

Towards Semantically Aggregating Indian Open Government Data from data.gov.in

Asha Subramanian, Anmol Garg, Omang Poddar, and Srinath Srinivasa

International Institute of Information Technology, 26/C, Hosur Rd, Electronics City Phase 1,
Electronic City, Bengaluru, Karnataka 560100
{asha.subramanian, anmol.garg, omang.poddar}@iiitb.org,
sri@iiitb.ac.in

Abstract. Knowledge representation of “open data” involves aggregation of disparate information in a semantically meaningful context. This task is challenging as such datasets are arbitrarily structured and fragmented with no overarching contextual framework in which the datasets are uploaded. The utility of such datasets is determined by the “context” in which they are presented and the same dataset can be viewed and consumed in various contexts depending on the consumer. We present open data from data.gov.in in ‘Many Worlds on a Frame (MWF)’ - a framework where knowledge is organized within one or more thematic worlds each of which in turn relate to one another to form the global knowledge frame.

Keywords: Open Government Data, Linked Open Data Cloud, Semantic Integration, Knowledge Aggregation

1 Introduction

A large chunk of open data is made available through open government initiatives such as data.gov¹, data.gov.in² etc mostly in the form of CSV files. Since open data is generated with no pre-conceived data models, there is no overarching data model that can be used for integrating such datasets. This is a non-trivial task and we call such problems as *divergent aggregation* problems. The semantic integration and aggregation process involves extracting the various contexts or themes along which these datasets can be integrated and representing the semantic integration in an appropriate framework that not only identifies different perspectives from which the data can be aggregated, but also depicts how the perspectives can be inter-related. In this paper we present a knowledge aggregation application using Many Worlds on a Frame (MWF) that allows for rich representation of data across two aspects, namely, the type hierarchy (is-a) relationship and the containment hierarchy (is-in) relationship supported by associations to transform the open datasets into a web of semantically interlinked themes and their associations.

To the best of our knowledge, our work is the first to enhance and extend the usage of LOD³ to Indian Open Government Data.

¹ U.S. Government’s open data: <https://www.data.gov/>

² Open Government Data (OGD) Platform India: <https://data.gov.in/>

³ The Linking Open Data cloud diagram: <http://lod-cloud.net/>

2 Many Worlds on a Frame (MWF)

Many Worlds on a Frame (MWF) is an intuitive knowledge representation framework loosely modeled on *Kripke* semantics⁴. It allows for facts to be represented, grouped and related across many inter-connected worlds. Each world is considered a concept and concepts are organised in hierarchies, represented as rooted, acyclic graphs. Every concept belongs to two hierarchies - '**is-a**' or concept hierarchy and '**is-in**' or containment hierarchy. a 'is-a' b denotes a 'is a kind of' b and a 'is-in' b denotes a 'is contained in' b . The concept hierarchy is used to inherit properties and associations and the containment hierarchy is used to manage visibility. The root of the concept hierarchy is a concept called **Concept**, and the root of the containment hierarchy is a concept called **Universe of Discourse (UoD)**. A concept that cannot be subclassed using the 'is-a' relation, is called an **Instance** or a **Record**. Only 'instance worlds' store data, while 'context worlds' or 'class worlds' only manage structure and relationships. Each concept in a MWF system acts as a local 'context world' and hosts a set of knowledge fragments in the form of **associations** across concepts. 'Class worlds' can be imported into other worlds, so that their instances can participate as data elements. **Associations** are triples of the form (source, predicate, target). Here source and target are concepts in some target world say C_w and predicate is a label describing the association. In any association contained in world C_w , if the target concept is the world C_w itself, such associations are called **Roles**. The source concept is said to be playing a role defined by the predicate label in C_w . Roles, Associations and Worlds can be associated with zero or more attributes. An attribute is of the form (*Key, Value*), where *Key* is the name of attribute and *Value* holds the information regarding the attribute. Further, *Value* can hold literal data or a basic 'type'. 'type' can be 'String', 'URL', 'Date' or a world. When a world is subclassed by another world, all the roles, associations and attributes are inherited by the sub classed world.

