aganagar Union Sub-center @en".
<http://www.skeim.org#111016103815>	<http://www.skeim.org#hasGroup>	<http://www.skeim.org#41038>.
<http://www.skeim.org#111016103815>	<http://www.skeim.org#hasFunction>	<http://www.skeim.org#316>.

Figure 9: The snippet of *HCN-BD* knowledge-base

In spite of the large number of data in our collection, these files also missing the information of non-government educational institutions, vocational education, business management education, qawmi madrasha education, *another category madrasah education in Bangladesh* and so on. As a result, we aggregate compatible missing information after validating from a number of accessible heterogeneous sources by introducing a crawler, HTML cleaning procedure along with parse information to extract relevant information from the web.

5. OUR APPROACH

The approach of our research on the semantic network of educational institutions decomposes into four stages to address the comprehensive knowledge that neither human intelligence nor present computing systems can answer. The first stage of our research develops an ontology model and populates this model by instances of educational institutions. We called our knowledge-base *EduBD*, *a semantic knowledge repository of educational institutions of Bangladesh*. The chronological step of our approach consolidates missing information while the successive step demonstrates how to produce our knowledge-base as *LOD* application to resolve interoperability issues effectively. Finally, the successive step describes the semantic information integration using *PCS* computing approach to locate the institutions on the map including their interoperable linked data. The overview of our approach is portrayed in the Fig. 10 to comprehend the essence of our system.

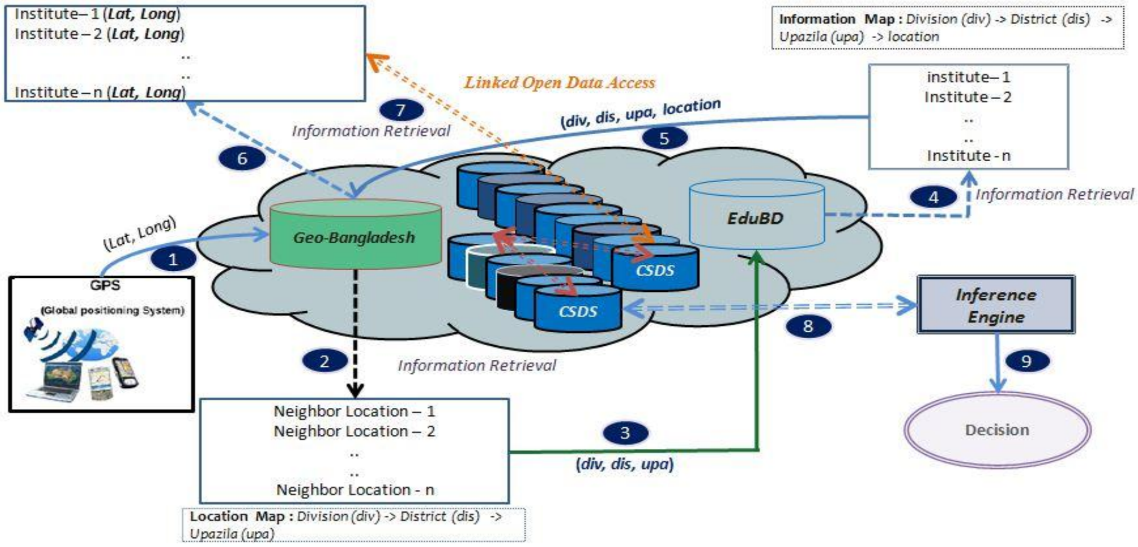
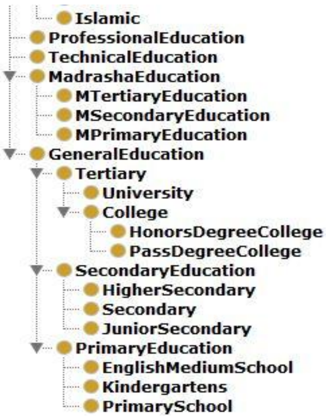


