Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 03, March-2019

Managing Data Provenance in the Semantic Web

Parth Sabnani
School of Computer Science
Lakehead University
Thunder Bay, Canada

Sarthak Kothari
School of Computer Science
Lakehead University
Thunder Bay, Canada

Vikas Trikha
School of Computer Science
Lakehead University
Thunder Bay, Canada

***Abstract*—Web is the most popular information source as it got precious information and is the most preferred choice. However, with the increasing amount of data on the web often poor quality, irrelevant and inaccurate data can also be found which further raises the question for credibility. Credibility can be one of the important criteria for the data quality which means "finding the origin of the data" and hence it leads to the provenance of the data. This thesis focuses on employing the provenance in the semantic web and further discusses different frameworks for querying and reasoning the RDF datasets to explore provenance with the help of SPARQL language to derive facts to find the origin. This approach would aim to add a layer of trustworthiness and reliability in the semantic web which would add up to the data quality.**

***Keywords*—RDF (Resource Description Framework), Semantic web, SPARQL query language, credibility, origin**

I. INTRODUCTION

The web is the most popular information source as it got precious information and is the most preferred choice. Since it is one of the most important information programs, a lot of information is being flooded on the web daily, and in today's digital market it influences the means of our living and doing commerce. Certainly, internet upholds a very precious knowledge and yet, poor quality and irrelevant data can be found on it. One of the causes for this can be the internet's flexible constitution where anyone can post and edit any data without any permission to form and publish information all over the world. One of the biggest elements of this scenario is Facebook and Twitter which have been emerged as primary news and marketing source. In such cases, wrong facts and figures can be circulated all over the media in a very short span. Further, it leads to dishonest inferences. This can be understood by various examples listed such as the departure of MH370, a popular news channel NBC declared that the plane had landed safely.

Another piece of misinformation was posted on www.guardian.com by printing the pictures of 2 suspects for the Boston Marathon bombing who had ultimately nothing to do with the case.

One of the most important means to find the authenticity and reliability and trustworthiness of data is to find its origin. It helps to find how genuine and valid the data is and above all how appropriate it is. The information received on this basis can be referred to as provenance.

Provenance information helps determine the trustworthiness and credibility of the information. It contains the source of information from where it has been generated. It can further explain if the information has been re-used or is integrated with other sources of information. The exact definition of provenance can be "Provenance of a resource is a record that describes entities and process involved in producing and delivering or otherwise influencing that resource. Provenance provides a critical foundation for assessing authenticity, enabling trust and allowing reproducibility. Provenance assertions are a form of contextual metadata and can themselves become important records with their provenance." [Ref: https://www.w3.org/2005/Incubator/prov/wiki/What_Is_Provenance]

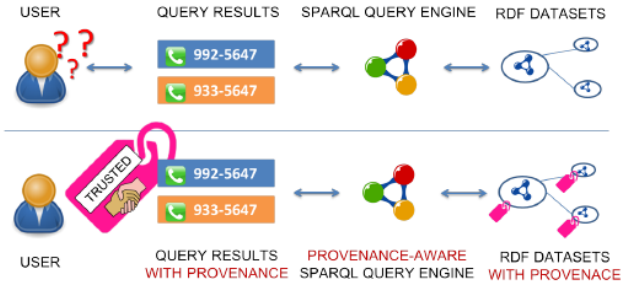

Fig1: provenance adding the web-proof layer of trust

Provenance helps in adding a layer of credibility in the semantic web which would help in upgrading the data quality and even would ease the process in data management. In this, we would study three major protocols which are used and supported by semantic web technologies: RDF (Resource Description Framework), SPARQL Query Language, Web Ontology Language (OWL). RDF can be referred to as a shared prototype where data can be readily blended and connected with the other data or information available on the web. This task can achieve by creating links between web resources which is why it can be coined as interlinked data resources. A good quality data would be able to provide the answer to the following questions to fulfill the needs of provenance:

- Where is the data from?
- Who provided the data?
- When was the data provided?
- Was the provider-specific about the authenticity of the data?
- Was this data believed by others?

