Knowledge Base Smarter Articulations for the Open Directory Project in a Sustainable Digital Ecosystem

Shastri L Nimmagadda
School of Information Systems
Curtin University, Perth, Australia
+61 8 9266 9780
shastri.nimmagadda@curtin.edu.au

Dengya Zhu
School of Information Systems
Curtin University, Perth, Australia
+61 8 9266 7056
D.Zhu@curtin.edu.au

Amit Rudra
School of Information Systems
Curtin University, Perth, Australia
+61 8 9266 7055
A.Rudra@curtin.edu.au

ABSTRACT

We examine the volumes and varieties of data sources of the Open Directory Project (ODP), which can endure, regenerate and flourish with new knowledge. The ODP motivates us in building a knowledge base smarter multidimensional data constructs and models. We articulate the models with new artefacts, addressing the heterogeneity and multidimensionality of the data. The conceptualization and contextualization of various entities and dimensions have emerged with innovation that led us to develop a digital ecosystem-based inventory. The ODP based domain ontologies support the warehouse repository, which accommodates multidimensional data relationships. The concept of a digital ecosystem in the ODP context is to bring the dimensions together and unite with multidimensional schemas. We explore the Big Data, incorporating their characteristics in the ODP constructs and models. The volumes and varieties of the ODP data are logically organized and integrated in the warehouse repositories. The multidimensional data modelling makes the ODP more smart and flexible in an environment, where varieties of business rules and constraints change rapidly. The visualization and interpretation are the other artefacts of the Big Data facilitating us use, reuse, test the interoperability and effectiveness of the data models for sustainable ODP digital ecosystem. We compute the polynomial regressions, based on the data fluctuations of the ODP as observed in the scatter plots, providing new data mining models for knowledge interpretation.

Categories and Subject Descriptors

E.1 [Data]: Data Structures; E.5 [Data]: Organization/Structure; H.2.2 [Database Management]: Physical Design; H.3.2 [Information Storage and Retrieval]: Information Storage, File Organization.

General Terms

Data Structures; Documentation; Design; Management; Performance and Standardization.

Keywords

The ODP; Digital Ecosystem; Multidimensional Data Modelling; Domain Ontologies.

WWW 2017 Companion, April 3-7, 2017, Perth, Australia.
ACM 978-1-4503-4914-7/17/04.
http://dx.doi.org/10.1145/3041021.3054769

1. INTRODUCTION

We develop the concept of a digital ecosystem, simulating the ODP framework. An ontology-based data warehousing and mining motivate us a mechanism for bringing a comprehensive, consistent, flexible and smart metadata together all in a single repository, encapsulated in a digital ecosystem. Managing the advancement of the human edited ODP [19, 29], with 91868 number of editors, 1, 031, 852 categories, 3, 871, 704 websites and 90 languages and the continuing effort of the web-based directory is a huge task. We need a more holistic and smart integrated framework with new data modelling artefacts. The generalization and specialization hierarchies [8, 24, 25] play roles on data relationships through various ontology descriptions. The domain ontologies further enable us the data integration process, formulating the integrated framework, in particular, the knowledge-based conceptualization and contextualization attributes as interpreted in various digital ecosystems. Additionally, keeping in view the current volumes and varieties of the ODP data, we exploit the use of Big Data concepts [1] in building knowledge-based constructs and models. The models are likely to deliver an efficient data mining and interpretation that can explore the connectivity in between the categories, sub-categories and their levels. We apply the statistical polynomial regression for establishing the models of data relationships between the categories, sub-categories and levels (web layers).

2. PROBLEM STATEMENT

In spite of major breakthroughs and advances in the internet technologies, identification and precise description of systems and their connectivity remain unresolved. This is partly due to poorly integrated multiple data sources and domains, in which the phenomenon of an ecosystem has not been readily descriptive. Heterogeneity and multidimensionality of data sources are the other major issues. The unstructured data complicate the concept identification, data integration and interpretation in different knowledge domains. Highly specialized data semantics [12, 14] make it infeasible to incorporate ideas within a consistent repository. The meaning of data is usually hard to define precisely [14, 16] because they are neither explicitly stated nor implicitly included in the database designs. An ontology description of an entity or a dimension is not a single, consistent scientific domain; it is composed of several dozens of smaller, focused research communities. It would not be a significant issue if researchers were able to access data from a single domain, but that is not usually the case. Typically, the researchers require data access from an integrated metadata of the ODP [6, 17], after resolving the terms that have different meanings and vocabularies across diverse communities or domains. The observations further complicate the metadata access [6, 11] at which the particular community whose terminology is being used by the data source, usually not explicitly identified and the terminology evolved over time. For many larger community data sources, the domain is obvious—the domain ontology handles the structured information, providing sequential information and useful annotation—but the terminology used may not be current and can reflect a combination of definitions from multiple domains. Though these challenges are inherent in the ODP, the new data integration approach exploits the common scientific domains and attributes, but not typically found elsewhere.

In general, the narration of a system in the ODP scale is elaborate, because of the existence of several categories, sub-categories, websites, web pages, documents and millions of words within the documents. Making up the systems in the ODP may possess a variety of attributes, and each attribute is characteristic in its representation, classification and conceptualization. The ODP is often interpreted as categorization and classification of a group of attributes in multiple domains [29]. The contextualization has significance, which has been ignored in several knowledge domains of a system. Each system comprises of group of data events, making up the system with hierarchies.