Figure 10: Overview of our System

5.1. Building an Ontology-Based Knowledge Repository

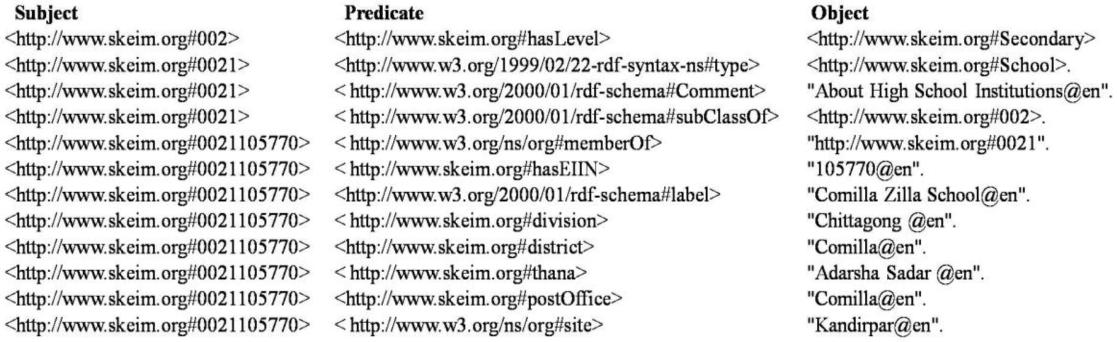
Ontology defines common vocabularies to share information of educational institutions with other compatible data sources on the web to address semantic interoperability issues. It is easier to publish data in RDF using vocabularies from available ontology, however in the case of unavailability of vocabularies, we can propose for suitable one to describe data on the web. The basic steps to develop an ontology are defining classes and arranging these classes in a taxonomic *(sub-class super-class)* hierarchy, defining relationship with other classes called *slots* and allowed values for these slots to develop a semantic data source [41]. We consider classes, sub-classes, relations, object property, data property, general axioms to develop our ontology using ontology environment Protégé for educational institutions to reuse these knowledge and called this ontology *ontoEduBD*. In this ontology, we encoded our domain data by following the World Wide Web Consortium *(W3C)* standards: a core ontology for organizational structure *(org, www.w3.org/ns/org#)* and Dublin Core *(dc, dublincore.org/documents/dces/)* to create interoperable data on the web. We examine a number of available vocabularies such as *type*, *label* and *site* to represent a resource type, feature names, objects location respectively of real world entities in the virtual world. The fragment of taxonomy of our ontology is portrayed in the Figure 11.



  - Islamic
- ProfessionalEducation
- TechnicalEducation
- MadrashaEducation
  - MTertiaryEducation
  - MSecondaryEducation
  - MPrimaryEducation
- GeneralEducation
  - Tertiary
    - University
    - College
      - HonorsDegreeCollege
      - PassDegreeCollege
  - SecondaryEducation
    - HigherSecondary
    - Secondary
    - JuniorSecondary
  - PrimaryEducation
    - EnglishMediumSchool
    - Kindergartens
    - PrimarySchool

Figure 11: The fragment of ontology hierarchy for ontoEduBD

Moreover, we populate our ontology by instances of educational institutes and evolve this knowledge-base, *EduBD* in N-TRIPLE format. The representation of this knowledge-base is demonstrated in the Figure 12 that exhibits the directed labeled graph structure. The graph pattern representation for a small number of triples is portrayed in the Figure 13 that exposes inexpressible visualization of our knowledge-base.



Subject	Predicate	Object
<http://www.skeim.org#002>	<http://www.skeim.org#hasLevel>	<http://www.skeim.org#Secondary>
<http://www.skeim.org#0021>	<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>	<http://www.skeim.org#School>.
<http://www.skeim.org#0021>	< http://www.w3.org/2000/01/rdf-schema#Comment>	"About High School Institutions@en".
<http://www.skeim.org#0021>	< http://www.w3.org/2000/01/rdf-schema#subClassOf>	<http://www.skeim.org#002>.
<http://www.skeim.org#0021105770>	< http://www.w3.org/ns/org#memberOf>	"http://www.skeim.org#0021".
<http://www.skeim.org#0021105770>	< http://www.skeim.org#hasEIIN>	"105770@en".
<http://www.skeim.org#0021105770>	<http://www.w3.org/2000/01/rdf-schema#label>	"Comilla Zilla School@en".
<http://www.skeim.org#0021105770>	< http://www.skeim.org#division>	"Chittagong @en".
<http://www.skeim.org#0021105770>	<http://www.skeim.org#district>	"Comilla@en".
<http://www.skeim.org#0021105770>	< http://www.skeim.org#thana>	"Adarsha Sadar @en".
<http://www.skeim.org#0021105770>	<http://www.skeim.org#postOffice>	"Comilla@en".
<http://www.skeim.org#0021105770>	< http://www.w3.org/ns/org#site>	"Kandirpar@en".