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 03, March-2019

Moreover, OWL ontologies can be used to derive the reliability of the information with the help of information derived in provenance. Ontologies define the set of concepts and relationships with the help of classes and relationships. The ontologies in the semantic web can be coined as "vocabularies" which is the basic building block of the inference procedures on the semantic web. SPARQL query language uses the knowledge set of ontology classes and extracts the information from the interlinked data spaces and helps to track the provenance in many dimensions by exploring the data.

II. RELATED WORK

Finding provenance is a topic of research for many years, and relatively different authors have provided with different theories. Broadly, it can be categorized into two approaches: data-oriented approaches and process-oriented approaches. The key focus in the data-oriented approach remained stuck to data items only whereas in the process-oriented approach it focuses on the knowledge about the processes that utilize and produces the data. Different provenance models have already been introduced by W3C (World Wide Web Consortium), and different documentations have already been published for different models. The open provenance model has been explained in the form of graphs in which nodes used are representing artifacts, processes, and agents. Open provenance model can be used to understand the information of parts of the provenance graph. The core components of the PROV model are entity, activity and agents. [Ref: https://www.w3.org/TR/2013/REC-prov-dm-20130430/]
The things which we want to derive the provenance of are called entities. An activity is something which over some time and appears on or with entities. A more technical concept of provenance is described in the form of RDF graphs which represents the RDF data in the form of RDF triples which are subject, object and predicate. A predicate defines the relation between a subject and object and in the semantic web, RDF graphs lays the foundation of semantic web which is interlinked to each other. This keeps the information in the semantic web more structured and tagged, and hence it helps in the extensibility of the data. PROV family of documents is supplied by W3C which defines the general overview over provenance, and this defines a set of specifications which consists of a data model and an OWL ontology with serialization for representing the aspects of provenance. Following documents have been included in the PROV model:

- PROV- PREMIER: offering an introduction to the provenance data model.
- PROV-O: defines lightweight OWL2 ontology.
- PROV-XML: defines an XML schema for the provenance data model.
- PROV-DM: provides conceptual data model for provenance including UML diagrams.
- PROV-N: provides a human-readable notation for the provenance model.
- PROV-CONSTRAINTS: defines a set of constraints on the PROV data model that specifies a notion valid provenance.
- PROV-AQ: defines how to use web-based mechanisms to locate and retrieve provenance information.
- PROV-DC: defines mappings of Dublin Core and PROV-O
- PROV-DICTIONARY: defines construct for expressing the provenance of dictionary-style data structures.
- PROV-SEM: defines a declarative specification in terms of first-order logic of the PROV data model.
- PROV-LINKS: defines extensions to PROV to enable linking provenance information.

III. PROVENANCE FOUNDATIONS

Since 2007, a large volume of provenance has already been done, and it has been quite diverse in the results. Provenance helps in constructing an additional layer of trustworthiness since it explains where the data has come from and who is providing you the data. Many surveys have been done so far to study the various aspects of provenance in terms of foundations, challenges, and opportunities of managing provenance in the semantic web.
Provenance was originally read as an extension of relational databases and was later inherited as RDF knowledge bases, and similarly, the semantic web is an extension of World Wide Web with interlinked web pages, but it goes from linked documents to linked data. It defines a shared framework that provides data available in a structured and tagged way and is readily available in a machine-readable format which makes it easy for data to get re-used and moreover to integrate the data in other platforms also. Its main purpose was to study the data origin which was later converted in RDF datasets. The provenance which was initially taken as the key focus was subcategorized into three classes:

- Where-Provenance: Where the given pieces of data are physically stored in data tuples.
- Why- Provenance: Which subset of the data tuple contributed to the result?
- How-Provenance: How the given data is helping forward to conclude for a result.

