ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

geoinformation
engineering
ikg eth zurich

IKG
Institut für Kartografie
und Geoinformation

Semantic description of spatial analysis data sets on the Web

Problem statement and Main Goal

A problem, which is often observed on the Web is that there is lack of provenance information about spatial and statistical data sets, such as their origin, the operations performed for their generation and the people involved in the procedure. Due to this lack of information, data analysts do not know which analysis method can be meaningfully applied to a specific data set and often implement methods, which are inappropriate. As a result, they draw wrong conclusions about the data set.

The main objective of this Master thesis was to investigate how geographic data sets can be described with linked data for improved analysis. In this line of research the following research questions have been developed in order to reach the main goal of the Thesis:

- Which are the available linked data vocabularies for the main purpose of the Thesis?
- Which are the differences between these vocabularies and meta-data standards?
- In how far do they serve this purpose, and what is missing?
- How could a linked data vocabulary for geographic and statistical analysis look like?

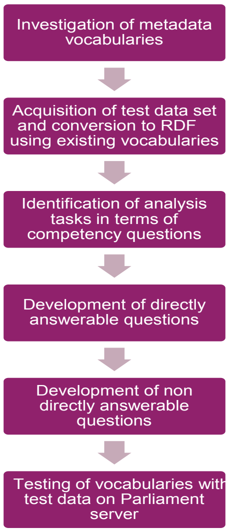

Investigation of metadata vocabularies
Acquisition of test data set and conversion to RDF using existing vocabularies
Identification of analysis tasks in terms of competency questions
Development of directly answerable questions
Development of non directly answerable questions
Testing of vocabularies with test data on Parliament server

Figure 1: Approach and workflow of the Thesis

Data and Methodology

- Two json files providing meta-information about the data sources and data sets included in the AURIN portal were used as data
- The principal objective of the methodology was to test vocabularies and whether they allow retrieval of data sets corresponding to answers to the competency questions that can be asked for analysis purposes
- Direclty answeable questions can be answered from the content of information in a database using content vocabularies
- Indirectly answerable questions cannot be answered directly from the content of information in the database, since there are some restrictions on the data to be retrieved based on some operations that should be carried out
- These operations are combination, comparison, transformation and calculation of statistics relevant to a particular data set.
- There is a kind of meta-information determining to which data sets the derivation can be meaningfully applied
- Translation of each competency question into a query
- Request of a particular data set and its retrieval were possible through the Parliament server

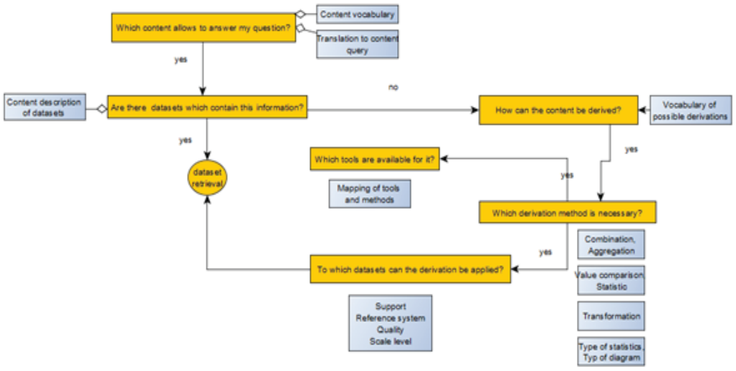
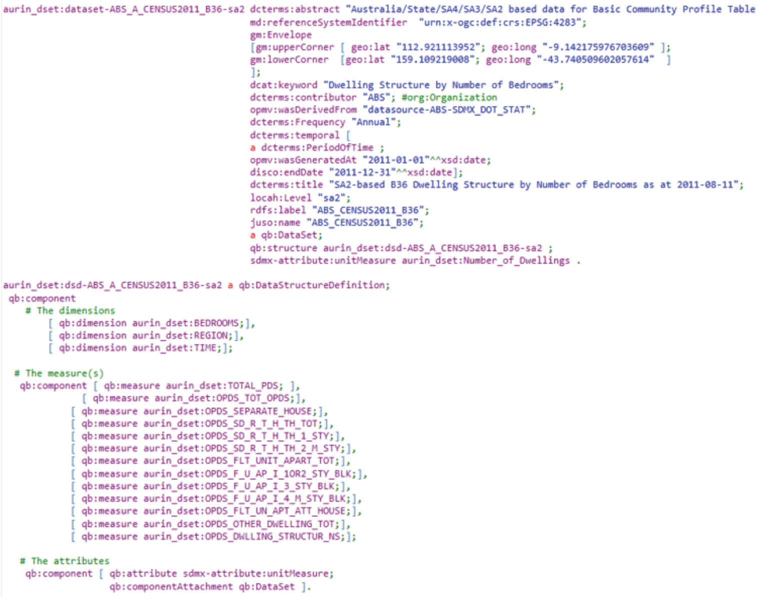