Figure 12: Machine understandable educational institutions data of Bangladesh in N-TRIPLE format

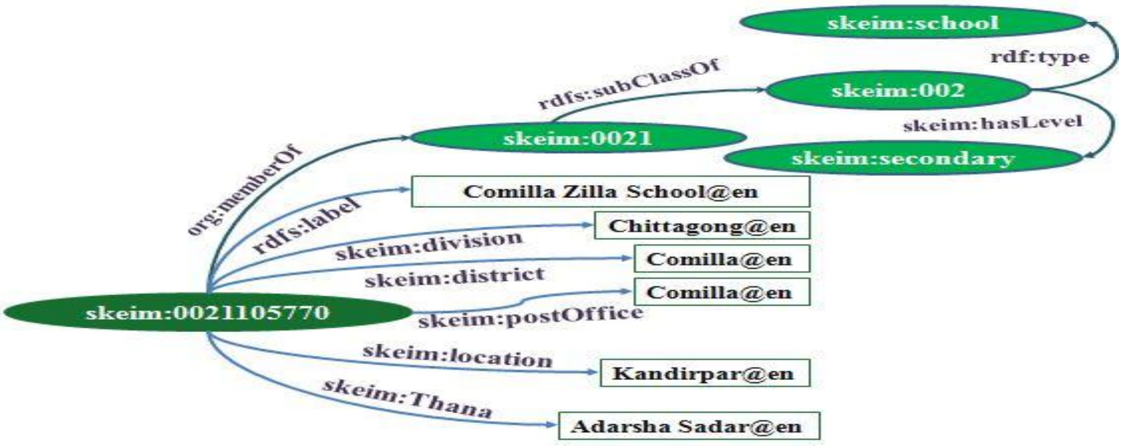
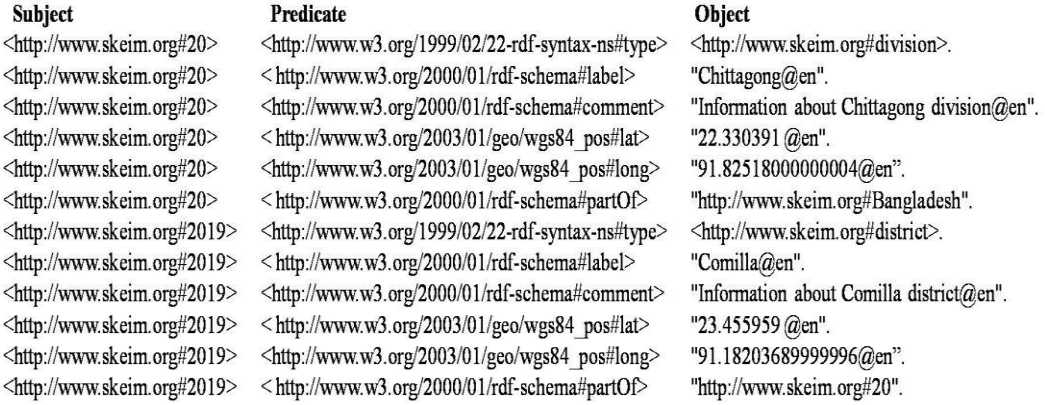


Figure 13: Directed graph structure for knowledge pieces from our EduBD

5.2 Published *EduBD* as Linked Open Data

The semantic web focuses on the ontological level to describe, discover and access web contents to develop interoperable global data-space to infer implicit knowledge-intensive operations that can be carried out automatically [42][43]. In this connection, our domain dataset *EduBD* published as *LOD application* by following the synopsis of Tim Berners Lee. The first step is to use *URIs* as names for things so that *HTTP-URIs* can explore those names on the web. When machine explore available *URIs*, these *URIs* are furnished with information on the web. These information are described using semantic web standards such as Web Ontology Language *(OWL)*, *RDF* and RDF Schema *(RDFS)*. Finally, these definition of data established links to resources of other datasets on the web to discover more compatible knowledge [29].