3 Semantic Knowledge Aggregation of Open Data in MWF

A separate model generates **Thematic** and **Schematic** integration outputs given a collection of open data using heuristic algorithms over LOD [1]. This model generates a set of dominant classes or themes (output of Thematic integration) that best explain the 'context' of the datasets. The Schematic integration generates for each table in the collection, anchoring column(s) or subject column(s) that associate with the themes generated in the Thematic integration and the relations of the anchoring column(s) with the other columns of the table. Thus the tuples (Anchoring column, Relation, Connected column) provide complete semantics for each table using the themes that explain the collection. The themes and relations are classes and properties from LOD respectively. We use three tables from data.gov.in to explain the semantic aggregation in MWF namely - *AgmarkRice2012.csv*, *NutrientContent.csv* and *IndianStates.csv*. These datasets contain market-wise rice prices in various Indian states and districts, nutrient content against various parameters in Indian food crops and geographical information regarding various Indian states respectively. Here, 'Yago/YagoPermanentlyLocatedEntity' and 'dbo:Food' are themes produced by the 'Thematic integration' process depicting

⁴ Kripke Semantics: <https://plato.stanford.edu/entries/logic-modal/>

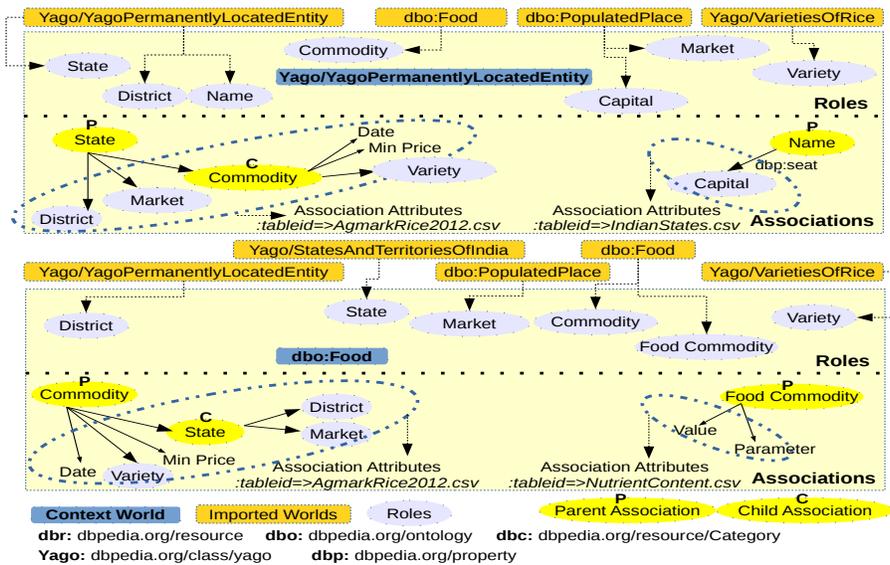


Fig. 1: Illustration of the working model of Many Worlds in a Frame (MWF) using open data tables

most pertinent contexts for the collection of tables (*AgmarkRice2012.csv*, *NutrientContent.csv* and *IndianStates.csv*). These themes translate into 'context worlds' in MWF. We have illustrated the components of these two 'context worlds' in Fig. 1. Note that table *AgmarkRice2012.csv* constitutes a complex subject determined by the columns - *State* and *Commodity* and has been consumed in two contexts. The context 'dbo:Food' shows the commodities that the various states sold while the context 'YagoPermanentlyLocatedEntity' shows the same table from the states perspective. The parent and child **associations** depicted by (*State*, *Commodity*) and (*Commodity*, *State*) in their respective contexts hold the complete semantics of the table *AgmarkRice2012.csv*. Similarly, table *IndianStates.csv* using the association *Name*, associates with the context 'Yago/YagoPermanentlyLocatedEntity'. The table *IndianStates.csv* constitutes a simple subject with the anchoring column *Name* explaining all the columns of this table. The semantics of this table is explained by the association *Name*.

