The Universal Data Cube

and the benefits it brings to Weave

Introduction

The Universal Data Cube is an idea which, when implemented within the Weave project and community, may begin to fulfill some of the wider visions of the Open Indicators Consortium, such as the notion of a “public common indicator taxonomy”, a “data commons”, a means to “data democratization”, including “nested indicators”, and deep integration of comprehensive and extensible metadata. The Universal Data Cube is a blueprint for a technology developed to investigate the possibility of leveraging Semantic Web technologies such as Ontologies and federated data repositories in the context of Weave. It was conceptualized by the IVPR group last spring and developed significantly by Curran Kelleher during his stay at the University of Konstanz from June to August 2010. The intention of this document is to introduce the work to the Open Indicators Consortium, and to begin a discussion investigating the decision to make the Universal Data Cube the standard data model and dissemination infrastructure for Weave.

The Current Weave Data Model

Weave is a powerful visual analysis tool which is groundbreaking in many ways, and has tremendous potential for integration with the kind of powerful and distributed data backend originally envisioned by the Open Indicators Consortium.

The current Weave data model can be characterized as follows:

- A dataset is a collection of mappings from keys (of standard “key types” such as FIPS Code) to values in attribute columns.
- Attribute columns are defined be name only (no metadata about units or meaning).
- Attribute columns are categorized within a hierarchical indicator taxonomy.
- The “type of thing” that a record is in an attribute column is defined only by a string, such as “FIPS” or “State”.
- Geometry collections are published with key types as well, which must match with the declared key type for related attribute columns.
- Multiple data servers can be queried simultaneously by a client, but the data is not interoperable unless both servers use common key column names.

Drawbacks of The Current Weave Data Model

The Universal Data Cube was designed specifically to address the disadvantages of the current Weave model, which are the following:

- Hierarchically related key types are not linked. For example, US Counties and US States are totally independent key types, though in fact they have a containment relationship and are adjacent levels in a hierarchy.
- Key types referring to the same things are not compatible. For example, the key type for “US State FIPS” and “US State abbreviations” are incompatible and have no connection.
- Attribute columns representing the same measure with different units are not compatible. For example, “Population in thousands” is not comparable with “Population in millions”

automatically within Weave.

- There is no way of resolving when two datasets provide comparable columns. For example, there is no way of automatically deducing that the attribute column “Pop” in dataset A is the same measure and unit as the attribute column “population” in dataset B.

The Universal Data Cube Model

A data cube is a way of structuring data such that it represents qualities of an intersection of many simultaneous hierarchies. For example, imagine a data set which contains the measures “Population”, “Average Income”, and “Total Wages”; for all US states and counties; for all years, months, and quarters from 1990 to 2008; and for all industry categories in the NAICS industry. This is the structure of the Bureau of Labor Statistics Employment dataset, and is a great example of a data cube. For each unique combination of geographic region, temporal region and industry category, the data cube contains numeric values for “Population”, “Average Income”, and “Total Wages”. The Universal Data Cube (UDC) Model is a metadata vocabulary for describing the contents of data cubes. The UDC System will provide the means to create and publish metadata about data cubes, as well as the means to expose existing data sets as data cubes within the metadata framework.

The Universal Data Cube (UDC) vision has at its core the metadata model represented in Figure 1. This model can be expressed as an ontology using the Web Ontology Language (OWL), a key Semantic Web technology for defining metadata vocabularies. Each word in the diagram represents an OWL Class, and the connecting “crows feet” represent one-to-many relationships between them. Instances of the “DataCube” class have corresponding databases which hold their content and can be queried. All other classes represent metadata describing what those data cubes contain and mean.