Furthermore, our knowledge-base *EduBD* reuse resources from LOD repositories, *Geo-Bangladesh* and *HCN-BD* to extract *geo-coordinates* of a location and information of medical institutions by reducing redundant information on the web. The resource sharing process at the same time achieved semantic interoperability issues effortlessly. The data structure of *LOD* repository, *Geo-Bangladesh* is given in the Figure 14 respectively to comprehend the essence of resource sharing approach.

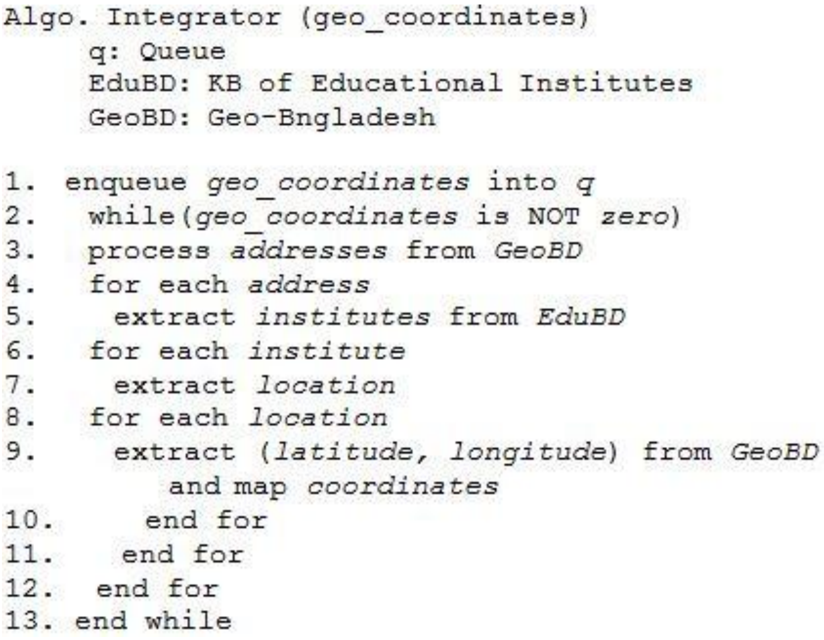


Subject	Predicate	Object
<http://www.skeim.org#20>	<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>	<http://www.skeim.org#division>.
<http://www.skeim.org#20>	< http://www.w3.org/2000/01/rdf-schema#label>	"Chittagong@en".
<http://www.skeim.org#20>	<http://www.w3.org/2000/01/rdf-schema#comment>	"Information about Chittagong division@en".
<http://www.skeim.org#20>	< http://www.w3.org/2003/01/geo/wgs84_pos#lat>	"22.330391 @en".
<http://www.skeim.org#20>	<http://www.w3.org/2003/01/geo/wgs84_pos#long>	"91.82518000000004@en”.
<http://www.skeim.org#20>	< http://www.w3.org/2000/01/rdf-schema#partOf>	"http://www.skeim.org#Bangladesh".
<http://www.skeim.org#2019>	<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>	<http://www.skeim.org#district>.
<http://www.skeim.org#2019>	< http://www.w3.org/2000/01/rdf-schema#label>	"Comilla@en".
<http://www.skeim.org#2019>	<http://www.w3.org/2000/01/rdf-schema#comment>	"Information about Comilla district@en".
<http://www.skeim.org#2019>	< http://www.w3.org/2003/01/geo/wgs84_pos#lat>	"23.455959 @en".
<http://www.skeim.org#2019>	<http://www.w3.org/2003/01/geo/wgs84_pos#long>	"91.18203689999996@en”.
<http://www.skeim.org#2019>	< http://www.w3.org/2000/01/rdf-schema#partOf>	"http://www.skeim.org#20".