4 Datasets and Demonstration

Currently, approximately 100 datasets from data.gov.in from various sectors such as 'Agriculture', 'Health and Family Welfare', 'Environment' etc have been aggregated using MWF. In the demonstration⁵, we will present "Sandesh" - the semantic data mesh of Indian Open Government Data. "Sandesh" seamlessly integrates the outputs from the 'Thematic and Schematic integration' model and populates MWF, given a collection of open data csv files. The demonstration is currently set up on a server with external IP and is powered by a SQLite database. During the demonstration, the implementation of

⁵ Demo: <http://wsl.iiitb.ac.in/sandesh-web>
 Video: <https://www.youtube.com/watch?v=pt1j2k1M97o>

MWF using datasets from data.gov.in will be presented using the ‘context worlds’ and ‘instance worlds’ that have been inferred from the datasets. Figure 1 and section 3 explain the MWF implementation in detail using a concrete example, that will be used during the demonstration.

5 Conclusion

In this paper, we have presented a rich and versatile application, that achieves semantic knowledge aggregation of open data. MWF intuitively represents the datasets as a semantic data mesh of interconnected worlds, roles, associations and attributes. The information from each table is represented in as many contexts as are applicable using ‘context worlds’. Information within each ‘context world’ is coherent and captures the various facets of an entity applicable within the boundary of that world. One can also traverse across inter-related worlds from a ‘context world’. We aim to integrate a reasoning engine in MWF to incorporate rules and reason new facts.

Other similar semantic integration efforts on open data include [2], [3], [4], [5] to name a few. Our semantic knowledge aggregation efforts largely differs from the cited work in that - the ‘theme identification’ is central to semantic integration. This form of knowledge representation allows data from multiple files or resources to be integrated using different contexts they represent. Inter-related contexts allows traversing through underlying resources in a seamless fashion. Other efforts to link government data use vocabularies to link metadata and provenance information regarding the datasets ([2]) or a custom vocabulary specifically meant to represent open government data ([5]). However, our model focusses on the subject or the context of the datasets to link related information. Our model currently presents aggregated contexts from multiple datasets, simultaneously representing a dataset in multiple perspectives. We are also able to provide a comprehensive picture of each ‘context world’ (a class or a concept in LOD) and how it relates to various tables from a collection of open data tables.

References

1. Subramanian, Asha, Ved Kurien Mathai, Vikkurthi Manikanta, Janaki Vinesh Joshi, and Srinath Srinivasa. "Semantic Integration of Open-Data Tables." In OTM Confederated International Conferences "On the Move to Meaningful Internet Systems", pp. 589-607. Springer International Publishing, 2016.
2. Ding, Li, Vassilios Peristeras, and Michael Hausenblas. "Linked open government data [Guest editors' introduction]." IEEE Intelligent Systems 27, no. 3 (2012): 11-15.
3. Böhm, Christoph, Markus Freitag, Arvid Heise, Claudia Lehmann, Andrina Mascher, Felix Naumann, Vuk Ercegovic, Mauricio Hernandez, Peter Haase, and Michael Schmidt. "GovWILD: integrating open government data for transparency." In Proceedings of the 21st International Conference on World Wide Web, pp. 321-324. ACM, 2012.
4. Heise, Arvid, and Felix Naumann. "Integrating open government data with stratosphere for more transparency." Web Semantics: Science, Services and Agents on the World Wide Web 14 (2012): 45-56.
5. Hoxha, Julia, and Armand Brahaj. "Open government data on the web: A semantic approach." In Emerging Intelligent Data and Web Technologies (EIDWT), 2011 International Conference on, pp. 107-113. IEEE, 2011.