3. RESEARCH OBJECTIVES

We identify the heterogeneity and multidimensionality of the data in multiple domains. The ODP possess volumes and varieties of data, interpreted in multiple domains. For example, "Health, "Regional", "Society", "Science", "Computers" dimensions of the ODP are such closely related and interconnected domains. We intend to address:

1. *Simulate an integrated ODP framework:* Multidimensional ontologies are described. We describe dimensions from categories, sub-categories, levels of downloaded websites, the number of documents and associated textual information and words. We intend to develop a robust and holistic methodological framework, simulating the ODP.
2. *To share common understanding of the structure of information and knowledge:* It is one of the common goals in developing ontologies. Several websites contain information that provides e-commerce services and products. If the websites share and publish the same ontological descriptions of the entities, the computer agents can be able to extract and aggregate information from different sites smartly and tidily. The agents can use the aggregated information to answer user queries or as input data to other applications of the ODP.
3. *To enable reuse of domain knowledge:* Models in several domains of ODP need to represent the view of space and time. This representation includes the notions of time-intervals, points in time, relative measures of time, and so on. If one group of researchers develops such ontology in detail, others can reuse it in their other domains. As an example, the domain knowledge acquired from a particular model made from a category or sub-categories, may be reused in the same ecosystem in the other categories, sub categories and or levels. Additionally, if a large ontology needs to be built in the ODP scale, several existing ontologies that describe portions of the larger domain can be used or reused in the integration process. Data views extracted from warehoused metadata [15] are visualized in the new knowledge domains.

4. RESEARCH OBJECTIVES

The ODP occupies large geographic regions worldwide, possessing many geographic based domains or information systems. Understanding their connectivity is crucial in making useful knowledge, based on business alliances and decisions. The existence of volumes and varieties of data and their heterogeneity has motivated us to develop new ideas of ontology-based data warehousing and mining [18], in particular when multiple dimensions and their attributes exist with the ODP. An integrated framework that may lead to the development of a digital ecosystem evolves with a new knowledge-based digital solution. The digital solutions are in growing demand in the integrated business and project management environments, in spite of the complexity, dynamically changing business rules and constraints. The data integration and understanding the connectivity among multiple domains or information systems are paramount and key motivating factors of the current research. Volumes and varieties of historical data in various geographic regions have motivated us to undertake the current research. We have had extensive consultations with variety of producing and service companies worldwide. We have consulted more than 100 websites that involved with the ODP and its applications [24, 29]. Based on our experiences with various business situations and published sources, we construe various pitfalls and ambiguities in the knowledge representation of the categories and sub-categories in a smarter way, keeping in view the heterogeneity and multidimensionality of their data. The atomicity and granularity are additional features needed to bring from denormalized multidimensional ontologies [18] for minimizing the ambiguities in the interpretation of the data relationships.

The digital ecosystems and their embedded systems are described in the context of the ODP, demonstrating the necessity of ontology modelling in the integrated workflows and their implementation in multiple domain applications. The specification of conceptualization and contextualization modelling and integration of multidimensional and heterogeneous data sources in the ODP context are new visions. In our view, "Health", "Science", "Society", "Regional" and "World" dimensions cannot be isolated, which are otherwise embedded, demonstrating an ecosystem within which the dimensions inherited from their connectivity. For this purpose, an integrated methodology is proposed simulating the ODP and enable to understand the ecosystem phenomena through interconnected digital ecosystems. Several data sources exist within the ODP and applications. As a part of demonstrating the ODP digital ecosystem, we adopt an ontology-based data warehousing, simulating multiple systems. Super-type and sub-type dimensions are interpreted within the ODP based on the conceptualization and contextualization including generalization and specialization features. The ODP is a motivating platform that can take numerous dimensions together within the ecosystems' sustainable integrated framework.

4.1 The ODP as a Sustainable Digital Ecosystem

The sustainability is a capacity to endure, regenerate and flourish through a period [4]. It underlies with the fact of understanding the governance, cultural systems and how they interact with ecological systems so that they can better be structured and managed to produce science and knowledge for policy making purposes. In the ODP perspective, the sustainability is regarded as a composite attribute meaning thereby the sustainability-related problems cannot be adequately addressed from a single domain perspective. Whether it is that of one demography or one culture of human ecosystems or an environment, the ODP is put up with other coexistent ecosystems to generate values of knowledge or economic gains. Although the concept of "ODP sustainability" is evolving and dynamic, the ambiguity of its perception and description of "sustainability" [4], as well as the communion, connectivity and interaction events among multiple domains and their associated systems can explicitly be interpreted in the data knowledge and engineering perspective.

4.2 The Digital Ecosystem Conceptualization of the ODP

The elements and processes [18] of a system benefit from each other's participation via symbiotic relationships (positive sum associations) [26, 28], is termed as an ecosystem. The ODP, in which the constructs and models, several domains and systems described, is a complex community, but the environment in its functioning as a single ecological unit is still a mystery. More realistically, it is a term of millions of data attributes and properties from volumes of DBs all that store in one place. At this stage, the concept of an ecosystem is introduced, in which, we describe several entities and or dimensions. In ecosystem situations, all elements and processes continuously interact and communicate each other. In the context of a broader notion of the ODP, integration of the entities or dimensions fits with the view of data warehousing, the actual concept of metadata, a smart representation. In the current research, an attempt is made to acquire the data sources from different domains of the ODP [17, 23, 22] and integrate them using the concepts of data warehousing.

The ODP ecosystem refers to an interdependent group of natural entities with associated categories, existing in a particular environment and the habitat within which these categories and sub-categories interact based on web layers and or levels. The ecosystems sustain in the natural world, providing humans to live and thrive sustainably. As an example, an ecosystem is described as an element of the biosphere, which has purposeful mechanism needed to sustain itself. Due to significant interchange between ecosystems, they survive next to each other. They share material and energy when adjacent systems interact each other. If an ecosystem collapses, the surrounding system is affected, or it could take with it. When human-made ecosystems, such as urban ecosystems, croplands and farms are encompassed, in which case, the humans alter the natural balance of ecosystems. Analogous to freshwater ecosystems, oceanic and terrestrial systems are part of an ecosystem broadly, a collective entity, in which several elements and processes interact both geographically and periodically. We take advantage of the conceptualization in simulating the ODP framework in which several articulations accommodate in the form of applications [24, 29]. While describing multiple domains in ecosystems, we emphasize the characterization and description of data sources that are critical in the modelling and mapping process. Millions of records from thousands of attributes are in one repository, which is termed as a digital ecosystem. Each domain is characterized and categorized by several sub domains with sub-categories. Each category has sub-category. Here a hierarchy [23] of generalization, on a broad ODP (with multiple domains) to specialization is interpreted with sub-categories. It is inherently an ecosystem, a system whose members are hierarchically connected and communicating each other. Similar participation of relationships and or positively summed relationships may benefit each other, which may be referred to as a self-sustaining system or an overall system participation with attributes of the neighboring systems. We propose an integrated framework, simulated as multidimensional ODP with many components as various artefacts.

