Collaborative Project

GeoKnow - Making the Web an Exploratory for Geospatial Knowledge

Project Number: 318159 **Start Date of Project:** 2012/12/01 **Duration:** 36 months

Deliverable 3.5.1

Initial Report On Spatial Data Quality Assessment

Dissemination Level	Public
Due Date of Deliverable	Month 20, 30/07/2014
Actual Submission Date	Month 20, 04/07/2014
Work Package	WP3, Spatial knowledge aggregation, fusing and quality assessment
Task	T3.5
Type	Report
Approval Status	Final
Version	1.0
Number of Pages	26
Filename	D3.5.1_Initial_Report_On_Spatial_Data_Quality_Assessment.pdf

Abstract: This deliverable provides a survey of the metrics used for measuring Spatial Data Quality and the initial results of the CROCUS, a semi-automatic tool developed to measure Data Quality in the context of GeoKnow.

Project funded by the European Commission within the Seventh Framework Programme (2007 - 2013)

History

Version	Date	Reason	Revised by
0.0	07/07/2014	First draft created	Muhammad Saleem
0.1	15/07/2014	Draft revised	Axel-Cyrille Ngonga Ngomo
0.2	15/07/2014	First version created	Muhammad Saleem
0.3	2/08/2014	Peer reviewed	Giorgos Giannopoulos
0.4	06/08/2014	Final version submitted	Muhammad Saleem

Author List

Organization	Name	Contact Information
INFAI	Muhammad Saleem	saleem@informatik.uni-leipzig.de
INFAI	Axel-Cyrille Ngonga Ngomo	ngonga@informatik.uni-leipzig.de
Unister	Didier Cherix	didier.cherix@unister.de
Unister	Ricardo Usbeck	ricardo.usbeck@unister.de
Unister	Christiane Lemke	christiane.lemke@unister.de

Executive Summary

The Linked Open Data cloud hosts over 300 publicly available knowledge bases covering a wide range of topics, with many of them containing geospatial annotations. Producing a single integrated geospatial data set from different RDF representations is one of the main challenges in the GeoKnow project.

In general, LOD knowledge bases comprise only few logical constraints or are not well modelled. Thus, merging geospatial features from different data sets is not straightforward. Common problems in this respect include distorted geometries and divergent meaning of meta-information associated with geo-spatial objects. With GeoKnow aiming to a address an extensive range of users including customers in an industrial environment, the quality of data sets resulting from fusing other data sets becomes a crucial factor for the acceptance and distribution of the project's results.

This deliverable addresses the issue of quality assessment for geo-spatial data sets. It presents a new semi-automatic approach called CROCUS developed in the scope of GeoKnow and how this approach can be used to assess and visualize the quality of geo-spatial datasets. To this end, we also present a survey of geo-spatial data quality features. The structure of this document is as follows: Section 2 provides a survey of the geospatial Data Quality standards and scientific contrifbutions. Section 2.3 describes a list of the geospatial Data Quality metrics. Section 3 motivates the need for a semi-automatic approach and summarises the advantages of CROCUS, section 4 describes its general approach in more detail. Section 5 shows application to a geospatial data set including visualisation before section 6 concludes.

Abbreviations and Acronyms

LOD	Linked Open Data
CROCUS	Cluster-based Ontology Data Cleansing
GIS	Geographic Information System
SDTS	Spatial Data Transfer Standard
OGC	Open GIS Consortium
CEN/TC 287	European Committee for Standardization Technical Committee 287
CGDI	Canadian Geospatial Data Infrastructure
NSDI	National Spatial Data Infrastructure
DIGEST	Digital Geographic Information Exchange Standards
SAIF	Spatial Archiving and Interchange Format
MUM	Multidimensional User Manual
QUIM	Quality Information Management Model
SOLAP	Spatial On-Line Analytical Processing
OSM	Open Street Map
CBD	Concise Bounded Descriptions
SCDB	Symmetric Concise Bounded Description
DBSCAN	Density-Based Spatial Clustering of Applications with Noise

Contents

List of Figures

List of Tables

1 Introduction

Data Quality is the degree of excellence exhibited by the data towards the actual scenario in-use. It is generally thought of as a multi-dimensional concept and is most commonly referred as "Fit-for-use", i.e., some applications are more critical towards high data quality and others may only require data of adequate quality [15]. For example, an application providing soccer players information using DBpedia may not require very high quality of data. On the other hand, prescribing a treatment to a cancer patient based on DBpedia information (such as drugs, people ethnicity and countries etc.) is simply not sufficient.

