

Beverage Graph: Connecting Data about Consumable Liquids

Jessica Singer¹, Robert Warren¹

¹Myra Analytics, Ottawa, Ontario

Abstract

We describe the design and ongoing update of a knowledge graph and its assorted ontologies which describe beverages and their commercial availability as products. Previous approaches have focused on beverage types or brands with limited support for tracing the product's content or identifying the specific product being consumed by a person. This inability to link the product and source has until now been a hindrance to nutritional studies and food traceability systems.

Keywords

Beverage Ontologies, Beverage Products, Beer Products, Juice, Consumable Liquids

1. Introduction

The Beverage Graph is an ontology-backed, knowledge graph focused on beverages, their styles, brands and the packaging in which they are commercially available. Initially created to support commercial brewing activities, it has been made available for public use and to encourage linking to other knowledge graphs. It is available online through data dumps at <https://rdf.ag/> or through a sparql server at <https://rdf.ag/sparql>. URIs are dereferencable and available in all RDF serializations through HTTP content negotiation.

Previous projects in this area have primarily been simple data dumps, without strong schema or ontological structures. Online web sites dedicated to beverage reviews occasionally have an external API to retrieve data but lack support for shared identifiers or linked data principles[1]. Bev-On¹ was an early OWL ontology attempt at building a structured representation of beverages but it is now unmaintained and an early knowledge graph project BevGraph² is no longer active. A missing element within all current beverage datasets is the relationship between the product that is actually handled by the consumer and the substance within the product: datasets seem to focus exclusively on one or the other. Most nutritional and dietary datasets themselves will reference a specified measured serving of a substance rather than that of a commercial product. Product nutritional labelling will itself reference a measured serving, sometimes disconnected from the container capacity, and may only be a statistical approximation of the


FOIS 2021 Ontology Showcase, held at FOIS 2021 - 12th International Conference on Formal Ontology in Information Systems, September 13-17, 2021, Bolzano, Italy

✉ singer@myraanalytics.ca (J. Singer); warren@myraanalytics.ca (R. Warren)

🆔 0000-0002-7066-1141 (R. Warren)



© 2021 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

¹<http://rdfs.co/bevon/latest/html>

²<https://github.com/bevgraph>

generic substance rather than an empirical measurement. The separate ontological structures representing substances, containers, commercial products and their individual instances allow the graph to be properly integrate with both empirically and statistically approximated nutritional datasets. As most consumers interact with nutritional substances through products, this ontological bridge will enable better end-user reporting of consumption which will lead to better nutritional analysis and recommendation.

2. Design

Knowledge Graph, vocabulary, schema, taxonomy and ontology are all terms that have come to be used interchangeably in the literature, causing no small amount of confusion. Beverage Graph is meant to be used as a RDFS/OWL ontologically-based Knowledge Graph capable of integrating with as much of the food supply chain as possible. It currently numbers over 50M triples, growing daily and available as a data dump at <https://rdf.ag> or through a Sparql endpoint at <https://rdf.ag/sparql>. All Beverage Graph URIs are de-referenceable and accessible in most data formats through HTTP Content Negotiation.

The core of Beverage Graph relies heavily on the schema.org and GS1 vocabularies. Schema.org[2] is arguably one of the most successful RDFS web vocabularies currently in use. It provides support for store inventory recording, commercial offering and product variant enumeration. While an official OWL version is available on experimental basis, we simply type the relevant terms and properties as OWL entities.

GS1 Global's Webvoc[3] is similarly available as an RDFS vocabulary that we augment using OWL classes. The GS1 Webvoc has its roots in commercial logistics and product management, providing support for the labeling of the product, branding and it's identification for inventory purposes. GS1 Webvoc also provides the `gs1:packaging` property and `gs1:PackagingDetails` class which permits the creation of standardized package descriptions including their dimension and weight. Currently, neither vocabulary provides a satisfactory solutions for "compound packaging" for bundled containers. We resolve this issue by having intermediate packaging listing parent item and count until such a time as a standardized solution be made available.

The largest issue in aligning these two vocabularies was the resolution of what a *product* is, as represented in Figure 1. We understand that Beer, Porter Beers and that a (hypothetical) Porter Beer brewed by ACME exist as facts, but that pragmatically, ACME Porter Beer can only really exist within a container.