Figure 14: The fragment of knowledge representation in Geo-Bangladesh

5.3 Semantic Information Integration using Physical-Cyber-Social *(PCS)* Computing Approach

The operation of data integration between *PCS* data-sources along with information of educational institutions is a challenging task to extract specific information for spatial analysis to apply various policies to improve quality of educational institutions. The algorithm to integrate data from various data sources is portrayed in the Figure 15.



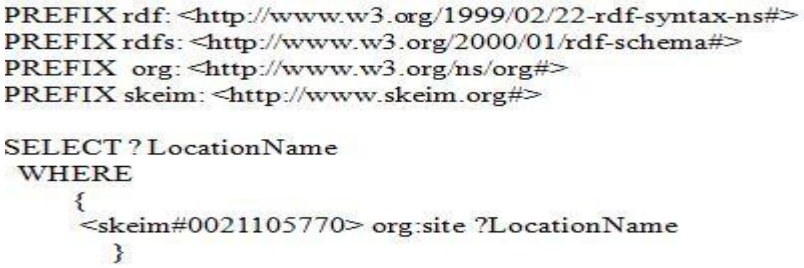
```
Algo. Integrator (geo_coordinates)
      q: Queue
      EduBD: KB of Educational Institutes
      GeoBD: Geo-Bngladesh

1.  enqueue geo_coordinates into q
2.   while(geo_coordinates is NOT zero)
3.   process addresses from GeoBD
4.   for each address
5.     extract institutes from EduBD
6.   for each institute
7.     extract location
8.   for each location
9.     extract (latitude, longitude) from GeoBD
           and map coordinates
10.      end for
11.    end for
12.  end for
13. end while
```

Figure 15: Pseudo code of how our data integrator works.

In the information integration process we use *GPS* enabled sensors as physical tools to extract *geo- coordinates* that are consistently connected and interacted with *Geo-Bangladesh* knowledge-base *(as lines 1 and 2)*. The *GPS* data in the pattern of *{latitude (Lat), longitude (Long)}* are used to fetch semantically related neighboring information in the form of *{division (div) → District (dis) → Upazila (upa)}* by eliminating natural language polisemy problems *(as line 3)*. In this stage we expand our extracted information pattern by examining *http://www:w3:org/ns/org#site* as *{division (div) → district (dis) → upazila (upa) → site}* where *site* indicates the exact location of the institute available *(as line 5 and 7)*. This advanced information pattern is used for getting further action from *Geo-Bangladesh* using semantic query language SPARQL to retrieve spatial coordinates to map these locality to increase search ability that shows the usefulness of our application *(as line 9)*. Moreover our system investigates accessible data sources along with intuitive human knowledge about education services and their respective service qualities of each institution. In our integration process we use an inference engine *(portrayed Figure 10, step 9)* to explore implicit knowledge that assist people in making their decisions about institutions.

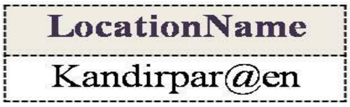
6. EXPERIMENTAL RESULTS AND EVALUATION

We perform a large number of experiments on our machine understandable dataset *EduBD* using semantic search engine, *SPARQL* to explore distinguishing information according to our query on the web. These experimental results based on the relationship of triples show the effectiveness of our research. One of the semantic searches is portrayed in the Figure 16 to fetch the relevant information of the URI, < *http://www:skeim:org#0021105770>*. The corresponding output demonstrates a location name of *Comilla Zilla School* that is illustrated in the Fig. 17.



```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX  org: <http://www.w3.org/ns/org#>
PREFIX skeim: <http://www.skeim.org#>

SELECT ? LocationName
 WHERE
     {
     <skeim#0021105770> org:site ?LocationName
        }
```

Figure 16: The semantic query using SPARQL to retrieve the location of an institution



LocationName
Kandirpar@en

Figure 17: The result of the query that is given in the Figure 16

These semantic query operations promote to address interoperability issues in multiple knowledge repositories of *EduBD*, *Geo-Bangladesh* and *HCN-BD* by using common vocabularies. Moreover, our ontology model, *ontoEduBD* incorporates 32 classes and sub-classes; however 7 of them are extracted at the phase of heterogeneous information integration of our research. Our ontology is populated by more than 41,000 educational institutions and these institutions are interoperable on the web with other compatible knowledge repositories. Our proposed knowledge-base consists of 0.75 million accessible triples to evaluate information sharing issues successfully.