Data Quality is important because we need[1]:

1. Accurate and timely information to manage services and accountability.
2. Good information to manage service effectiveness.
3. To prioritise and ensure the best use of resources.
4. To report to auditors and inspectors who will make judgements about our performance and governance.

Large Geospatial data sets are specially prune to errors because they contain data from multiple providers and use different assumptions about structure and semantics of data [13]. In many of the use cases (such as transportation, navigation, GIS guidance), a very high quality of Geospatial data is required. A lack of Geospatial data quality can results in sever accidents such as: 1998 ski-lift accident in Italy and 1999 accidental bombing of the Chinese Embassy in Belgrade [13]. In recent years, the concern for Geospatial data quality has increased due to a number of factors including [10]: (1) increased data production by the private sector and non-government agencies, which are not governed by uniform quality standards (production of data by national agencies has long been required to conform to national accuracy standards), and (2) increased reliance on secondary data sources, due to the growth of the Internet, data translators, and data transfer standards, making poor quality data ever easier to get.

This deliverable provides a survey of the different metrics used for assessing the spatial Data Quality. In the next section, we provide a short overview of the state-of-the art and present a ranked list of the set of metrics used for spatial Data Quality.

[1] Source: http://goo.gl/tbesUU

2 Geospatial Data Quality Metrics Survey

The state-of-the-art in Geospatial data quality can be divided in to two directions: Geospatial Data Quality standards, and Geospatial Data Quality scientific research contributions.

2.1 Geospatial Data Quality Standards

ISO/TC 211[2] provides a series of standards that deal with various aspects of Geospatial Data Quality. In particular, ISO 19115-1:2014[3], ISO 19113:2002[4], ISO 19114:2003[5], and the technical specification ISO/TS 19138 Data quality measures[6] are important to be considered. ISO 19115-1:2014 defines the schema required for describing geographic information and services by means of metadata. It provides information about the identification, the extent, the quality, the spatial and temporal aspects, the content, the spatial reference, the portrayal, distribution, and other properties of digital geographic data and services. ISO 19113:2002 establishes the principles for describing the quality of geographic data and specifies components for reporting quality information. It also provides an approach to organizing information about data quality. ISO 19114:2003 provides a framework of procedures for determining and evaluating quality that is applicable to digital geographic datasets, consistent with the data quality principles defined in ISO 19113. It also establishes a framework for evaluating and reporting data quality results, either as part of data quality metadata only, or also as a quality evaluation report. ISO/TS 19138:2006 defines a set of data quality measures. These can be used when reporting data quality for the data quality sub-elements identified in ISO 19113. Multiple measures are defined for each data quality sub-element, and the choice of which to use will depend on the type of data and its intended purpose.

The Spatial Data Transfer Standard (SDTS[7]), is a robust way of transferring earth-referenced spatial data between dissimilar computer systems with the potential for no information loss. It is a transfer standard that embraces the philosophy of self-contained transfers, i.e. spatial data, attribute, georeferencing, data quality report, data dictionary, and other supporting metadata all included in the transfer.

Other organisations that provide Geospatial Data Quality standards includes: the Open GIS Consortium (OGC[8]), the European Committee for Standardization Technical Committee 287 (CEN/TC 287[9]), Canadian Geospatial Data Infrastructure (CGDI[10]), National Spatial Data Infrastructure (NSDI[11]), Digital Geographic Information Exchange Standards (DIGEST[12]), and Spatial Archiving and Interchange Format (SAIF[13]).