Furthermore, there is more than one size of container (variant) and each physical container is filled from beer from a specific lot (beer batch). The specific arrangement of Figure 1 leverages the strengths of both schema.org and GS1 vocabularies to represent all aspects of a beverage product. The actual contents of beverages is represented using the Beer[4] and FoodOn[5] ontologies which gives the Beverage Graph coverage for Beers, Ciders, Meads, Juices and "Flavored" Juice Drinks, with support for coffees, teas, wines and hard liquors to come at a later date. Even in the case of untreated spring water, beverages are created through a process that transforms ingredients into a product. The design of the Beverage Graph allows the use of multiple ontologies to discover these processes. As an example, a query of the Porter class will reveal linkages to the Hops[6] ontology which lists Golding hops as a common ingredient to the

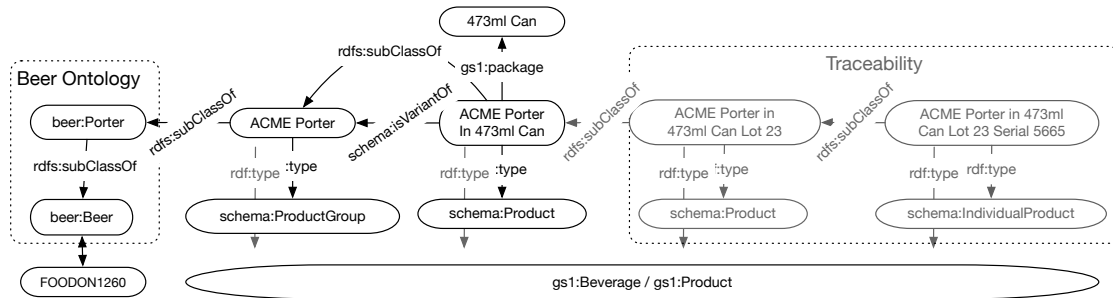


Figure 1: A fictitious ACME Porter is represented as modeled within the graph using the Beer, GS1 and schema.org ontologies. Note that we can readily link this data to specific lots or instances to deal with beverage recalls and traceability.

beer style. Similarly for fruit juices, we leverage the FoodOn ontology to ontologically reveal ingredients, as in Figure 2.

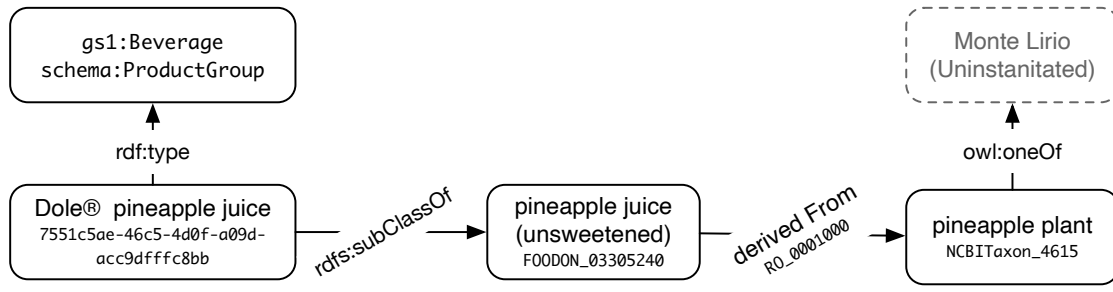


Figure 2: A similar instantiation of a pineapple juice, using FoodOn pineapple juice classes. Note FoodOn’s use of the Relation ontology and a possible expansion to take into account the specific varietal grown for juicing purposes.

The Dole®brand pineapple juice product group is a subclass of the FoodOn pineapple juice class which references its source fruit. It is evident that the structure has enough flexibility to reference the specific cultivar instead of a generic pineapple plant. Figures 1 and 2 are simplifications of the data available within the graph and do not represent properties such as packaging, manufacturer, brand and product description and location information.

The beverage graph uses the W3C Provenance [7] ontology is to trace sources and beverage processes, the OGC Time [8] ontology provides temporal annotations, the SKOS[9] vocabulary used for descriptions and the OGC GeoSparql[10] vocabulary provides geolocation information. We deliberately choose these mature, well engineered vocabularies over simpler solutions to better support the complexity of the real world data being represented.

