An Ontology and Semantic Metadata based Semantic Search Technique for Census Domain in a Big Data Context

Sanjay Ajani
PG Scholar, G.T.U. P.G. School,
Gujarat Technological University, Ahmedabad-382 424, India

***Abstract*-In the current world, we are faced with a massive data sets having complex, varied and large volume which is generated and captured in the form of digital resource with multiple sources. The big Data concept uses in the current and future world where scientific pursuits and human endeavours will be aided by not only the financial, and physical assets, but also digital data assets. The main challenges in making use of Big Data dealing with the complexity, the diversity and the heterogeneity of data. Now semantic web technologies concepts like ontologies and semantic metadata are dealing with these issues. In this paper, I proposed a semantic metadata and ontology based semantic search technique for large census data in the context of Big Data which may helpful for government as well as third party agencies to enhancing the semantic search on census data to perform various actions and taking decision.**

Keywords—Big Data, Ontology, Semantic Metadata, Census.

I. INTRODUCTION

Big Data refers to a massive datasets whose size are very large and it is beyond the ability of typical database software tools. It is difficult to capture the data, storing the data, managing the data and analysing the data. There is no explicit definition of how large a dataset should be considered as the Big Data.

Big Data is not just about the size of data but also refers to several other characteristics like data volume, data variety and data velocity. Combination of these three attributes form the three V's of Big Data [1]. These parameters leads to complexity as well.

Volume: It is synonymous with the "Big" in the term, "Big Data". Volumes of data have increased exponentially in recent times and will continue to grow, regardless of the organisations size [1, 2]. Many of these companies have their datasets are within the terabytes range today but, soon they could reach Peta-bytes or even Exabyte.

Variety: Data can come from a variety of sources and in a variety of types. With the explosion of social networking, smart devices as well as sensors data in an enterprise has become more complex because it includes not only structured traditional relational data but also semi-structured and unstructured data. It is not easy to put massive dataset into a relational database.

Dealing with a variety of structured and unstructured data, it is greatly increases the complexity of both storing and analysing Big Data [1, 2].

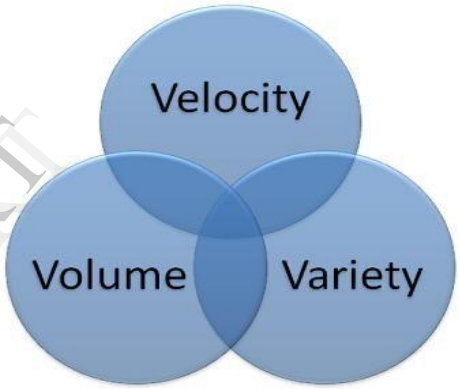

Figure 1-Three V's of Big Data. [1]

Velocity: The frequency of generation of data and delivery of data referred as velocity is also a characteristic of big data.Basic understanding of velocity is how quickly the data arrives and is stored, and how quickly it can be retrieved [1, 2]. Big Data is not just about the volume of datasets but, also important is the rate of change of the data.

A. Big Data Inconsistencies

In circumstances where big data from various resources are produced, acquired, aggregated, transformed or represented, inconsistencies are invariably find their way into massive datasets. Inconsistencies can also been produced in heuristics, reasoning methods or problem-solving approaches. These are deployed for various analysis tasks, resulting in complications for big data analysis and decision-making process.

Inconsistencies can be attributed to a number of factors in the decision-making process and in the human behaviours. Once inconsistencies captured in Big Data, it occur at various granularities of knowledge base, from digital data, information

and knowledge to expertise the domain knowledge. It can affect the quality of the outcomes of Big Data. In circumstances where big data from various resources are produced, acquired, aggregated, transformed or represented, inconsistencies are invariably find their way into massive datasets. Inconsistencies can also been produced in reasoning methods, heuristics or problem-solving approaches which are deployed for various analysis tasks, resulting in complications for big data analysis and decision-making process. Inconsistencies can be attributed to a number of factors in decision-making process and in human behaviours. Once inconsistencies captured in Big Data, conflicting phenomena occur at various granularities of knowledge base, from digital data, information, knowledge and meta-knowledge, to expertise the domain knowledge, and can adversely affect the quality of the outcomes of Big Data in its analysis process.

