The Universal Data Cube

Curran Kelleher

Abstract

Visualization and analysis tools often lack metadata support, and do not explicitly support hierarchical data cubes. The Universal Data Cube is a vision for a world wide web in which hierarchical data cubes are first class citizens and rich web-based data visualization and analysis tools are commonplace. The Universal Data Cube Ontology is the data model at the core of this vision, which supports rich descriptions of hierarchical data cubes. In the Universal Data Cube System, existing semantic web technologies are used to publish metadata and structural descriptions of data sets, while a hierarchical data cube query endpoint exposes the data itself.

1 Introduction

2 Related Work

2.1 The Semantic Web

The semantic web and linked data movement provide a foundation for the distributed metadata infrastructure required to realize the Universal Data Cube. A rudimentary understanding of semantic web technologies such as XML, RDF, and OWL is an outline of existing semantic web technologies.

2.2 Resource Description Framework

The Resource Description Framework (RDF) [2] is the fundamental data model of the semantic web. It allows one to encode metadata about resources as "triples" of the form `subject predicate object` where `subject`

represents a resource, `predicate` represents a property, and `object` represents either a resource or a literal value. The standard syntax for representing RDF data is RDF XML, a specification published by the World Wide Web Consortium (W3C). The first specification was published in 1999, and a revised specification was published in 2004.

2.3 Web Ontology Language

The Web Ontology Language (OWL) [3] allows one to define ontologies for use in RDF descriptions. OWL was also created by the W3C in an effort to standardize ontology specification. Ontology design using OWL is much like object oriented design. One can specify class hierarchies defining resource types, and properties of those classes. For some ontologies, OWL classes and properties can be used to inform design of classes in an object oriented software system as well as a corresponding relational database schema.

2.4 Linked Data

Tim Berners Lee put forth a set of principles and practices which would enable the semantic web to exist as a browsable distributed data system. The principles are as follows:

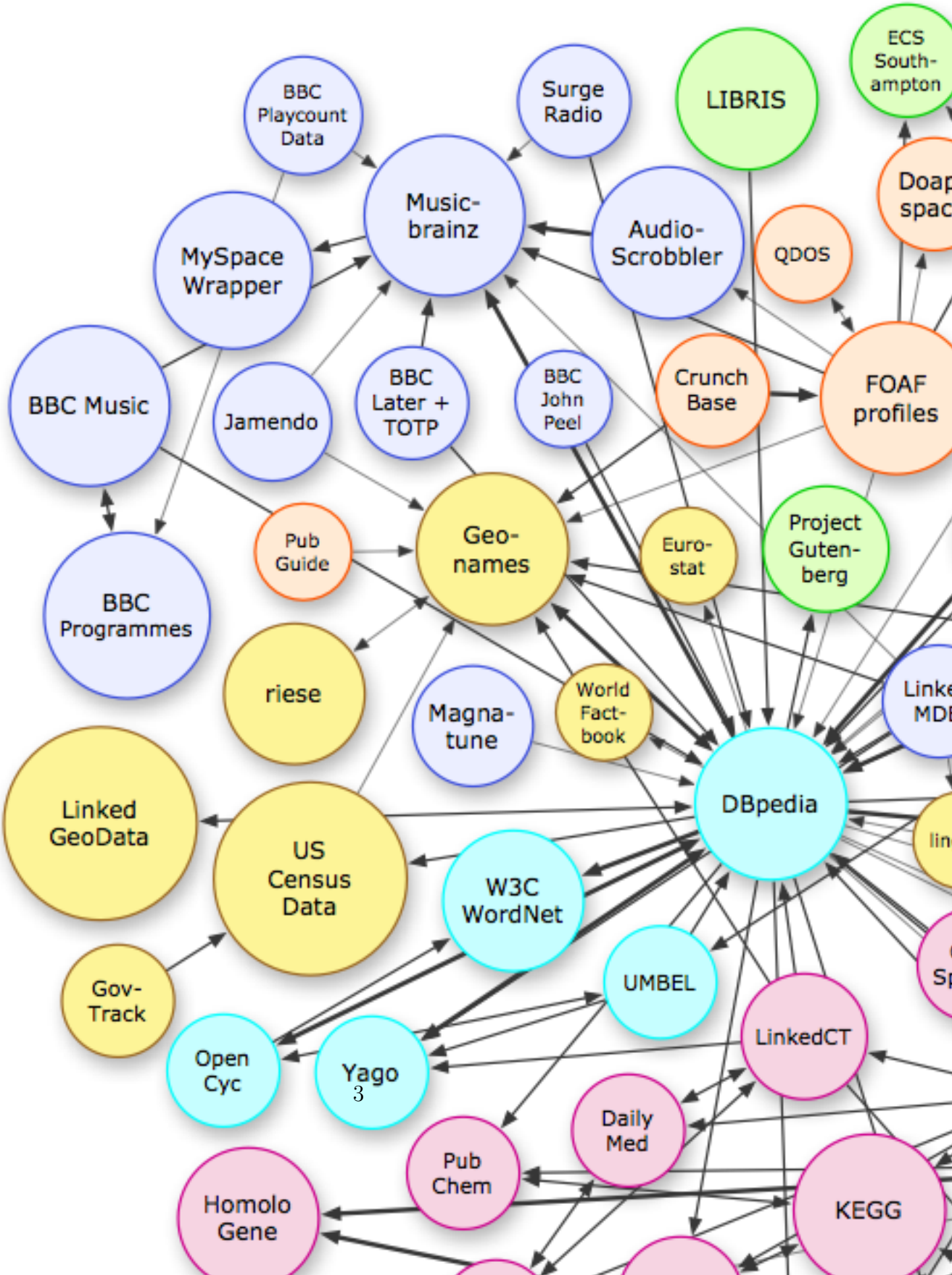
1. Use URIs as identifiers
2. Use derefencable URIs
3. Provide structured data about a thing when its URI is dereferenced
4. Include references to other dereferencable URIs in the structured data