Example of Database Tuple 1 (programs and channel)

channel	program	Provenance
TVOntario	Eco Engineering	A1
Star Sports 1	Cricket Premier League	A2
CTV News Channal	In City News	A3
Star Sports 1	FIFA	A4
HBO Canada 1	Man of Steel	A5
CTV News Channel	Morning News	A6
Neo Sports	FIFA	A7

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 03, March-2019

Example of database tuple 2 (Program and genre)

Program	Genre	Provenance
In city News	News	b1
FIFA	Sports	b2
Man of Steel	Movie	b3
Morning News	News	b4
Cricket Premiere League	Sports	b5

The progress took its pace with the formation of RDF graph models and its query language which is SPARQL and further representing ontologies with the web ontology language (OWL). RDF model was providing a platform to interchange data on the semantic web. Since RDF graphs were represented in nodes, it made it more flexible to describe resources. Nodes were meant to cover internationalized global identifier (IRI) which served as a global identifier which uniquely identified any resource globally. Through the RDF model and its foundation, more formalism was attained in the world of provenance information. RDF data consists of PREFIX declarations, dbo (DBpedia ontology classes), RDF type and its functional properties. SPARQL lays its foundation from the roots of SQL, is a graph matching query language. Following syntax has been used by SPARQL for querying and extracting the data:

- Prefix declarations: defines URI prefixes.
- Dataset clause: It is done using the FROM clause which tells from where the data has to be extracted. It can be performed on the union of one or more named graphs.
- Query forms: specifies what type of query is being performed by using keywords SELECT or CONSTRUCT, it tells which data has to be extracted from the table.
- Query clause: specifies the query patterns that are matched against the data and used to generate variable bindings.

Following is the example of a sample RDF graph constructed:

```
# PREFIX DECLARATIONS
@prefix dbo: <http://dbpedia.org/ontology/>
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
@prefix dbp: <http://dbpedia.org/property/>
# DATASET CLAUSE
FROM <http://www.mythesis.de/ToyScenario/FilmsIWantToSee.rdf>
# RESULT CLAUSE
SELECT DISTINCT ?film ?duration
# QUERY CLAUSE
WHERE { ?film rdf:type dbo:Film ;
        dbp:runtime ?duration.}
# SOLUTION MODIFIER
LIMIT 2
```

Fig 2: RDF graph in the form of XML

Linked data uses the principles of the semantic web to ease the data re-usability. The structure data could contain links to further data which enables users to do more exploration on data. To deal with the problems of linked data, a new approach known as crowdsourcing has been used to analyze the data on a given basis. Crowdsourcing makes use of the metadata which is certainly supplied by agents.

IV. RESOURCE DESCRIPTION FRAMEWORK DATA MODEL

RDF stands for resource description framework. It was first originated as structuring the metadata about web sites, pages, etc. which collected the information about the authors, creator, publishers, editors and the data about them like email, phone, job, etc. The first version of RDF was recommended by W3C in 1999and it was specifically designed in XML. The creators or designers who proposed the idea behind RDF were Berners-Lee, Hendler, and Lassila which described this model as a platform for data exchange on the semantic web. A typical RDF model would be consisting RDF graph which would be having nodes and edges, and the key elements of this graph would be: Subject, Predicate, and Object.

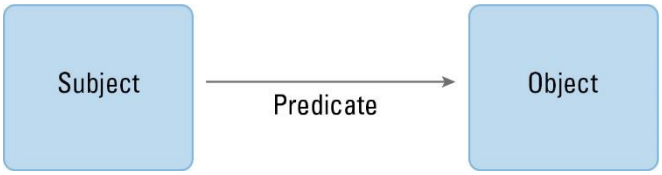

Fig3: Model for subject, predicate, and object

On the web, we have typically kind of identifiers: URL's, URI's and IRI's.
URL stands for Uniform Resource Locator, and it provides the web address for an information resource like website, blog, etc. URI stands for Uniform Resource Identifier which was earlier also known as URN (Universal Resource Name), in some cases it looks like URL only but it might identify something else, and ISBN support it. Every URL is URI but not vice versa. IRI stands for international Resource Identifier, and it uses Unicode instead of ASCII. Every IRI can be turned in to URI with the help of encoding. RDF is accompanied by QNames which is nothing but short hand for long URI's. For example, if prefix foo: is bound to http://example.com then foo: bar expands to http://example.com/bar. It is not the same as XML namespaces. URI's helps to remove ambiguity in naming conventions. The same identifier means the same thing. Resources and relationships are named with URI's. Resources can be referred to IRI's where the subject, the object can act as resources. Sometimes "triples" can also be referred to as "statement."