II. MOTIVATION

Currently, various technologies are available and that are being applied to handle the Big Data. All organizations struggle to handle the Distributed Information effectively and which is costly to their business. The current technologies doesn't solve or answer the question of what is going to apply those resources. To solve this problem of the Distributed Information, Semantic Web Technologies were created. To extracting meaningful information from structured, unstructured and semi -structured data of large datasets, Semantic Web Technology is used for that. The relationship with data is handled in more flexible way as compared to traditional database by Semantic Web Technologies.

A. The Role of Ontology and Metadata

Ontology: An Ontology is an explicit representation of knowledge. It is a formal, explicit specification of shared conceptualizations, representing the concepts and their relations that are relevant for a specific domain of discourse [16]. It consists of a representational vocabulary with precise definition of the meaning of the terms plus a set of axioms.
To represent ontologies, several ontology languages have been proposed. OWL (Web Ontology Language) is based on a logic of description [15]. OWL is designed to use by various applications which uses it to process the web-based content. OWL (Web Ontology Language) facilitates an Interpretability of web-based content gathered from web resources. An example representation of OWL is below here:

```
<GeographicalRegionrdf:ID="Saurashtra">
<hasName>Saurashtra</hasName>
<inside>
<State rdf:ID=Gujarat">
<hasName>Gujarat</hasName>
</State>
</inside>
<include>
<Province rdf:ID="Rajkot">
<hasName>Rajkot</hasName>
</Province>
</include>
<include>
<Province rdf:ID="Jamnagar">
<hasName>Jamnagar</hasName>
</Province>
</include>
<include>
<Province rdf:ID="Surendranagar">
<hasName>Surendranagar</hasName>
</Province>
<include>
<Province rdf:ID="Porbandar">
<hasName>Porbandar</hasName>
</Province>
</include>
</GeographicalRegion>
```

OWL Representation for Saurashtra Region.

Here, Saurashtra is a geographical region and it is an instance of Gujarat State. The State includes four provinces named Rajkot, Jamnagar, Surendranagar and Porbandar. These all are an instance of province.
In OWL, the relations between classes can be defined as properties. OWL can represent the meaning of the domain terminology. It allows to perform reasoning on particular domain specific document and also used as the language to represent the ontology.
Metadata: An important parameter of the information quality is how well its user understand it. Semantic Metadata called "Machine-readable metadata" is an important contributor to understanding structured data. To integrate data, model driven techniques uses the metadata component. Metadata enabling the current crop of graphical data mapping tools. These tools and techniques provides integration of structured data less costly and error-prone.

B. Census Information

The Census of any country or region is a leading source of facts and figures about a country's development and decision making process in all sectors. Census provides the detailed information about a country's population census and the increasing-decreasing ratio of the peoples who lives there and the people who operate them. It is the only reliable source of uniform, comprehensive and accurate census data for every state and county in the country. Census information is used by many who provide services to peoples and other communities including state governments and municipal corporations, businesses etc. The census data uses in a variety of social and economic areas. This census data is mainly used to monitoring the progress and in analysing population's gender, marital status, languages, religion and geographic region issues.

III. PROPOSED TECHNIQUE METHODOLOGY

Semantic Web Technology used by Semantic Data Model for Census Domain is to help the government. It is based on RDF (Resource Description Framework), XML (Extensible Mark-up Language) and OWL (Web Ontology Language) language data format. It contains raw unstructured and semi-structured census government data which is spreads across various resources. Protégé tool Framework can be used to design and develop the conceptual model into RDF (Resource Description Framework), XML (Extensible Mark-up Language) and OWL

(Web Ontology Language) language data format dataset. It can be used to get the suitable datasets from massive data. The protégé tool framework helps to define and describe the massive or large datasets and correlating the massive datasets in suitable among the part of datasets. It is used to establish the relationship between Subject-Predicate-Object and the inherent feature of Resource Description Framework. It helps to inferring the class as well as reasoning and validating the subclass of parent class and their relationship between them.