In addition, the advanced operation on our application examines the location in the form of *{latitude (lat), Longitude (long)}* for each institution from *Geo-Bangladesh*. However, we use the spherical law of cosine to calculate the distances between sensors and the location of each institution by using *GPS*. This formula is given in the *Eq. 2* where Φ_1 is the latitude of first position while Φ_2 is the latitude of the second position. The notation $\Delta\lambda$ express the difference between longitude value of two positions and *R* indicates the radius of the earth that is approximately 6371 kilometers.

$$distance = acos\ (sin\ \Phi_1.\ sin\ \Phi_2 + cos\ \Phi_1\ .\ cos\Phi_2.cos\Delta\lambda).R \qquad (2)$$

To evaluate this numerical calculation we consider the *(latitude; longitude)* pairs of *Comilla Zilla School*, *Comilla University*, *Comilla Victoria Government College* are *(23:463689; 91:181114)*, *(23:430282391:1361569)* and *(23:4598; 91:1823)* respectively. The distance calculation results obtained by spherical cosine formula from Comilla Zilla School are 6.71 kilometers and 357 meters respectively. This approach performs on the core of distance calculation to retrieve information of specific institutions from our dataset along with useful information from the web of data.

Furthermore, we use Jena, *a java based semantic web application development framework* for rule based inference and Fuseki, *a SPARQL server* on our data repository for our semantic application.

7. CONCLUSION

Our proposed system is significant for a country to develop a network of educational institutions for originating collaborative framework of information sharing that promote to infer knowledge regarding of their services to improve their individual institution. Moreover, the semantic metadata set *EduBD* decrease duplication of similar data for other applications and this machine understandable knowledge-base address the semantic interoperability that shows its application and performance through the significant quantitative outcomes, both in terms of concepts and educational entities of Bangladesh. In addition, the *PCS* computing on generic knowledge-base *EduBD* extend our research on spatial human sentiment analysis for educational institutions.

8. REFERENCES

[1] T. Bray, J. Paoli, C. M. Sperberg-McQueen, E. Maler, and F. Yergeau, “Extensible markup language (xml),” World Wide Web Consortium Recommendation REC-xml-19980210. http://www. w3. org/TR/1998/REC-xml-19980210, vol. 16, 1998.

[2] J. Shanmugasundaram, K. Tufte, C. Zhang, G. He, D. J. DeWitt, and J. F. Naughton, “Relational databases for querying xml documents: Limitations and opportunities,” in Proceedings of the 25th International Conference on Very Large Data Bases. Morgan Kaufmann Publishers Inc., 1999, pp. 302–314.

[3] A. Silberschatz, H. F. Korth, S. Sudarshan et al., Database system concepts. McGraw-Hill Hightstown, 1997, vol. 4.

[4] C. Martinez-Cruz, I. J. Blanco, and M. A. Vila, “Ontologies versus relational databases: are they so different? a comparison,” Artificial Intelligence Review, vol. 38, no. 4, pp. 271–290, 2012.

[5] O. Olubadewo, I. Abdulkarim, and M. Ahmed, “The use of gis as educational decision support system (edss) for primary schools in fagge local government area of kano state, nigeria,” Academic Research International, vol. 4, no. 6, p. 614, 2013.

[6] D. Barstow, “An introduction to gis in education,” in First National Conference on the Educational Applications of Geographic Information Systems (EdGIS) Conference Report, vol. 14. TERC Communications Cambridge, Massachusetts, 1994, p. 19.

[7] R. H. Audet and G. L. Abegg, “Geographic information systems: Implications for problem solving,” Journal of Research in Science Teaching, vol. 33, no. 1, pp. 21–45, 1996.