Because several commercial sources are used to update the Beverage Graph, Entity resolution is an important process due to the overlap between commercial data sources. The graph nature of the data provides a ready made structure for a statistical record linkage[11] model to be build and GeoSparql containment properties provides a quick means of obtaining coarse location

matching when combined with the GeoNames[12] RDF dataset. When any two entities are determined to be the same, the SKOS-XL[13] vocabulary is used to convert the most recent entity node into a skos-xl:Label node which points to the authoritative entity. This approach preserves the original data provenance and allows us to “walk back” erroneous merges if needed.

We note with disappointment that vocabulary reuse seems to be a “do as I say and not as I do” principle and that similar properties are often re-implemented. Concurrently, few graph databases provide the facilities, or are configured, to make use of ontological equivalencies when querying data. For this reason, the Beverage Graph often contains redundant properties in order to make data consumption as simple as possible. A small series of ontological statements³ is also maintained here as a means of aligning temporal statements between PROV-O, Time and schema.org as well as documenting equivalencies between common properties such as gs1:organizationName, schema:name and foaf:name. The data can be consumed with or without these ontological axioms.

3. Discussion

The construction of the graph highlighted the complexities of commercial data management, the benefits of ontological backing and the complexities of integrating different ontological backed datasets. In acquiring external data, commercial API design reflect the needs and views of their owners which can result in unexpected data representations. A product variant should reference product instances that vary on explicitly defined, specific dimensions. Consumer facing API will often return product variants based on undefined conditions which may include similar packaging type, volume or store location which makes automated integration difficult.

Issue in entity resolutions have highlighted the usefulness of generic terms such as the Geonames 6295630 “planet earth” entity as a generic stand-in for the locality of a brewery as this information is not always available. This avoids the sort of issues that would occur in relational databases with null values, in that the data is always logically consistent and schematically complete even through it is factually imprecise. Operationally, this greatly reduces the complexity of entity resolution queries as fewer exceptions must be handled.

Ontology quality literature focuses on ontological completeness, logical consistency and structural issues[14] that are not always relevant to the actual operational use of the ontology itself. Some ontologists view “enumerative completeness” as an (unrealistic) primary objective, other rely on a reasoner reporting logical consistency and still others insist on over-constrained ontological constructs. Again, from an end-user perspective design consistency is the most important aspect through current tools and approaches may not enforce it. Our concern with ontology reuse is poor high-level documentation and the lack of consistency (or curation) in the ontological structures used across instances. FoodOn as an example is a collaborative project curated by multiple people and one that has chosen to import non-ontological datasets in bulk. Coordination across multiple designers can be difficult without close coordination and the large amounts of imported terms can make it difficult to identify the curated parts of an ontology and those still under review.

³<https://rdf.ag/o/BeverageGraph>

In this case, the consequence is that there are two mechanisms for defining a fruit juice and some confusion as to whether it derives from the plant or the fruit. Both mechanisms are ontologically consistent, but it makes querying the FoodOn ontology more difficult and potentially duplicates terms. Much has been made of “the code being the documentation” but at scale, ontological integration must be done programmatically and these issues will not be discovered without a high level overview of how specific real world objects are modeled. Too often, “suggestions” are made about the proper use of an ontology when it should be clearly specified. While well intentioned, the cost of flexibility in solving too many problems is a series of poor solutions instead of one good one.

A parallel can be made with the early experiences of the Dublin Core standards which failed to provide an official structure for citing a bibliographic work in RDF while simultaneously publishing a `dcterms:bibliographicCitation` term. The only way known to the authors to use Dublin Core coherently is the Bibo[15] ontology which provides a minimal structure to `dcterms` and which is being supplanted by the SPAR ontologies. To this end, we wish to highlight the requirement for term labels, term descriptions and ontological object narratives that can explain an ontology at a high-level. Too often, we read ontology documentation that focuses on itself rather than on its uses and without commenting on the instantiation of classes or how to solve actual problems within the domain.

Lastly, the actual semantic power of OWL2 ontologies is immense which, in a parallel to software design, can tempt designers to use overly complex technical solutions to simple problems. Ontology end users that wish to solve their own problem will naturally gravitate to the simplest, most documented solutions as it has the lowest cost of implementation.

4. Applications

Beyond its initial focus on supporting beer brewers, the Beverage Graph is flexible in its design to support additional information as to the product, the generic beverage and detailed packaging information, including whether the packaging is recyclable and its composition.