Finding & Justifying Relevant Concept:

The Conceptual model can be identified and justified based on the information, the facts and the data. The massive datasets are available from various sites and from various resources. Current system having some problems/drawbacks which will be helpful to find and justify the new way to provide a better solution. To design a domain specific ontology, the process steps are described here:

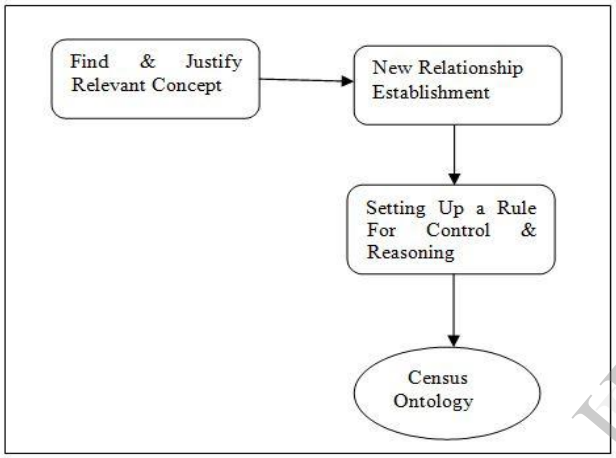

Figure 2- Process Steps to Design Census Ontology.

Semantic web technology based semantic data model will be promising to provide a better solution as in terms of cost and benefits and more. Most of distributed infrastructure which are supporting and they faced the failure of the current system after resulting process. Here the results of semantic data model can be completed through the graph based semantic search which is based on optimization process. The result may give justification to adopting the semantic web technology for new generation.

New Relationship Establishment:

The basic building block of Semantic Web Technology is to find the strong relationship between subject-predicate-object formats which is creating a triplet. The triplet which is inferring the Uniform Resource Identification (URI) to the web sites makes the strong relationship between various nodes which resides on graph network and of its interest. To provide the semantic search on specific datasets, Query is running and completed through specific query running-handling process which is having valid criteria to getting desired results.

Setting Up a Rule for Control and Reasoning:

Various constraints imposed to set up the laws and SWRL axioms which are used to provide guidance to the semantic search criteria and it also used for specific purpose and gives a movement to the right destination choice dynamically. In the proposed solution, OWL can be used for describing rules to guide the searching path, reasoning and inferring during clients request is based on query.

IV. FUNCTIONAL ARCHITECTURE OF PROPOSED ONTOLOGY BASED SEMANTIC SEARCH IN CENSUS DOMAIN

Following figure 3 shows the functional flow diagram of the ontology based system with the use of domain specific ontology described above. Here, K.B. stands for Knowledge Base.

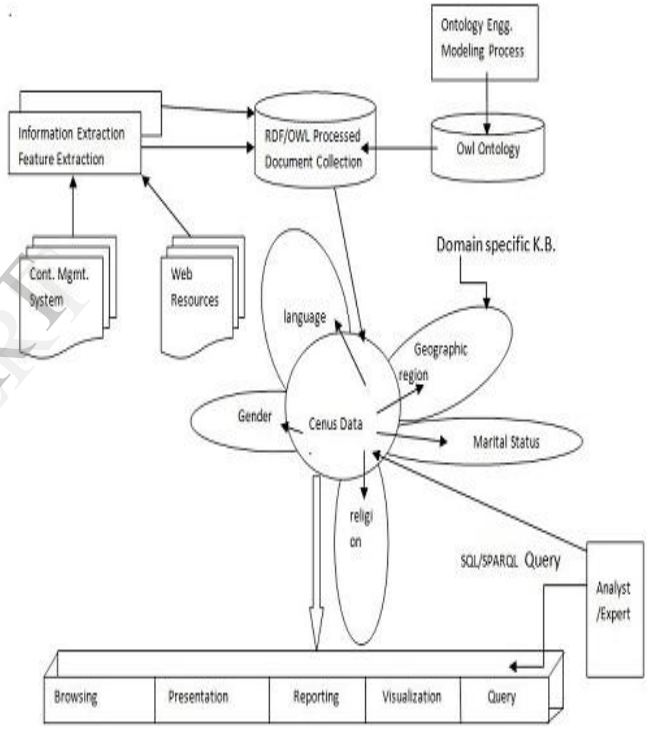

Figure 3- Functional diagram for proposed Ontology Based semantic search in Census domain using domain specific ontology.