4.3 The Components of an Integrated Framework

Several artefacts articulated in the integrated framework are domain bound and data modelling, schema selection, data warehousing and mining, visualization and interpretation/knowledge-based models. These constructs and models are critical in addressing the heterogeneity and granularity of the multidimensional ODP logically.

As highlighted in Figure 1, broadly we describe the data acquisition, data modelling and information analysis stages in the proposed framework. In each description, how the acquired data can quality control, how the modelled data can organize the information that is processed and interpreted for further evaluation. It is worth mentioning that all the events, such as data acquisition, data modelling, and information analysis are interconnected in a way to achieve the integration and the connectivity process. The methodology is vital in connecting and integrating multiple ecosystems. Using the concepts of the ecosystem and embedded ecosystems, we aim at exploring the connectivity between systems and or domains.

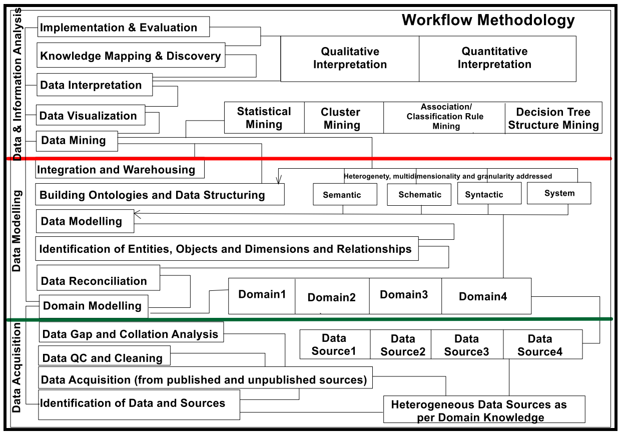
Workflow Methodology
Data & Information Analysis
Implementation & Evaluation
Knowledge Mapping & Discovery
Qualitative Interpretation
Quantitative Interpretation
Data Interpretation
Data Visualization
Statistical Mining
Cluster Mining
Association/ Classification Rule Mining
Decision Tree Structure Mining
Data Mining
Data Modelling
Integration and Warehousing
Heterogenety, multidimensionality and granularity addressed
Building Ontologies and Data Structuring
Semantic
Schematic
Syntactic
System
Data Modelling
Identification of Entities, Objects and Dimensions and Relationships
Data Reconciliation
Domain Modelling
Domain1
Domain2
Domain3
Domain4
Data Acquisition
Data Gap and Collation Analysis
Data Source1
Data Source2
Data Source3
Data Source4
Data QC and Cleaning
Data Acquisition (from published and unpublished sources)
Identification of Data and Sources
Heterogeneous Data Sources as per Domain Knowledge

Figure 1: An integrated methodological framework

The acquisition of new knowledge through conceptualization and contextualization surrounding the ODP is assessed for a holistic modelling methodology. Within the digital ecosystem scenarios, we emphasize that all the logical and interrelated data sources existing within a single and broad ecosystem (such as ODP) and their data events are made good use in the modelling process. The nomenclature and vocabularies associated with the content, the semantics of the dimensions and their associated attributes of the ODP events, are handled by ontology descriptions [27, 29]. For example, “Regional” “Health” and “Science” dimensions of the ODP have varied scope of coherent and inherent semantics [28], providing a space connecting to the related geospatial dimensions.

For the purpose of multidimensional ontology descriptions in the context of categories/dimensions of the ODP, we interpret the description of concepts in a domain of discourse (entities or dimensions sometimes called concepts). Each concept describes various features and attributes (slots, at times called roles), and restrictions on slots (facets, rules sometimes called constraints). An ontology constitutes a knowledge-based set of relationships among individual instances of entities or dimensions. Various dimensions are interpreted in categories and sub-categories in each domain of the ODP. A dimension can have sub-dimensions that represent concepts that are more specific than the super-type dimension. In real situations, there is a fine line where the ontology ends and the knowledge base starts. In this study, we focus on ontology constructs in different knowledge domains of the ODP. For the convenience of the nomenclature, we describe all the terms or events as dimensions in place of entities, since dimensions are the focus of most multidimensional ontology descriptions. Besides, we highlight the involvement of a programmer and ontology designer in perceiving the design aspects and requirements for categorization of dimensions and their levels in the ODP.

In the ODP, several categories are described with sub-categories and various web layers, in terms of levels. Integration of data events with multiple categories such as "Regional, "Society", "Business", "Health", "Science" and "Computer" dimensions is challenging. Connecting and integrating various such categories and sub-categories from local to global geographic dimensions are characteristic features of the ODP. We intend to demonstrate the data events of ODP, where large scale unstructured data sources need structuring and integration through ontological descriptions.