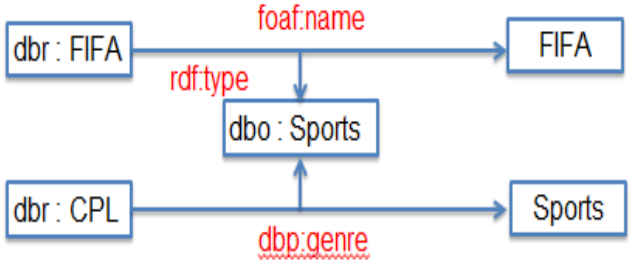

Fig4: RDF graph for interconnected data

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 03, March-2019

We can represent RDF in the form of TURTLE which stands Terse RDF Triple Language which defines the simple syntax for RDF. W3C standardized it in 2014, and it was published by Dave Beckett as a subset of Tim Berners Lee's Notation (N3) language. Literals are used in RDF graph which represents some data values, and these are encoded with strings. A literal without a type is called Plain Literal. A plain literal may have a language tag. A literal can be interpreted using data-types. Any literal without a datatype would be considered as same as string. There are some nodes in RDF which are not having any IRI with them which can be considered as Blank Nodes. The representation of these kinds of nodes is syntax dependent. In TURTLE it is represented by an underscore followed by ":" RDF/XML was revised in 2004 since its adoption was standardized by W3C in 1999. It then worked with XML tools to get a proper standard. Some of the tools which can be used in relevance to RDF can be Apache Jena, Mobi, FRED, Outdated-ARC RDF store, Outdated-adobe's XMP. A sample RDF graph can look like the following figure:

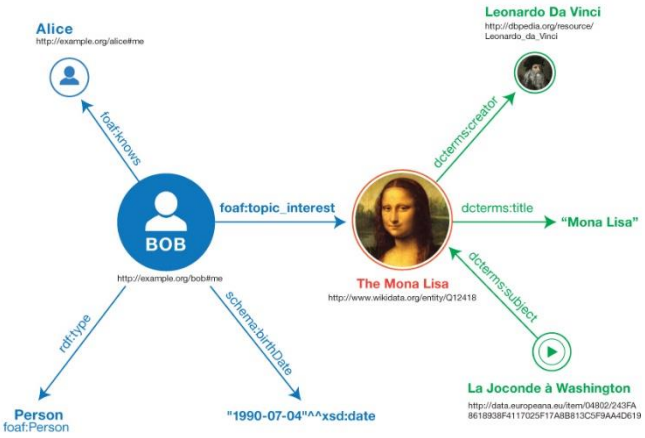

Fig5: RDF graph in terms of graphical representation

V. QUERYING RDF DATASETS WITH PROVENANCE

The Semantic Web depends on getting to and reusing RDF information from numerous different sources, which one may appoint different dimensions of authority and believability. Existing Semantic Web query languages, like SPARQL, have targeted the retrieval, combination, and reuse of facts, but have so far ignored all aspects of provenance, such as origins, authorship, recentness or certainty of data. In this chapter, we present an original, generic, formalized and implemented approach for managing many dimensions of provenance, like source, authorship, certainty, among others.

The approach re-uses existing RDF modeling possibilities to represent provenance.

Then, it extends SPARQL query processing in such a way that given a SPARQL query for data, one may request provenance without modifying the query properly. Thus it helps in querying the interconnected data with a highly flexible approach, and it's quite adaptable for relating it to the arenas of provenance. Provenance provides knowledge that can be used to quantify the value which can further help us to get indicators like where, why and whom, etc. Establishing relationships between knowledge and provenance requires appropriate mechanisms for supporting the statements about statements. This can be complex which would involve large chunks of data, but our main objective is to extract the information on original data with the help of queries. Our study deals with the data linked in www.dbpedia.org where the user can have direct access to the information and have direct access to the provenance where the user can design his queries for extracting the provenance based on his convenience. This allows SPARQL to fetch results with provenance with user intervention.