Domain Specific Ontology Design & Creation:

The Ontology created for specific domain is represented in XML format. Ontology models knowledge for census model is Logical Model. It may possible to attach some useful components with the logical model of census domain. These components can be generalized by depicting it. As per the characteristics and functionality of system, some components can attached on various resources and other concerning dependent bodies.

Content Management System: All the features and terms, local datasets in structured/unstructured/semi structured format, web resources which are available can be handled by Content management system.

Ontology Engineering Process: To get the data into RDF-OWL format, resources gets processed and it can be stored into database. Each of component process will guided by OWL with demand and its proper analysis. The reasoning is done through Ontology engineering Process.

Expert: The resources can be availed and query of interest is done by Expert.

A. Query Processing for Semantic Search in Census domain:

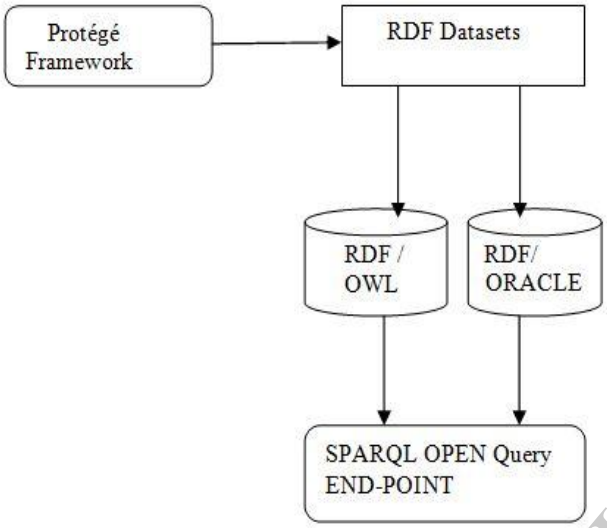

Figure 4- Query Processing for Census Domain.

Protégé Framework: It helps to define and describe the massive or large datasets and to correlate the massive datasets in suitable among the part of datasets. In the proposed system, SPARQL end-point can be used to generating the query to get the desired result. From massive census data, to mining data of specific interest and to see the desired query result, Query is generated in triplet form.

RDF/OWL: The Data in RDF format can be taken by SPARQL end-point and the desired query result will be generated by it if any criteria is suitable and set for the query**.**

SPARQL end-point: It can be used for the query processing.

B. Benefits of the Proposed Technique

This technique can be very helpful to government as well as Municipal Corporation to enhancing semantic search about the census with specific purpose.

An ontology is created for census domain which plays an important role in organizing distributed information. Semantic metadata provides data integration with distributed datasets. Using an ontology and metadata based semantic search technique, Admin/expert bodies of government or organization have no need to go to monitor the traditional database of census individually to search the desired information. As shown in figure 3, proposed technique can show the process of it.

An approach facilitates to analyse population's gender, marital status, languages, geographic region, and religion issues. If government wants to examine some particular location for analysis of census with specific intent, it can be done with directly from the large datasets by extracting information with the proposed technique.

V. CONCLUSION

The Census of any country or region is a leading source of facts and figures about a country's development and decision making process in all sectors. This census data is mainly used to monitoring the progress and in analysing population's gender, marital status, languages, geographic region, and religion issues. Census information is used by many who provide services to peoples and other communities including state government, municipal corporations, businesses etc. The census data uses in a variety of social and economic areas. It is necessary to enhancing search for information retrieval form census data. To fulfil the need, an ontology based approach is proposed.