4.4 The Ontology Descriptions in the ODP Contexts

We equate ontologies in the present application scenarios with taxonomic [7] hierarchies of classes, class definitions, and class conceptualizations of relationships described among multiple dimensions. To specify conceptualizations, business rules and axiom constraints need to be committed during contextual interpretations of the conceptualizations. In the context of an integrated workflow, the concept of an ecosystem is benefited with several multi-disciplinary entities or dimensions or events participation in the integration process through conceptualized relationships (in other words through symbiotic relations, positive sum relationships). More realistically, it is a term of volume of attributes gathered from multiple sources (both geographical and periodic) all in one place. A similar analogy is applied in the broader and larger size of the ODP (global scale), with multiple systems with several hundreds of attributes connected to large size domains elsewhere at a global level, where coexistent data exist with no boundaries. As it applies to any business, an ecosystem in the case of a scheme, can be viewed as a system in which the relationships established across different dimensions represent new data events of the ODP can become mutually beneficial, self-sustaining and (somewhat) closed.

Several conceptualized data relationships [25, 27] exist among different entities and attributes to build the conceptual ontology models. Each data event is again composed of groups of other associated events making up the system, interpreting events in a particular knowledge domain and describing the leverage of human edited ODP. Each system, within a broader ODP context, is an information system. All the elements of the local system share their attributes and strengths with elements of other systems, extendable to ODP. Several data models are deduced representing the ontologies among ODP categories and sub-categories. Data structures are constructed in different star-schemas (Figure 2) using dimensions of categories, sub-categories and levels. Known multiple dimensions are logically structured in a way to understand the unknown domain knowledge. For example, the knowledge-based conceptualization and contextualization attributes have significance in understanding the knowledge of a particular system, its use and reuse for a period of time. Ontology models use the data events, interpreted as dimensions to connect to the factual data with one-to-many relationships as described in a star-schema in Figure 2. Even UML models [2] can also be used in building attribute relationships among ODP dimension data tables.

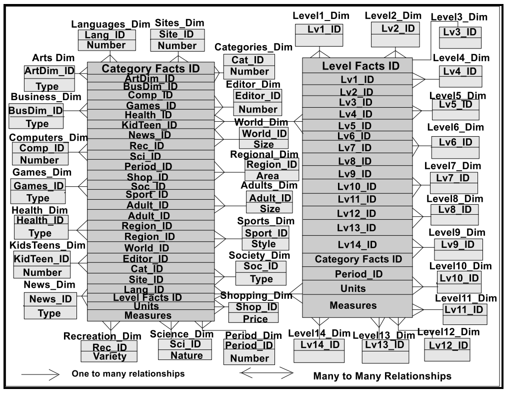

Figure 2: Star schema model for category and level attributes

5. INTEGRATED METHODOLOGICAL FRAMEWORK

As discussed in [6, 14, 19] star, snowflake and fact constellation schemas are open for constructing multidimensional logical data models. As suggested in [13] warehoused data are hierarchically structured in different knowledge domains. Figure 3 describes one of the initial hierarchical structural views. Several such hierarchies are described in [29] for the ODP. The categories and sub-categories are used in building the hierarchical ontologies among various attributes and their relationships.

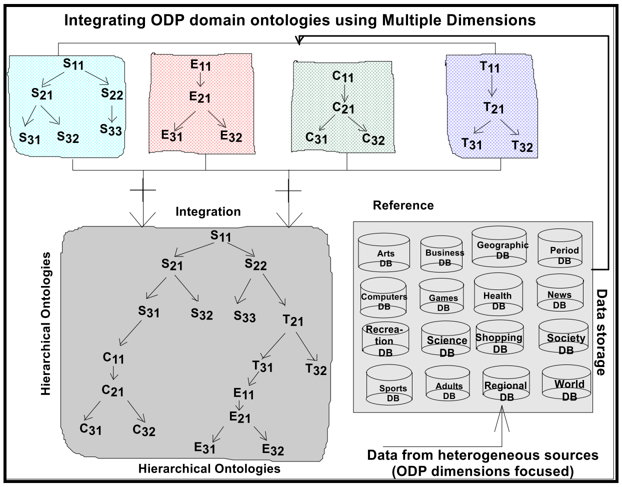

Figure 3: A hierarchical ontology view of the ODP

The ontologically structured data are warehoused through multidimensional structuring process. Hierarchical ontologies are intended with a goal of fine-grain multidimensional data structures. The process of integrated interpretation of the domain knowledge from fine-grained metadata is a significant measure of the suggested methodological framework.

The design of an integrated information system for an open directory depends on the individual design of conceptual schemas in multiple domain applications involving "entities/objects/dimensions". Integration of schemas belonging to various sub-systems is a requirement in the ODP to accomplish the legality and validity of data. Intelligent and expert data systems [22, 29] are used in domain applications. Ontology-based data modelling, data warehousing, mining, visualization and data interpretation, articulated all in combination in a single canvas in an integrated framework, are envisaged in developing the digital ecosystem in ODP context. Ontological structuring explores the connections among multiple domains that use for interpretation and evaluation of data events and the validity of categories and sub-categories of the ODP. Investigation and interpretation of digital data of a system or number of systems, existing within the ODP scale can lead up to a new digital information solution.

5.1 The ODP as a Digital Ecosystem Framework

The data warehouse (DW) approach brings together the systems' data from different categories and sub-categories to accommodate within ODP framework. The DW approach is used to benchmark, tracking the effectiveness of systems' productivity with both periodic and geographic dimensions. It also allows processing the data shared (knowledge-domain models) among information system professionals and the ODP experts worldwide. The need to integrate the data from multiple systems and sources is well known [1, 3, 8]. It has significance in the data warehouse design to define the scope, depth, comparability and accuracy of the data entering the warehouse repository. The range of data refers the level of the categories and sub-categories and the depth of information in the ODP. The depth of data indicates the level of details of the categories and sub-categories. To be comparable, data from multiple dimensions and websites should adopt the consistent classification as much as possible. No matter how differently data are collected across sites, they are significantly altered dynamically (based on the nature of data types) for integration before moving into the data warehouse. To reduce the burden of alteration, systems analysts and data modellers make vital changes in the models, making compatible to software systems to send metadata [20] to the repositories. It is also imperative to standardize the data collection processes. The accuracy of data is paramount in all types of data (that undergo intelligent storage in the global ODP schema) in any given situation which is a fundamental requirement of veracity and documentation of the data.