[8] J. Chomicki and P. Revesz, “Constraint-based interoperability of spa-tiotemporal databases,” Geoinformatica, vol. 3, no. 3, pp. 211–243, 1999.

[9] Z. Peng, “A proposed framework for feature-level geospatial data sharing: a case study for transportation network data,” International Journal of Geographical Information Science, vol. 19, no. 4, pp. 459– 481, 2005.

[10] Y. Bishr, “Overcoming the semantic and other barriers to gis interop-erability,” International Journal of Geographical Information Science, vol. 12, no. 4, pp. 299–314, 1998.

[11] F. Harvey, W. Kuhn, H. Pundt, Y. Bishr, and C. Riedemann, “Semantic interoperability: A central issue for sharing geographic information,” The Annals of Regional Science, vol. 33, no. 2, pp. 213–232, 1999.

[12] C. Jones, H. Alani, and D. Tudhope, “Geographical information retrieval with ontologies of place,” Spatial Information Theory, pp. 322–335, 2001.

[13] A. Cohn, “The challenge of qualitative spatial reasoning,” ACM Com-puting Surveys, vol. 27, no. 3, pp. 323–325, 1995.

[14] J. Brodeur, Y. Bedard, G. Edwards, and B. Moulin, “Revisiting the concept of geospatial data interoperability within the scope of human communication processes,” Transactions in GIS, vol. 7, no. 2, pp. 243– 265, 2003.

[15] F. Fonseca, M. Egenhofer, P. Agouris, and G. Camara,ˆ “Using ontolo-gies for integrated geographic information systems,” Transactions in GIS, vol. 6, no. 3, pp. 231–257, 2002.

[16] T. Berners-Lee, J. Hendler, O. Lassila et al., “The semantic web,” Scientific american, vol. 284, no. 5, pp. 28–37, 2001.

[17] A. Sheth, P. Anantharam, and C. Henson, “Physical-cyber-social com-puting: An early 21st century approach,” Intelligent Systems, IEEE, vol. 28, no. 1, pp. 78–82, 2013.

[18] M. Arenas and J. Perez,´ “Querying semantic web data with sparql,” in Proceedings of the thirtieth ACM SIGMOD-SIGACT-SIGART sympo-sium on Principles of database systems. ACM, 2011, pp. 305–316.

[19] M. H. H. Rahman, S. Chakraborty, and M. H. Seddiqui, “Machine understandable information representation of geographic related data to the administrative structure of Bangladesh,” in Computer and Information Technology (ICCIT), 2013 16th International Conference on. IEEE, 2014, pp. 236–241.

[20] E. D. Kaplan and C. J. Hegarty, Understanding GPS: principles and applications. Artech house, 2005.

[21] M. Mouly, M.-B. Pautet, and T. Foreword By-Haug, The GSM system for mobile communications. Telecom Publishing, 1992.

[22] M. Rahnema, “Overview of the gsm system and protocol architecture,” Communications Magazine, IEEE, vol. 31, no. 4, pp. 92–100, 1993.

[23] T. Gruber et al., “Toward principles for the design of ontologies used for knowledge sharing,” International journal of human computer studies, vol. 43, no. 5, pp. 907–928, 1995.

[24] R. Studer, V. R. Benjamins, and D. Fensel, “Knowledge engineering: principles and methods,” Data & knowledge engineering, vol. 25, no. 1, pp. 161–197, 1998.

[25] T. Berners-Lee and M. Fischetti, Weaving the Web: The original design and ultimate destiny of the World Wide Web by its inventor. DIANE Publishing Company, 2001.

[26] A. Maedche and S. Staab, “Ontology learning for the semantic web,” Intelligent Systems, IEEE, vol. 16, no. 2, pp. 72–79, 2001.

[27] M. Ehrig, Ontology alignment: bridging the semantic gap. Springer, 2006, vol. 4.

[28] G. Antoniou and F. Van Harmelen, A semantic web primer. MIT press, 2004.

[29] C. Bizer, T. Heath, and T. Berners-Lee, "Linked data-the story so far," International journal on semantic web and information systems, vol. 5, no. 3, pp. 1–22, 2009.