In this paper, I have proposed a semantic metadata and ontology based semantic search technique for large census domain in the context of Big Data. It may helpful for govt. as well as third party agencies to enhancing the semantic search on census data and to perform various actions and taking decision. From large census data, Gender monitoring, Region wise monitoring, Language wise Monitoring, Marital status, Religion wise monitoring can be possible with semantic search using the proposed technique. An ontology can play an important role in organizing distributed information related to the specific domain of interest. In the Proposed Technique, Protégé framework is used to provide ontology-based semantic search for specific domain. RDF/OWL dataset has potential to establish Relationship with URLs in natural way. SPARQL end-point can be used to generate the query for the desired result from massive census data, to mining data of specific interest.

This paper is focusing on building the census domain specific ontology and performing semantic search on it to get desired result from large datasets.

REFERENCES

[1]http://www.ida.gov.sg/~/media/Files/Infocomm%20Landscape/Technology/TechnologyRoadmap/BigData.pdf
[2] Anett Hoppe, "Automatic Ontology-based User Profile Learning from Heterogeneous Web Resources in a Big Data Context", Checksem Group, LE2I Universite' de Bourgogne Dijon, France.
[3] Fabian Suchanek, Gerhard Weikum, "Knowledge Harvesting in the Big Data Era", Max Planck Institute for Informatics, D-66123 Saarbruecken, Germany.
[4] Yaoyong Li,KalinaBontcheva, "Hierarchical, Perceptron-like Learning for Ontology Based Information Extraction", University of Sheffield211 Portobello StreetSheffield, S1 4DP, UK, WWW 2007 / Track: Semantic Web.
[5] SwaranLata, Bhaskar Sinha, Ela Kumar, Somnath Chandra, Raghu Arora, "Semantic Web Query on e-Governance Data and Designing Ontology for Agriculture Domain", Department of Electronics and Information Technology, New Delhi, India.
[6] GuntarsBumans, "Mapping between Relational Databases and OWL Ontologies: an Example", Department of Computing, University of

Latvia. Raina bulv. 19, Riga, LV-1586, Latvia. Scientific Papers, University of Latvia, 2010.

[7] Christian Bizer, Peter Boncz, Michael L. Brodie, OrriErling, "The Meaningful Use of Big Data: Four Perspectives – Four Challenges", Web-based Systems Group, FreieUniversität Berlin; Centrum Wiskunde&Informatica, Amsterdam; Verizon Communications USA, OpenLink Software, Utrecht.

[8] AhmetSoylu, Martin Giese, Ernesto Jimenez-Ruiz, "OptiqueVQS – Towards an Ontology-based Visual Query System for Big Data", University of Oslo, Noway, University of Oxford, Oxford, UK.

[9] Jihyun Lee, Jun-Ki Min, Chin-Wan Chung, "An Effective Semantic Search Technique using Ontology", Korea Advanced Institute of Science and Technology Yuseong-gu, Guseong-dong Daejeon, Republic of Korea. April 22, 2009.

[10] Djamel NESSAH, Okba KAZAR, "Document Analysis to Provide Semantic Metadata based Ontologies", Department of Computer Science, Abbes Laghrour University CenterKhenchela, Algeria, IEEE 2012.

[11] Du Zhang, "Inconsistencies in Big Data", Department of Computer Science California State University Sacramento, CA 95819-6021, IEEE 2013.

[12] Seref SAGIROGLU, Duygu SINANC, "Big Data: A Review", Gazi University, Department of Computer Engineering, Faculty of Engineering Ankara, Turkey, IEEE 2013.

[13] David S. Frankel,"The Role of Semantic Metadata in Improving Information", Lead Standards Architect, Technology Strategy Group, SAP Labs, LLC, The Fifth MIT Information Quality Industry Symposium, July 2011.

[14] Jihie Kim, Yolanda Gil and VarunRatnakar, "Semantic Metadata Generation for Large Scientific Workflows", Information Sciences Institute, University of Southern California 4676 Admiralty Way, Marina del Rey CA 90292, United States, The 5th International Semantic Web Conference, Athens, GA, USA, , ISWC-2006.

[15] Xin Wang, Howard J. Hamilton, "Towards an Ontology-BasedSpatial Clustering Framework", Department of Computer ScienceUniversity of Regina, Regina, SK, Canada S4S 0A2.

[16] Gruber, T. R.: A translation approach to portable ontologies. Knowledge Acquisition,