We populate the key fact instances and dimensions in tables so as to organize the relationships among multidimensional models through their common attribute instances and simulate the interaction among multiple dimensions. The facts may be dimensions and dimensions may also be the facts. Relationships identified based on logical concepts and contexts may also be the facts and or dimensions. In ecosystems settings, there are several tools, procedures and processes to connect dimensions and integrate their ontology descriptions as simulated in a framework in Figure 4. These facts are characterized in the ODP as global digital ecosystem conceptualization, at which several categories and sub-categories at various web layers and levels constantly interact and communicate, sharing resources among each other. The attributes in the categories and sub-categories are interrelated each other so that if one element is missing in one sub-system, another feature is shared from the other sub-systems of the ODP.

Several facts and dimensions are organized and structured in a way to get the knowledge of interconnectivity among the sub-systems. The data dimensions are modelled in different star-schemas. In these data schemas, there are multiple dimensions narrated and data relationships interpreted conceptually (logical data organization) among several dimensions [14, 20] and fact data tables, physically organized, as shown in Figure 4.

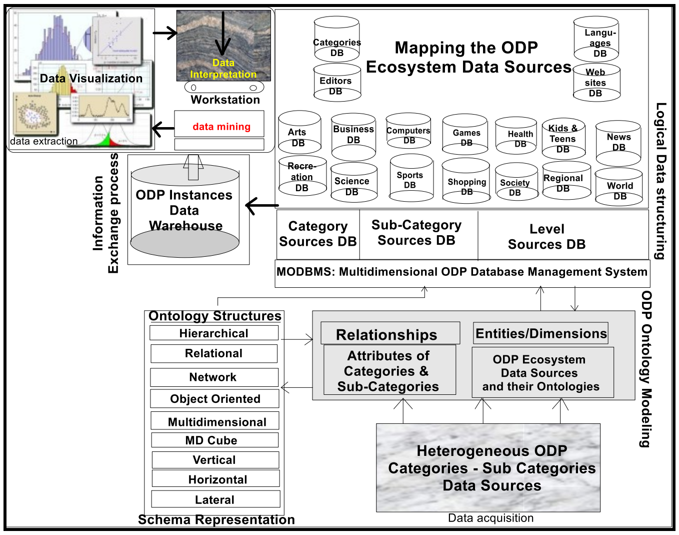

Figure 4: An integrated framework describing the categories and modelling of their relationship attributes

It is significant to design intelligently/logically the attributes among the multiple dimensions in a way to understand or increase the domain knowledge that is unknown. The Oracle database procedures are used for mapping the ecosystem models. The ontology models are constructed for systems, in which several dimensions are described for connecting the factual data in one-to-one, one-to-many and many-to-many relationships [10, 16, 17].

The "Regional" and "World" data facts and their associated dimensions, for example, play roles in the ODP ecosystem situations in integrating different data from various systems and their sub-systems. These facts can better be corroborated with actual data acquired in different categories and domains [17, 29]. The "Regional", "World", "Science" and "Health" data facts also depend on other facts of associated either super-type and or sub-type dimensions of the other domains. An ontology model is described for building the relationship of the facts and dimensions. The models built based on the relationship facts are used for extracting and mining the data views of the knowledge-based interpretations. The phenomena of interconnectivity establish through ontological conceptualization and contextualization. A multidimensional model drawn with location hierarchies can ontologically interact with the models that have relations with other ecosystems [18, 27]. The warehouse schema has different data sources of the same type of data, but common in the semantic framework [12]. If there is no common attribute information or data property among data dimensions, different DBs may have survived side by side within a warehouse environment, without merging the data cubes (Figure 5). Based on common data property information that exists in data structures, DBs get merged. The ontology in agreed domains such as “Regional”, “Health”, “Science”, is created and used as a basis for specification and development of data warehouse schemas or even using unified modelling language (UML) models [3] of various categories and sub-categories. The benefits of the approach include documentation, maintenance, reliability and knowledge use and reuse during data mining and interpretation stages. It is a much smarter way of presentation with sustainable storage of ODP data attributes.

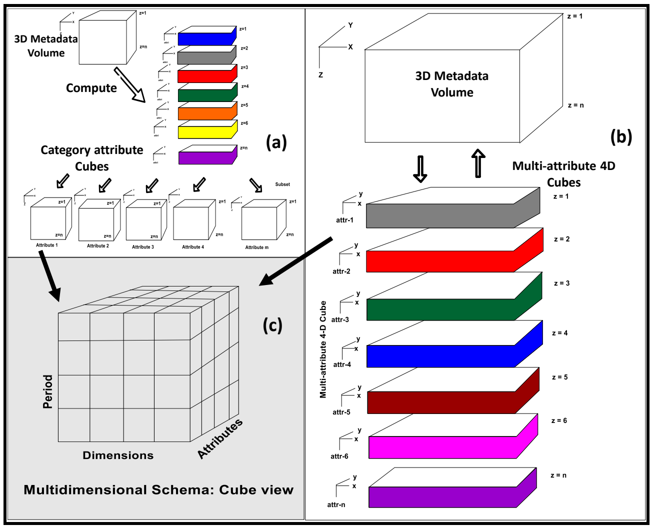

Figure 5: Smarter presentation of the attributes for variety of categories and sub-categories of the ODP

In the present work, we import the data from various categories and sub-categories of the ODP [6, 29] using programming applications such as MS Excel and Access. The data cleaning and loading of data are key operations done before the data arrived in storage devices. The data models are further ready in the form of relational data structures to represent in an Oracle environment. We run the SQL queries to obtain data views for interpretation and analysis. The database has been populated with existing data quickly using text, ASCII, SQL and control files (data loaders). We use various grapher and surfer programs and tools for visualization and data interpretation.