This opens the door to low hanging fruit studies on the prevalence of reusable packaging versus recyclable packaging and their relative volumes within specific markets. As other datasets also report the nutritional / calorimetric content of products, it becomes possible to quickly generate a partial but accurate nutritional profile of a person’s diet simply by scanning the barcode located on their beverage as they consume it. A direct application is in the resolution of the product content on an ontological basis based on its nomenclature. Consider the case of “Cider” which can mean an alcoholic beverage from fermenting apples or unfiltered apple juice or (confusingly), an “Non-alcoholic Cider” sold in the context of alcoholic beverages that contains no or only trace amounts of alcohol. As the Beverage graph reports the commercially mandated alcohol by volume (`beer:abvValue / gs1:percentageOfAlcoholByVolume`) for beverages an appropriate determination can be made.

The addition of linkages to legislative ontologies may be most interesting from an analytical viewpoint as legislation is heavily dependant on context and local culture. A direct example is the contrast between beer, an alcoholic beverage, and vanilla extract, a baking ingredient. While beer may contain alcohol, it is not mandated to and may actually be non-alcoholic. Vanilla

extract is mandated to contain a certain percentage of alcohol in order to be considered an extract but is regulated as a food and not an alcoholic beverage. Legislatively, their intended use dictates the regulatory regime under which they are controlled. From a public health perspective, it is their compositional properties that will dictate their capacity to be abused.

In closing, the Beverage Graph provides a core from which other datasets and ontology can link against or extract a working set. It is freely available, well labeled and aims to be as open as possible.

5. Conclusion

The Beverage Graph is a maintained collection of instances and ontological classes that document commercially available beverages, their contents and their packaging. It's construction allows for integrations with other external data sets and lends itself to dietary, commercial and food production analysis. As additional upstream data sources are acquired, the graph will be expanded to more brands and beverage types such as coffee, tea and hard liquors.

References

- [1] C. Bizer, T. Heath, K. Idehen, T. Berners-Lee, Linked data on the web (ldow2008), in: Proceedings of the 17th international conference on World Wide Web, 2008, pp. 1265–1266.
- [2] R. V. Guha, D. Brickley, S. Macbeth, Schema.org: Evolution of structured data on the web, *Commun. ACM* 59 (2016) 44–51. doi:10.1145/2844544.
- [3] GS1 Web Vocabulary Standard, 1.6.1, GS1, 2015. URL: https://www.gs1.org/docs/gs1-smartsearch/GS1_Vocabulary_Standard.pdf.
- [4] R. Warren, J. Singer, Beer ontology, 2021. URL: <https://doi.org/10.5281/zenodo.4672337>.
- [5] D. M. Dooley, E. J. Griffiths, et al., Foodon: a harmonized food ontology to increase global food traceability, quality control and data integration, *npj Science of Food* 2 (2018) 23.
- [6] R. Warren, J. Singer, Hops ontology, 2021. URL: <https://doi.org/10.5281/zenodo.4672692>.
- [7] Prov-o: The PROV Ontology, 2013. URL: <https://www.w3.org/TR/prov-o/>.
- [8] S. Cox, C. Little, Time Ontology in OWL, 2013. URL: <https://www.w3.org/TR/owl-time/>.
- [9] D. Brickley, A. Miles, SKOS Core Vocabulary Specification, W3C Working Draft, W3C, 2005. URL: <http://www.w3.org/TR/swbp-skos-core-spec/>.
- [10] OGC GeoSPARQL - A Geographic Query Language for RDF Data, 2012. URL: <http://www.opengis.net/doc/IS/geosparql/1.0>.
- [11] W. E. Winkler, Advanced methods for record linkage, Technical Report rr945, Statistical Research Division, U.S. Bureau of the Census, 1994.
- [12] M. Wick, Geonames ontology, 2015. URL: <http://www.geonames.org/about.html>.
- [13] A. Miles, S. Bechhofer, SKOS-XL Simple Knowledge Organization System eXtension for Labels, Technical Report, 2009. URL: <https://www.w3.org/TR/skos-reference/skos-xl.html>.
- [14] S. M. Gurk, C. Abela, J. Debattista, Towards ontology quality assessment, in: MEP-DaW/LDQ@ESWC, 2017.
- [15] B. D’Arcus, F. Giasson, Bibliographic ontology specification, 2009. URL: <https://bibliontology.com/specification.html>.