Moreover, provenance requires an extension of querying mechanisms which can be serialized into different levels and can be achieved with different application-specific interfaces. The syntax of RDF+ is based upon the building blocks of RDF only where:

- U is covering URI's
- L defines RDF literals.
- G covers graph names
- P covers the set of properties.

An RDF+ dataset D+ is referring to a set of literal statements and is further associated with provenance statements. The next challenging situation is to map both RDF and RDF+ to one another to define proper semantics to the datasets to avoid ambiguity and to refine granularity of the representation. The serialization of RDF+ data into RDF knowledge is straight forward. Now extraction of any knowledge needs to a query language to extract any information out of it which has been provided by SPARQL. In our case, we can access the SPARQL query engine via:

www.dbpedia.org/snorql

This gives us the interface to generate a query for the information available on DBpedia and to exploit the capabilities of SPARQL to derive every aspect of provenance in the data. It works on graph pattern matching and fetching the information from the ontology classes and functional properties. The standard SPARQL query is containing keywords like SELECT, CONSTRUCT, FROM, WHERE, LIMIT, FILTER which can help to eradicate the unwanted information. The keyword DISTINCT can be used to eliminate the duplicate entries in the table which again helps to hide the unwanted data. Before defining any query we need to declare prefixes and adding to it more emphasis can be laid upon various terminologies like FOAF (friend of a friend), RDF:type, RDF: property, rdfs:resource and many other attributes can be added upon in the query to fetch detailed information out of any dataset.

The pure design and flow of the RDF and SPARQL language have been explained in the below figure:

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 03, March-2019

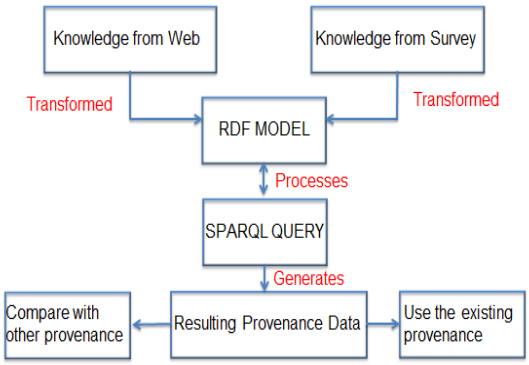

Fig6: Flow of the methodology

All the knowledge from the web and surveys is embedded into the RDF model with a defined framework and structure. Then comes into play the SPARQL engine with help to generate the query and it is being processed on the RDF model to extract the required information on it. Further the results of the provenance can be compared with the other provenance results, or it can be used with the existing provenance results.

Talking about the Provenance, there are few provenance attributes which can help us to get a clear picture of the derived information:

- was derivedFrom: It talks about the derivation of the information from where the information has been derived from.
- Was ended by Describes when an activity is deemed to have been ended by an entity. It may refer to a trigger that has ended or terminated by activity.
- WasGeneratedBy: It describes the generation of the entity by an activity.
- Was influenced by Describes how an entity or activity can influence that may affect the character, development or behavior of another source.
- Was informed by Describes the communication between two entities that how the information is getting exchanged between the two.
- WasInvalidatedBy: Describes the start of the destruction of an entity. Any entity which is generated is preceded by its invalidation.
- was quotedFrom: Describes the quotation of an entity such as image or text which may or may not be its original author.
- WasRevisionOf: Describes the revised version of the derivate entity of the resulting entity.
- WasStartedBy: Describes when an entity has been started by, and it did not exist before the start of this one.

VI. SEMANTIC WEB

The semantic web is an extension of World Wide Web which has been standardized by W3C (World Wide Web Consortium) which provides a mutual platform to have common data formats to make it easy for users to exchange data and it can reuse on various other platforms also. It contains the data in a very structured and tagged way which makes it more extensible. As the name suggests "semantic," this means adding the logic into anything to provide more formalism. Adding a degree of formalism to the web would help in making data more dimensional. The semantic network model was established in the 1960s by Alan M Collins to represent semantically structured knowledge, but the term "semantic web" was coined by Tim Berners Lee which helped further in adding more standards to it. Semantic web solutions were using three different languages, and those were: RDF, OWL (web ontology language) and XML (extensible Markup Language). These technologies were combined to make the web content more descriptive. It further helped to establish links in the open source data which was further coined as Linked Open Data (LOD's) which resulted in a giant global graph. Some of the challenges that are faced as of now in this are:

1. Vagueness
2. Vastness
3. Uncertainty
4. Inconsistency

The next solution which is being predicted can be in form Web3.0, and the semantic web is an essential component of it. Sometimes web3.0 can be used as a synonym for the semantic web. The following figure helps to define the stack diagram of the semantic web:

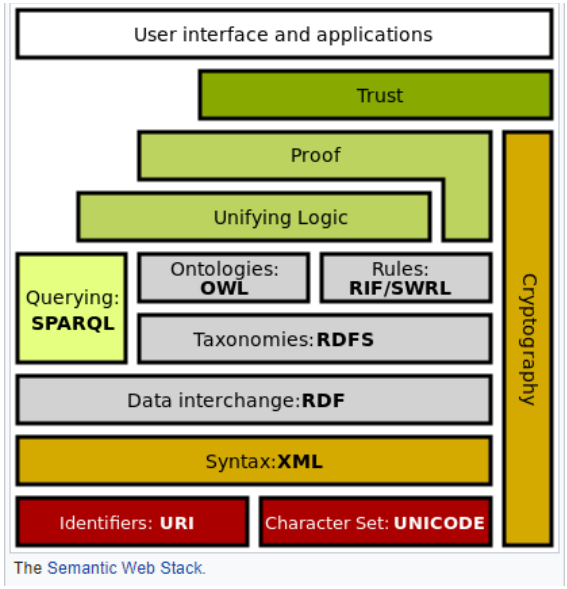

Fig7: Full stack diagram of the semantic web

The figure describes a certain set of elements which combine up together to form web3.0. The various component of the semantic web are:

1. RDF
2. RDF Schema
3. Simple knowledge Organization system
4. SPARQL
5. Notation3
6. TURTLE
7. OWL
8. Rule Interchange Format

The intent to create the entire component is to extend the use and usefulness of the web data and interconnecting all the web sources by creating semantic web services. It rather would help

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 03, March-2019

in dissolving the ambiguities in terminologies and would help in information retrieval. Adding on to it, this can help in decision making support in various ways.

VII. ONTOLOGY AND ITS SIGNIFICANCE IN ASPECTS OF PROVENANCE

OWL stands for Ontology Web Language which has been designed to fetch loud and complicated knowledge about the things that are available on the web. It is a computational logic-based language, and the knowledge is represented in a semantic web language. The current version of the OWL is OWL2.0. Ontology evolves into open, multi-user editing environments where users can create links to edit the information available and establish the connectivity between the data. All the ontology classes for DBpedia can be looked at the following link:
http://mappings.dbpedia.org/server/ontology/classes/

It contains 685 ontology classes and 2795 different functional properties. The DBpedia ontology is created based upon the information available within Wikipedia and with the release of DBpedia3.2; a new infobox extraction method based would be introduced. The DBpedia ontology contains 4,233,000 instances from over all resources. Some figures in the below table throw more light on the information available in the ontology:

Instances per class

Class	Instances
Resource (overall)	4,233,000
Place	735,000
Person	1,450,000
Work	411,000
Species	251,000
Organisation	241,000

In open settings, changes can be quite conflicting since users would be contributing to the knowledge in various ways at different points of time and two unwanted situations can arise:

1. Undesired inferences
2. Inconsistencies

To find the error, one has to debug the inferences made in the ontology, and it becomes necessary to question upon when, why and how? Since the provenance can be tracked in many dimensions like when was the last time data modified and certainly more. There are some annotations used in ontologies that may refer to as axioms. Such axioms cannot be answered by predefined algorithms which are already mentioned for debugging ontologies as it requires much expensive reasoning. With the approach presented in this paper, we will show how to represent provenance and efficiently reason in OWL with provenance. Our approach supports the user in coping with the complexity and dynamics of evolving ontologies.
Various approaches to the problem of debugging with provenance have been proposed. They can be grouped into three categories:

- Extensions of given logical formalisms that deal with a particular type of provenance. Examples include extensions for debugging with uncertainty, such as fuzzy and probabilistic
- Flexible extensions for systems allowing for algebraic query evaluation (e.g., as relational databases and SPARQL engines)
- A two-step evaluation for provenance, which is very expressive, but which does not assign a uniform semantics to the definition and composition of provenance in the knowledge base.