6. ANALYSIS AND DISCUSSIONS

The data schemas and attribute cube views are simple structures, flexible enough to modify smartly as per the knowledge representations and users’ requirements. If the data schemas are too large, complexity is a significant barrier to widespread adoption of the warehouse technology, because users may find the schema so difficult to understand that they may be unable to write queries and application programs. But the schemas are evolved to grow and support the new data types. Limiting the scalability to more data sources though has a definite advantage, but in the ODP context, the data structures are often large with hundreds of attribute variables. The data views are analysed for knowledge discovery, evaluating the veracity of data relationships from ODP metadata [6]. In such cases, cubes and cuboid schema views simplify smartly in representing the data with more details within small storage spaces. In our analysis, the data relationships among attributes variables are not linear. When plotted with the number of categories versus the number of documents or words attributes (downloaded from the ODP), we visually find scattered observations on the scatter plots, in which case, we interpret a curvilinear data relationship. For curvilinear trends, the polynomials are appropriate for fitting the observed data. We use the orthogonal polynomial regression in place of polynomial regression between various levels, from which the documents downloaded in different categories of the ODP possess different features in different web layers. At places, the data instances fluctuate around the polynomial trend line. The orthogonal polynomial regression is appropriate and at times necessary for higher order polynomial fits, if we need to explore the deeper level of knowledge from the category of websites of the ODP. If the resulting polynomial coefficients are large, relatively small Y values cannot be accurately calculated. We use orthogonal polynomial regression, because the polynomial regression oscillates excessively or when Y values calculated from the polynomial equation are not approximating the fit curve closely enough.

The orthogonal polynomial regression statistics [5, 21] contain standard statistics such as fit equations with polynomial degrees (change with fit plot properties). We use data instances for various categories and subcategories of the ODP [5, 19] with statistics specific to the orthogonal polynomial such as B[n], Alpha[n], and Beta[n]. Since this is an orthogonal method of calculating the polynomial regressions, each degree’s orthogonal polynomial factors are independent of each other. The degree zero results are the optimal zero order fit; the degree one results are the optimal first order fit, and so on. Adding more degrees to the fit does not change the previous degrees’ orthogonal polynomial factors. For example, if a fourth level of the orthogonal polynomial regression is calculated from a data set and a separate eighth-degree orthogonal polynomial regression can be determined [5] from the same data set, the orthogonal polynomial factors remain the same for degrees zero through four in both statistics results. The Polynomial regression fits a curve based on the equation as described in Eq.1. The polynomial degree can be set from zero to 10. A polynomial degree of zero is the average Y value, degree one is a linear fit, degree two is a quadratic fit, degree three is a cubic fit, and degree four is a quadric fit.

$$\boxed{Y = a_0 + a_1 X^1 + a_2 X^2 + \cdots + a_n X^n} \quad \text{Eq.1}$$

The Orthogonal polynomial regression fit is an alternate method of calculating the polynomial regressions. The Orthogonal polynomial equation is converted to "normal" polynomial form so, Y can be calculated from a given X with the equation. There are two options for calculating Y from a given X with the orthogonal polynomial regression statistics. The simplistic approach is to use the equation provided in the fit statistics. The orthogonal polynomial factors have been converted to polynomial regression equation (with coefficients) from which Y is calculated from X. Alternatively, Y can be calculated from X by using the orthogonal factors: X shift, X scale, B[k], a[k], and b[k].

$$\boxed{Y_{[k]} = B_{[k]} + XY_{[k+1]} - \alpha_{[k+1]}Y_{[k+1]} - \beta_{[k+1]}Y_{[k+2]}} \quad \text{Eq. 2}$$

Where X = XScale*(X-XShift); k = n, n-1,...,0 where: n is the polynomial degree; Y[n+1] and Y[n+2] = 0 where: n is the polynomial degree. The original X value is scaled before using it in the equation (Eq.2). The highest order equation is calculated first, then the results from that equation are used in the next lower order equation, and so on until the zero degree equation is solved. In the context of the ODP, we compute the polynomial fitting between the number of categories (dimensions) vs the number of documents, average lengths of levels, maximum lengths of levels and the number of documents attributes from the ODP. The number of sub-category attributes is plotted for each it's corresponding category as shown in Figure 6(a) and the categories "Arts" and "Regional" display higher number of sub-categories more logically. As presented in Figure 6(b), we plot the number of categories and the number of document attributes for each category and level 2, level 3 and level 4 attributes.

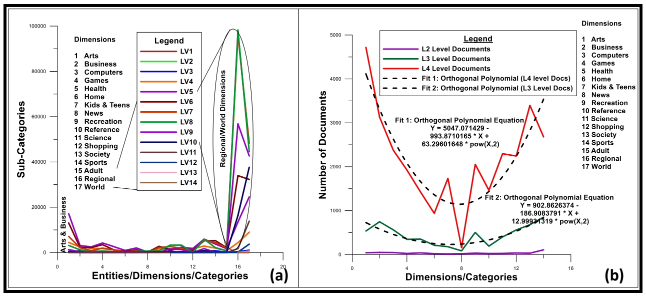

Figure 6: (a) A scatter plot between categories and sub-categories (b) scatter plots drawn for L2, L3 and L4 levels with ODP categories' polynomial regression models

We find scatter plots as shown in Figures 6 and 7, with fluctuations among the downloaded ODP document data, with smoothly fitting polynomial regressions.

The level 4 (or web layer) provides much narrower orthogonal polynomial regression, suggesting deeper and smarter knowledge of the ODP compared with level 2, which has broader polynomial regression, signifying a shallow knowledge of the OPD at level 2 or (web-layer).