[30] T. Heath and C. Bizer, "Linked data: Evolving the web into a global data space," Synthesis Lectures on the Semantic Web: Theory and Technology, vol. 1, no. 1, pp. 1–136, 2011.

[31] A. Geraci, F. Katki, L. McMonegal, B. Meyer, J. Lane, P. Wilson, J. Ra-datz, M. Yee, H. Porteous, and F. Springsteel, IEEE standard computer dictionary: Compilation of IEEE standard computer glossaries. IEEE Press, 1991.

[32] M. F. Goodchild, Interoperating geographic information systems. Springer, 1999.

[33] W. Wayne, "The good news and the bad news (embedded computing column," IEEE Computer, vol. 40, no. 11, p. 104, 2007.

[34] H. Ning and B. Wang, "Rfid major projects and state internet of things," 2010.

[35] H. J. Scholten and M. J. de Lepper, "The benefits of the application of geographical information systems in public and environmental health," World Health Stat Q, vol. 44, no. 3, pp. 160–170, 1991.

[36] O. O. Alharaki, F. S. Alfieri, and A. M. Saki, "The integration of gps navigator device with vehicles tracking system for rental cars firms," International journal of Computer Science and Information Security, 2008.

[37] S. Redl, M. Weber, and M. Oliphant, GSM and personal communica-tions handbook. Artech House, Inc., 1998.

[38] S. Chakraborty, M. H. H. Rahman, and M. H. Seddiqui, "Linked open data representation of historical heritage of Bangladesh," in Computer and Information Technology (ICCIT), 2013 16th International Conference on. IEEE, 2014, pp. 242–248.

[39] H. H. Rahman, S. Chakraborty, and M. H. Seddiqui, "Semantic information integration of health care network for physical-cyber-social computing approach," in Computer and Information Technology (ICCIT), 2014 17th International Conference on. IEEE, 2014, pp. 305–313.

[40] J. Hedley, "jsoup: Java html parser," 2010.

[41] N. Noy, D. McGuinness et al., "Ontology development 101: A guide to creating your first ontology," 2001.

[42] Y. Ding, D. Fensel, M. Klein, and B. Omelayenko, "The semantic web: yet another hip?" Data & Knowledge Engineering, vol. 41, no. 2, pp. 205–227, 2002.

[43] M. C. Daconta, L. J. Obrst, and K. T. Smith, The Semantic Web: a guide to the future of XML, Web Services and knowledge management. John Wiley & Sons, 2003.

Authors

Shima Chakraborty obtained B.Sc. in 2009 and MS (Engg.) in 2012 in Computer Science and Engineering from University of Chittagong. She is presently working as a lecturer in Department of Computer Science and Engineering at University of Chittagong. Her research interest includes Big Data, Semantic Web, Data Mining, Artificial Intelligence, and Machine Learning. She has published research articles in various national and international conferences.

Md. Hasan Hafizur Rahman completed his B.Sc. degree in 2009 and MS (Engg.) in 2012 in Computer Science and Engineering from Chittagong University. Now, he is working as a faculty member of the department of Computer Science and Engineering, Comilla University, Bangladesh. His current research interest on Semantic web, Artificial Intelligence and Machine Learning.

Dr. Md. Hanif Seddiqui received his B.Sc. Eng. degree in Electronic and Computer Science from Shahjalal University of Science and Technology, Sylhet, Bangladesh in 2000 and his M.Eng. and D.Eng. degree in Computer Science from Toyohashi University of Technology, Japan in March 2007 and in March 2010 respectively. He is currently working as a Professor at the Department of Computer Science and Engineering, University of Chittagong, Bangladesh. His current research interest includes Ontology Alignment, Knowledge Engineering, Bioinformatics and Semantic Web Techniques in Information Retrieval and Big Data.

Sajal Chandra Debnath completed B.Sc. (Hons) and M.Sc. in Geography and Environmental Studies from Chittagong University. Now he is working as a GIS and Remote Sensing Analyst in National Land Zoning Project at Ministry of Land, The People’s Republic of Bangladesh.