Debugging frameworks use on tree bases derivations to derive consistency for checking and querying. The process of finding explanations is used for finding inconsistency in the existing axioms which can be further referred to “PinPointing” which aims to cover up the provenance in all dimensions up to the full extent possible.

VIII. RESULT ANALYSIS

After understanding the whole mechanism of semantic web and ontologies, the familiarity was introduced in big subjects like RDF and SPARQL and yet we were able to determine the provenance with the help of SPARQL queries which helped in extracting the information out of the encyclopedia like Wikipedia and to study this we took a case study of 2 websites which were holding the information of Wikipedia in the form of tuples. Below mentioned are two sites which were available for the study:

1. www.dbpedia.org
2. www.wikidata.org

In these websites, we were able to run the SPARQL queries through SPARQL engines and check information on various elements and were able to cross verify the same to attain the proper provenance for the same. We learned the mechanism of how to flow of the whole model is working. It gives the immense power to check the information available in the ontology classes and helps us to understand the schema of all the classes and functional properties available in the ontology and how the concepts of mapping are done between the informational elements. Various models of provenance have been studied under this to understand the different architecture and functionalities of each model to check the provenance attributes of each model. With all the reasoning and analysis from different sources on the semantic web, applications need to track the various axioms of ontology classes and further use of pinpointing should be done necessarily to avoid ambiguity keep track of regular updates. This approach makes it highly scalable and acts as a building block of a web proof and trust layer. Under this paper, we studied various development tools and study tools like APACHE JENA, Mobi, etc which can help us to understand the concepts of provenance in terms of RDF and ontology. Moreover how data changes would be tracked in Linked open data sources to track the unwanted changes and ambiguities.

IX. CONCLUSION

This work demonstrates the need and necessity of provenance in the world of semantic web which can try to decay the redundant data which is meant to be inaccurate and spreads the

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 03, March-2019

misinformation about a certain thing. Provenance can play a big role in upgrading the data quality and data management in every dimension to make the base for a layer of trustworthiness and credibility because of the reliability of the data matters on the web. New approaches help in making kinds of data analysis for the quality assessment for the raw data available, and that makes it more extensible. We believe in publishing the correct provenance information which uses the correct quality parameters and provides valid provenance parameters which are required in measuring the degree of trustworthiness of data which would be adding up the quality of data. The whole process may involve querying RDF datasets with the help of SPARQL query language and may request for provenance out of the whole data available on the internet. This topic is still in its initial stages, and a huge potential resides in this topic which can eventually change the world of web. This dissertation motivates the web sources to publish more accurate information based upon their validity and quality supporting all the application needs. Ontology classes need to have more advanced algorithms for debugging and tracking the changes in them since it provides effective reasoning and logic to the data that prevails.

X. FUTURE RESEARCH WORK

Managing provenance itself can incur a huge cost and huge storage capacity since provenance information can sometimes be very large in comparison to the information itself, so provenance model has to be extended or changes has to be introduced to provide more abstraction to it. More optimizations are needed to track the ontology changes which would aim to add more formalism to it so that more formal structure can be introduced to the data. The effective use of metadata among applications requires more conventions about syntax, semantics, and structure. Provenance may be used by audits to establish accountability, but more work needs to be done to safeguard the information hub and enhance information handling policies.

ACKNOWLEDGMENT

I wish to express my sincere gratitude to Dr. Jinan Fiaidhi, Professor at Lakehead University for providing great assistance in the research. She also provided her expertise and wisdom on this topic of research and reviewers who provided great feedback for this research and helped me in improving my final reports.

REFERENCES

[1] Managing and Using Provenance in the Semantic Web - Renata Queiroz Dividino)
[2] www.dbpedia.org/snorql.
[3] www.wikidata.org.
[4] www.w3c.org.
[5] www.learningsparql.com.
[6] Provenance information in the web of data – Olaf Hartig