Which content allows to answer my question?
Content vocabulary
Translation to content query
yes
Content description of datasets
Are there datasets which contain this information?
no
How can the content be derived?
Vocabulary of possible derivations
yes
dataset retrieval
Which tools are available for it?
Mapping of tools and methods
yes
Which derivation method is necessary?
Combination, Aggregation
Value comparison Statistic
Transformation
Type of statistics, Typ of diagram
yes
To which datasets can the derivation be applied?
Support
Reference system
Quality
Scale level

Figure 2: General methodology for retrieving data for analysis tasks


```
aurin_dset:dataset-ABS_A_CENSUS2011_B36-sa2 dcterms:abstract "Australia/State/SA4/SA3/SA2 based data for Basic Community Profile Table
                                            md:referenceSystemIdentifier  "urn:x-ogc:def:crs:EPSG:4283";
                                            gm:Envelope
                                            [gm:upperCorner [ geo:lat "112.921113952"; geo:long "-9.142175976703609" ];
                                            gm:lowerCorner  [geo:lat "159.109219008"; geo:long "-43.740509602057614"  ]
                                            ];
                                            dcat:keyword "Dwelling Structure by Number of Bedrooms";
                                            dcterms:contributor "ABS"; #org:Organization
                                            opmv:wasDerivedFrom "datasource-ABS-SDMX_DOT_STAT";
                                            dcterms:Frequency "Annual";
                                            dcterms:temporal [
                                            a dcterms:PeriodOfTime ;
                                            opmv:wasGeneratedAt "2011-01-01"^^xsd:date;
                                            disco:endDate "2011-12-31"^^xsd:date];
                                            dcterms:title "SA2-based B36 Dwelling Structure by Number of Bedrooms as at 2011-08-11";
                                            locah:Level "sa2";
                                            rdfs:label "ABS_CENSUS2011_B36";
                                            juso:name "ABS_CENSUS2011_B36";
                                            a qb:DataSet;
                                            qb:structure aurin_dset:dsd-ABS_A_CENSUS2011_B36-sa2 ;
                                            sdmx-attribute:unitMeasure aurin_dset:Number_of_Dwellings .

aurin_dset:dsd-ABS_A_CENSUS2011_B36-sa2 a qb:DataStructureDefinition;
 qb:component
    # The dimensions
        [ qb:dimension aurin_dset:BEDROOMS;],
        [ qb:dimension aurin_dset:REGION;],
        [ qb:dimension aurin_dset:TIME;];

  # The measure(s)
   qb:component [ qb:measure aurin_dset:TOTAL_PDS; ],
                 [ qb:measure aurin_dset:OPDS_TOT_OPDS;],
               [ qb:measure aurin_dset:OPDS_SEPARATE_HOUSE;],
               [ qb:measure aurin_dset:OPDS_SD_R_T_H_TH_TOT;],
               [ qb:measure aurin_dset:OPDS_SD_R_T_H_TH_1_STY;],
               [ qb:measure aurin_dset:OPDS_SD_R_T_H_TH_2_M_STY;],
               [ qb:measure aurin_dset:OPDS_FLT_UNIT_APART_TOT;],
               [ qb:measure aurin_dset:OPDS_F_U_AP_I_1OR2_STY_BLK;],
               [ qb:measure aurin_dset:OPDS_F_U_AP_I_3_STY_BLK;],
               [ qb:measure aurin_dset:OPDS_F_U_AP_I_4_M_STY_BLK;],
               [ qb:measure aurin_dset:OPDS_FLT_UN_APT_ATT_HOUSE;],
               [ qb:measure aurin_dset:OPDS_OTHER_DWELLING_TOT;],
               [ qb:measure aurin_dset:OPDS_DWLLING_STRUCTUR_NS;];

   # The attributes
    qb:component [ qb:attribute sdmx-attribute:unitMeasure;
                   qb:componentAttachment qb:DataSet ].
```

Figure 3: Example of the resulting RDF model

Conclusions

- Development of general methodology to answer analysis questions based on the content of information included in the data sets and other tools enabling users to conduct analysis operations
- Research questions were answered to a great extent
- Only two provenance properties were used, since no specific information provided in the json metadata files about the procedures, through which the data sets were generated or whether any transformations were conducted

Outlook

- Integration of the resulting RDF model in a data portal
- Creation of metadata descriptions by the users or by automatic translation tools
- Possibility for the users to annoatate their own data with the proposed vocabularies
- Development of a running recommender system for retrieval of the most appropriate data set for analysis puproses

Author: Ariadni Gaki
Head: Prof.Dr.Martin Raubal, Advisor: Dr.Simon Scheider

Master Thesis
Autumn Semester 2015