Figure 1 shows the state of the linked open data cloud as of July 2009. Notice that DBPedia is emerging as a prominent hub.

2.5 Existing Ontologies and Knowledge Bases

2.5.1 DBPedia

DBPedia is a linked data project publishing information extracted from Wikipedia. They have published their own ontology for describing Wikipedia articles.


BBC Playcount Data
Surge Radio
LIBRIS
ECS South-ampton
Music-brainz
Audio-Scrobbler
QDOS
Doap spac
MySpace Wrapper
BBC Music
Jamendo
BBC Later + TOTP
BBC John Peel
Crunch Base
FOAF profiles
Pub Guide
Geo-names
Euro-stat
Project Guten-berg
BBC Programmes
riese
Magna-tune
World Fact-book
Linke MDB
Linked GeoData
US Census Data
DBpedia
W3C WordNet
Gov-Track
UMBEL
Open Cyc
Yago
LinkedCT
Daily Med
Pub Chem
Homolo Gene
KEGG

2.5.2 GeoNames

GeoNames is a linked data project which publishes descriptions of geographic regions.

2.5.3 OpenCyc

2.5.4 WordNet

2.5.5 Freebase

2.5.6 Umbel

2.5.7 SameAs.org

2.5.8 Sindice

2.5.9 data.gov

2.5.10 data.gov.uk

2.6 Data Cubes

A data cube, or OLAP (Online Analytical Processing) cube, is a data abstraction capable of representing multiscale multidimensional data sets. The fundamental elements of a data cube are measures and dimensions. A data cube is an n-dimensional array of cells, each cell containing values for each measure. Dimensions define spaces which contain regions over which measures can be aggregated.

2.7 Hierarchical Data Cubes

Any data cube dimension may be hierarchical, as it represents a space over which measures can be aggregated. A data cube containing hierarchical dimensions is called a hierarchical data cube. Each unique combination of dimension hierarchy levels defines a non-hierarchical data cube. Therefore a hierarchical data cube can be represented by a lattice of data cubes, linked together by parent-child relationships among dimension levels.

2.8 MDX

MDX is a query language for defining hierarchical data cube projections.

2.9 Existing Software Tools

2.9.1 OpenLink Virtuoso

2.9.2 Pubby

2.9.3 Mondrian

Mondrian is an open source OLAP database implementation in Java which supports hierarchical data cubes and MDX.

3 The Universal Data Cube Ontology

The Universal Data Cube Ontology defines a vocabulary for describing the structure and metadata of hierarchical data cubes. It is designed in such a way that knowledge and data are published separately.

4 The Semantic Web as a Hierarchical Data Cube

Semantic graphs can be automatically converted into hierarchical data cubes. Each class hierarchy is taken as a dimension. Each class is taken as a level. Each resource is taken as a record. Each

References

[1] LinkingOpenData W3C Community. Linking open data. World Wide Web electronic publication, 2010.

[2] Ora Lassila, Ralph R. Swick, World Wide, and Web Consortium. Resource description framework (rdf) model and syntax specification, 1998.

[3] D.L. McGuinness, F. Van Harmelen, et al. OWL web ontology language overview. *W3C recommendation*, 10:2004–03, 2004.