KDI
A Methodology for Data Integration

Fausto Giunchiglia and Mattia Fumagallli
University of Trento
Overview of the Model

Generalized Queries

Etypes Model

Evaluation

Case Studies
Overview of the Model

Components of the Model

“Data wrangling”
Components of the Model

- Datasets
- Standards
- Generalized Query

KDI Methodology

Schema

Language

MODEL

Application
Components of the Model

Datasets
- European Open Data Portal
- Open Data Trentino
- Open Street Map

Standards
- SIRI
- INSPIRE
- GTFS
- GQ 1, 2, 3 -n

Generalized Query

Ontological principles

Schema

Language

KDI Methodology

MODEL

Application
“Data wrangling”
Generalized Queries

- Application Scenario
- Identify the Concepts
- Queries Collection Mechanism
Choose the application scenario
Start with a set of ground queries:

Given the application scenario, a set of queries will arise which place demands on an underlying ontology.

- Give a list all the Hotels in X City which has facility for disable?
- Identification of general query pattern

  Give me all X in Y AND WHERE.property.True

- Identification: Concepts and Properties
  
  **Entity:** Hotel, City
  
  **Property:** Hotel.name, City.name, facilityForDisable.

  **Boolean**
Identify all the **core concepts** which are needed to answer the **generalized queries**.
Query generation methodology

1. via a user study, for instance via questionnaires or focus group
2. via a benchmarking analysis of existing sites and data
3. heuristically based on the understanding of the domain developer
4. from datasets – (see rapidminer tree example… see also http://quepy.machinalis.com/)
5. a combination of the above
EER Model

- Schema Level
- Language Level
The Plaza Hotel

Building

Country

Status

Date of construction: 1907
EER Model (example)
Alignment with Upper Ontology and Classification

Physical Place
Location
Hotel
Train
Bus
Building
Artifact
Country
Road
Mountain
Restaurant
Statue
Elevator
Social Object
Agency
Trip
Party
Dinner
Mental Object
Movie
Recipe
Address
Property
House No
Weather
Cold
Wi-Fi
Price
Ticket
Wheelchair Accessibility
Ontology Design

Formal Modelling

- building
  - agriculturalBuilding ≡ farmBuilding
  - commercialBuilding
  - educationalBuilding
  - factory ≡ productionBuilding ≡ industrialBuilding
  - farmBuilding ≡ agriculturalBuilding
  - governmentBuilding
  - healthcareBuilding ≡ healthcareFacility
  - healthcareFacility ≡ healthcareBuilding
  - industrialBuilding ≡ productionBuilding ≡ factory
  - placeOfWorship ≡ religiousBuilding
  - productionBuilding ≡ factory ≡ industrialBuilding
  - religiousBuilding ≡ placeOfWorship
  - residentialBuilding
  - transportationBuilding

- administrativeDivision
  - borough
  - city
  - country ≡ self_governing
  - county
  - federalDistrict ≡ unionTerritory
  - prefecture
  - region
  - state
  - town
  - unionTerritory ≡ federalDistrict
  - ward
Issue_1: Attributes and DataProperties

Complex

Simple

Yes

Wi-Fi

Address

Street Name

Post code

AddressCountry

AddressCity

House No

Schema Level
Issue_2: Relation and ObjectProperties

Hotel (The Plaza) ➔ AddressCountry ➔ Country (USA) ➔ City (New York) ➔ PartOf (New York)

Schema Level
Freeway

Highway

Class

Name

ID

Etype

has

Attribute Definition

n

contains

1

Quantitative Attribute

Qualitative Attribute

Relational Attribute

DataType

Semantic Relation

IS_A

a broad highway designed for high-speed traffic

expressway, freeway, motorway, pike, state highway, superhighway

synonym

hyponym

Language Level
Evaluation

- Inconsistency check
- Incompleteness check
Evaluation of Ontological Model

- Inconsistency
  - circularity errors: [ex. Traveler subclassOf Person; Person subclassOf Traveler; ]
  - semantic inconsistency errors: [ex. Airbus or Waterbus subclassOf Bus]
  - partition errors: [ex. Non stop Flight SubClassOf InternationalFlight and DomesticFlight where International and Domestic flight are disjoint]
- Incompleteness: On traveling domain, if we classify only beach and mountain location, and we do not consider cultural heritage site
- Redundancy
  - Identical formal definition of some class
  - Identical formal definition of instances
Case Studies

Evaluation of Methodology

Result
Case Studies (example)

Topics

- Geospatial
- Event in Trento
- Where to eat in Trento
- Tourism in Trento
- Transport in London
- Emergency Response in London
- Real Estate

Case Studies
• Technique
  • Used standard Human Computer Interaction (HCI) technique
  • Open Ended questions mixed with Likert scale closed questions
• How: Balanced Questioners
• Number of participant: 18
• Participants Information
  • Nationality: Italian, Indian, Germany, Brazil, Ukraine, Ethiopia, Mexico, Uganda, Cameroon
  • Gender: Male 13  Female 5
  • Age Range: 18-25 (14), 26-30 (4)
  • Level of education: Undergraduate (3)  Postgraduate (15)
**Perspicuity**: How easy it is to get familiar with the methodology

**Efficiency**: How effectively user can perform the process

**Dependability**: Can user control the process

**Stimulation**: Is it exciting and motivating

**Novelty**: Is it innovative and creative
Pros
• Well Structured
• programmatically durable
• It practically allows describe the world
• Provides methods to minimize the distance between the real world and the abstraction
• Helps finding out eventual defects of the ontology and helps correcting them: taxonomic errors, inconsistencies, reliability

Cons
• You need many practice to build something very well
• Needs more time to master
• difficult to identify class for to align with top level
• Necessary to write documentation to clarify choices and terms
• Formalizing DERA to DL
Data on the Web Best Practices W3C Recommendation 31 January 2017
https://www.w3.org/TR/dwbp/