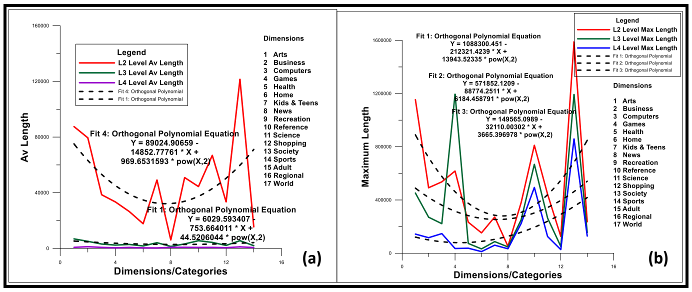

Figure 7: (a) ODP Categories Polynomial Regression Model with respect to Average and (b) Maximum Lengths

The average and maximum lengths are plotted for different categories/dimensions as shown in Figure 7. In both cases, the orthogonal polynomials fit well with the observed data of average lengths. From the polynomials computed for three different L2, L3 and L4 levels, we observe broad orthogonal trends for Level 4 and a narrower trend as observed for L2 layer. In the maximum length plot, L4 layer has a general trend, and it is narrower for layer L2.

As shown in Figure 8, a scatter plot is made with the orthogonal polynomial fit. The observations are plotted in between the number of categories and the number of words (of the downloaded documents). A steep and narrow curvilinear polynomial is observed with data fluctuations from the number of words. More words are reported from documents of "Arts" and "Regional" categories. Least number of words are documented for "News" category. The T IV sub-categories [26] and categories are plotted for maximum, minimum, µ and σ attributes. The peaks and troughs of the trends are different for different categories. The categories/dimensions from 4 to 8 ("Games" to "News") have troughs, the "Business and "Computers" have peak values of T IV sub-categories.

The sub-category in the 14/17 top level ODP categories representation appears symmetrical as is evident in Figure 9(a). The number of categories and the number of layers are plotted as shown in Figure 9 (a), in which the mean, median and mode appear equivalent because of the symmetry and normal distribution. As shown in Figure 9 (b), the number of levels attribute is plotted for maximum (Max), minimum (Min), mean (µ) and standard deviation (σ) ratios. The Max/Min ratio instance interestingly is higher at lower levels (or web layers) of the ODP and µ/σ is lower at higher levels of the ODP.

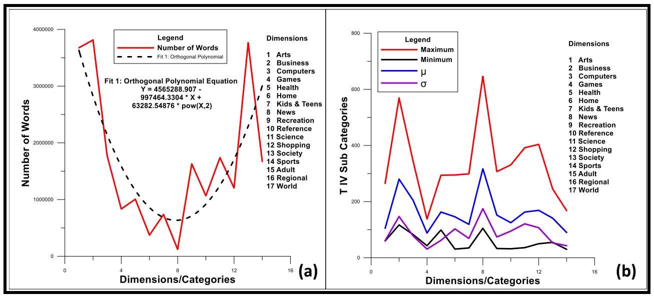

Figure 8: (a) ODP Categories Polynomial Regression Model 3 (b) "Max/Min" and "µ/σ" attributes

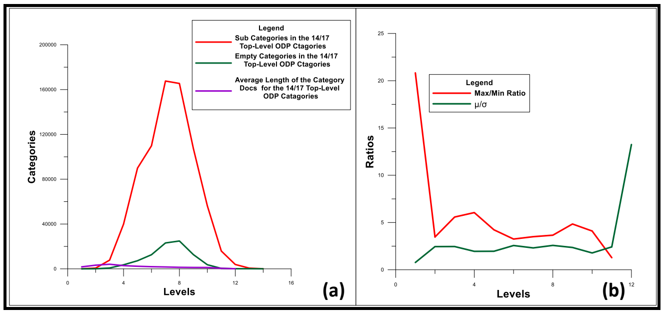

Figure 9: (a) Scatter plot between "levels" and "Categories" and (b) "Max/Min" and "µ/σ" attributes

7. THE CONCLUSIONS

The current research embodies the results of an integrated framework, simulated from the most comprehensive human-edited web directory, the Open Directory Project (ODP), using the multidimensional ontology descriptions and their logical constructs and models. Several categories, sub-categories and levels are interpreted as dimensions with various attribute instances and facts in the modelling process. The purpose of the paper is to provide an integrated framework accommodating domain ontologies involving the characteristics of the categories/dimensions, sub-categories, various levels, documents and words associated with the ODP. The ODP website users can share, use, reuse the multidimensional ontology structures including the aggregated classified information of user queries more smartly. The ODP new domain knowledge and categorization of the textual data are analysed for visualization and interpretation. The orthogonal polynomial models are computed in between various attribute dimensions of the ODP. These attributes include the number of categories, the number of sub-categories, and the number of levels including the number of downloadable documents and their associated words to establish an efficacy of the polynomial regression approach in representing the new knowledge hidden in larger levels or web layers of the ODP.

The measures of variation in the attributes of the ODP data are analyzed in terms of *mew, sigma, maximum* and *minimum* features. The category "News" has the least sub-category, downloaded documents and words. Whereas the sub-categories, the number of documents and number of words, downloaded from categories "Arts" and "Regional" are substantially high. We find the ODP as the most comprehensive and widely distributed warehoused metadata. There is an immense scope of the sustainable integrated framework in terms of a broad range of web information retrieval, web mining, semantic, schematic and syntactic analysis from smart constructs and models of the ODP. The digital ecosystem conceptualization and contextualization facilitate in managing the ever-increasing volume, the variety of information and knowledge obtained from the ODP sustainably and smartly.

8. REFERENCES

[1] Agrawal D., Bernstein P., Bertino E., Davidson S., Dayal U., Franklin M. and Widom J. (2012). Challenges and Opportunities with Big Data: A white paper prepared for the Computing Community Consortium committee of the Computing Research Association. http://cra.org/ccc/resources/ccc-led-whitepapers/.

[2] Berson, A. and Smith, J. S. (2004), "Data warehousing, data mining & OLAP", Mc Graw – Hill Education (India) Pty Ltd, pp. 205-219, 221-513.