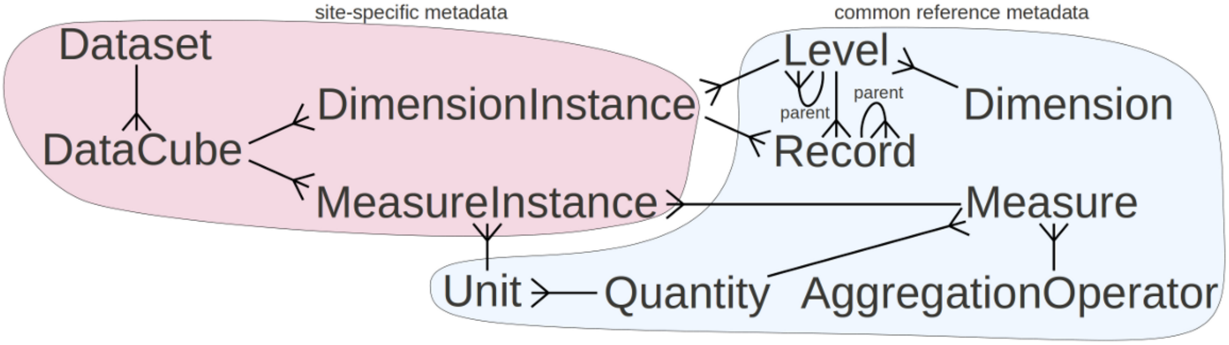

Figure 1. The Universal Data Cube Ontology.

Some classes are globally defined and embody a kind of “common knowledge”. These classes are: Dimension, Level, Record, Measure, AggregationOperator, Quantity, and Unit. Definitions of these should be made publicly available as web resources for others to reference now and in the future. Other classes are defined only locally to a specific site, and describe the content of a set of data cubes. These classes are: Dataset, DataCube, DimensionInstance and MeasureInstance.

The UDC framework uses Semantic Web technology to publish all metadata. In other words, a UDC server instance would expose persistent instances of all ontology classes as Linked Data, and provide a SPARQL query endpoint as well. Servers which expose instances of the DataCube class are expected to also expose a service for querying the contents of the data cubes, which typically resides in a database. The data cube query service is the only additional public-facing service introduced by the Universal Data Cube system. This service can be viewed as an extension of the Semantic Web, and as the main functional specification of the UDC framework.

Advantages of the Universal Data Cube Model

The UDC Model has the following advantages over the current Weave data model:

- Hierarchically related records are intrinsically linked within a record hierarchy of a Dimension, so hierarchically nested data is well supported.
- Multiple data sets published by different sites which refer to the same objects resolve to the same entities, because each Record has its own unique URI. This means that when different measures are provided by different sites for the same sets of records, the data is able to be automatically integrated on the client side and analyzed as though it were one data set.
- When two datasets provide data for the same measure, the client can resolve this fact and integrate the data dynamically.
- When two sites provide data using the same measure with different units, a client can resolve this fact and compensate by performing unit conversions on the fly. In this way, the two data sets can be transparently integrated and visualized as a single whole.

Effect on Storage and Performance

The Universal Data Cube framework was originally designed as purely a data presentation layer. When used as such, the UDC framework queries the existing data directly from a relational database table, and has no more performance overhead than any web based data retrieval system. However, the data cube and rich metadata structure enables automatic aggregation, a feature not present in the previous Weave data model. For example, if you have a data set for all towns in New England, but not for all counties or states in New England, the UDC system can automatically generate these aggregations and present them as additional data.

With automatic aggregation comes the question: when to compute the derived aggregations? They could either be precomputed and permanently stored (no performance impact, requires additional storage), computed on demand (heavily impacts performance, requires no additional storage), or computed on demand and cached on disk (lightly impacts performance, requires a bounded amount of storage). The additional storage required to precompute and store the aggregations permanently is always less than the size of the original data, so we believe this would be the best option.

The Larger Vision

Aside from being a significant contribution to the Weave project, the Universal Data Cube system has the potential to become the enabling technology at the epicenter of a boom in data democratization. Should this happen, there would be tremendous motivation for developers to create new UDC client and server tools, and it would become a web standard, with an open source reference implementation. The potential future “web of data” landscape would have the following characteristics:

- Organizations can publish data on the Internet in a scalable and standard way.
- People who want to consume such data can do so using freely available standard clients (such as Weave), which are able to dynamically integrate data from an arbitrary number of sources.
- Application developers can use UDC clients in their own programs, and develop many independent applications for data visualization, analysis, or other uses.
- Server developers can embed the UDC server component into their own server-side software, enabling integration of a data analysis process directly from within their own server and database configuration.
- It will be common for tools such as Weave to navigate geographic regional data at arbitrary scales, such as country, state, county, city, and neighborhood, and at each scale be able to see what data is available and who provides it.

Conclusion

The Universal Data Cube framework is at a nascent stage of development, but shows significant potential as the data infrastructure for the Weave project and beyond. Building on the Semantic Web, it it intrinsically extensible and can scale across many servers and ownership domains. Because it leverages existing database technology for data storage and retrieval, it is scalable in terms of data set size as well. It will take significant time and effort to see the vision through to completion, but we believe that it will be well worth it.