[3] Bennet, S. Mc Robb, S. and Farmer, R. (2010), Object-Oriented Systems Analysis and Design using UML, 4th Edition, McGraw Hill, UK, p. 16-137 and 363-445.

[4] Burke, G.R. (2013), Making Viability Sustainable, PhD Thesis, Department of Humanities, Curtin University, Perth, WA, Australia

[5] Chen R. Polynomial Regression Models, Chapter 7, Institute of Statistics, National University of Kaohsiung, Taiwan, www.stat.nuk.edu.tw/Ray-Bing/regression/regression/Chapter7.ppt

[6] Chirita, P. Nejdl, W. Paiu, R. and Kohlschuetter. C., (2005), Using ODP Metadata to Personalize Search, In SIGIR'2005.

[7] Christophi, C. Zeinalipour-Yazti, D. Dikaiakos, M. D. and Georgios Paliouras, G. (2007), Automatically Annotating the ODP Web Taxonomy, 11th Panhellenic Conference in Informatics, Web Search and Mining – Information Retrieval, p. 397-407.

[8] Coronel, C., Morris, S., and Rob, P. (2011), Database Systems, Design, Implementation and Management, Course Technology, Cengage Learning, USA.

[9] Damiani, E. (2008), Key note address on 'Digital Ecosystems: the next Generation of Service Oriented Internet", IEEE-DEST, Phitsanulok, Thailand, Feb 2008.

[10] Deitel, H.M. and Deitel, P.J. (2001), "C++ How to Program (Introducing Object-Oriented Design with the UML)", Upper Saddle River Publishers, 3rd Edition, Prentice Hall Publishers, NJ, USA.

[11] Ding, Y., and Fensel, D. (2001), Ontology library systems: the key for successful ontology reuse. Proceedings of the first Semantic Web Working Symposium, Stanford, CA, USA. August.

[12] Maguitman, A.G. Menczer, F. Roinestad, H. Vespignani, A. (2005), Algorithmic Detection of Semantic Similarity, *WWW 2005*, May 1014, 2005, Chiba, Japan. ACM 1595930469/05/0005

[13] Marakas, M. G. (2003), "Modern Data Warehousing, Mining, and Visualization Core Concepts", Prentice Hall Pub.

[14] Meersman, R.A. (2004), Foundations, implementations and applications of web semantics, parts 1, 2, 3, lectures at School of Information Systems, CBS, Curtin University, Australia.

[15] Meng, W. Yu, C. and Liu, K.L. (2002), Building efficient and effective metasearch engines. ACM Computing Surveys, 34(1):48–89, March 2002.

[16] Moody, L. D and Kortink, M.A.R. (2003), From ER Models to Dimensional Models: Bridging the gap between OLTP and OLAP Design, Part1 and Part 2, *Journal of Business Intelligence*, Summer Fall editions, Vol. 8(3), http://www.tdwi.org

[17] Nimmagadda, S. L. and Dreher, H. (2007), DESIGN OF PETROLEUM COMPANY'S METADATA AND AN EFFECTIVE KNOWLEDGE MAPPING METHODOLOGY, a paper presented and published in the proceedings of *IASTED* conference, held in Cambridge in USA.

[18] Nimmagadda, S. L. and Dreher, H. (2012), On new emerging concepts of Petroleum Digital Ecosystem, *Journal Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery,* 2012, 2 (6): 457–475 doi: 10.1002/widm.1070.

[19] Perugini, S. (2008), Symbolic links in the Open Directory Project. Inform. Process. Manag. 44 910-930.

[20] Plastria, F. Bruyne, S. D. and Carrizosa, E. (2008), Dimensionality reduction for classification: Comparison of techniques and dimension choices, published *in the 4th International Conference, ADMA 2008,* Chengdu, China.

[21] Pujari, A.K. (2002), "Data mining techniques", University Press (India) Pty Limited, Hyderabad, India.

[22] Qiu, G. Liu, K. Bu, J. Chen, C. and Kang, Z. (2007), Quantify Query Ambiguity using ODP Metadata, *SIGIR'07,* July 23–27, 2007, Amsterdam, The Netherlands. ACM 978-1-59593-597-7/07/0007.

[23] Reddy, V. S. and Chaudhary, A. B.D. (2006), Hierarchy of Search Engines based on ODP Concepts, Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence (WI 2006 Main Conference Proceedings) (WI'06), 0-7695-2747-7/06.

[24] Ševa, J., Schatten, M. and Grd, P. (2015), Expert Systems with Applications, Expert Systems with Applications 42 (2015) 6306–6314.

[25] Shanks, G., Tansley, E. and Weber, R. (2004), Representing composites in conceptual modelling, Communications of the ACM, Vol. 47 (7), pp. 77-80, ACM, NY, USA.

[26] Thomas, J.L, Yannick, P. Valerie, W., Gupta, P. Stringer-Calvert, D.WJ, Tenenbaum, J.D and Karp, P.D Biowarehouse, P.D. (2006), a bioinformatics database warehouse toolkit; *BMC Bioinformatics*, 7:170, p.1-14, UK; http:///www.biomedcentral.com/1471-2105/7/170

[27] Wand, Y. (2000), An ontological analysis of the relationship construct in conceptual modelling, ACM Transactions on Database Systems, Vol. 24 (4), pp. 494-528.

[28] Wu, B., Davison, B.D. (2006), Detecting Semantic Cloaking on the Web. In: Proceedings of the 15th international conference on World Wide Web (WWW 2006), 819-828. ACM Press, (2006).

[29] Zhu, D. and Dreher, H. (2010), Characteristics and Uses of Labeled Datasets – ODP Case Study, 2010 Sixth International Conference on Semantics, Knowledge and Grids, DOI 10.1109/SKG.2010.84, IEEE Electronic Society, USA.