2.2 Geospatial Data Quality Scientific Contributions

[23] explore measurement standards on the quality of open Geospatial data with the purpose of optimizing data curation and enhancing information systems in support of scientific research and related activities. A set of dimensions for data quality measurement is proposed in order to develop appropriate metrics. Pipino et al. [19] describe principles that can help organizations develop usable spatial data quality metrics. [1] addressed

[2] ISO/TC 211: http://www.isotc211.org/
[3] ISO 19115-1:2014: http://goo.gl/y4nbTG
[4] ISO 19113:2002: http://goo.gl/QUApX8
[5] ISO 19114:2003: http://goo.gl/A5Db23
[6] ISO/TS 19138: http://goo.gl/aYxyjC
[7] SDTS: http://www.fgdc.gov/metadata/csdgm/02.html
[8] OGC: http://www.opengeospatial.org/
[9] CEN/TC 287: http://goo.gl/DfW0Bn
[10] CGDI: http://goo.gl/MHbpI9
[11] NSDI: http://www.fgdc.gov/nsdi/nsdi.html
[12] DIGEST: https://www.dgiwg.org/digest/
[13] SAIF: http://archive.ilmb.gov.bc.ca/crgb/pba/saif/

some of the issues in spatial Data Quality, especially the need to incorporate visualisation of Data Quality into graphics and maps. [22] provides a detailed discussion of various spatial Data Quality components. Boin et al. [4] question whether or not the quality information that is typically provided in such spatial metadata is actually effective. This research employs qualitative research approaches to explore how users of spatial data determine the quality of a dataset. Consumer feedback emails and semi-structured interviews have been analyzed to discover the perceptions, actions and goals of individual data consumers from a range of professional backgrounds. Multidimensional User Manual (MUM) [9] allows the management of geospatial data quality and the communication of the quality information using indicators that can be analysed at different levels of detail. [2] presents GIS related issues and a set of spatical Data Quality metrics. Caprioli et al. [5, 6] examined the quality approaches in GIS contexts and some of most meaningful spatial Data Quality standards. Quality Information Management Model (QIMM) [8] allows a user to easily and rapidly navigate into the quality information using a Spatial On-Line Analytical Processing (SOLAP) client-tied to its GIS application. [11] discusses the implementation of various ISO spatial Data Quality standards. Mooney et al. [18] provide measures of quality for Open Street Map (OSM) which operate in an unsupervised manner without reference to a "trusted" source of ground-truth data. Other recent contributions [3, 16, 17] discusses the various spatial Data Quality metrics and their implementations.

In the next sub-section, based on above scientific contributions and standards, we present the set of spatial Data Quality metrics and rank them according to the number of citations.

2.3 Geospatial Data Quality Metrics

Accuracy: Accuracy is critical Geospatial Data Quality metric for location information services. Georeferencing helps align spatial entities to an image, which requires accurate transformation of data. It includes: *positional accuracy, attribute accuracy*, and *temporal accuracy*. *Positional accuracy* refers to the accuracy of the spatial component (e.g., point, line) of a database, *attribute accuracy* refers to the accuracy of thematic component (e.g., surface), and *temporal accuracy* refers to the agreement between encoded and actual temporal coordinates.
Consistency: It refers to the extent to which data is consistent and presented in same format. Data values from various sources referring to the same geospatial feature need to be consistent.
Completeness: The extent to which data is not missing and is of sufficient breadth and depth for the task at hand. Key data fields and the other types of data supporting spatial analysis and presentation should be associated with each geospatial object to ensure usability and appropriateness of data values.
Reputation: It refers to the extent to which data is highly regarded in terms of its source or content. Sources of geospatial data indicate quality. Authoritative sources can come from research institutions and the government. Examples include population counts, census tracts, and satellite imagery provided by the government.
Currency: It refers to the extent to which data is sufficiently up-to-date for the task at hand. We need to make sure that changes to geospatial data are updated, both on maps and in text. For example, even personally used GPS devices need frequent updates for current road conditions.
Objectivity: It refers to the extent to which data is unbiased, unprejudiced, and impartial.
Relevancy: It refers to the extent to which data is applicable and helpful for the task in hand.
Security: It refers to the extent to which data is restricted appropriately to maintain its security.
Accessibility: It refers to the extent to which data is available or easily and quickly retrievable. When open access is taken into consideration, data accessibility and understandability have a new meaning. For geospatial data, this can also mean an evaluation on the sources of availability, e.g., satellite imagery from government agencies like NASA versus imagery from commercial companies.
Sufficiency: It refers to the extent to which the volume of data is sufficient for the task at hand.
Compatibility: It refers to the extend to which one device data can be used in another device. Newer devices

can retrieve, read, understand and interpret the data retrieved from older devices; or vice versa. Data also need to be compatible between different technological systems.
Discover-ability: It refers to the extent to which data is in appropriate language, symbols, units, and the definitions are clear. It is also known as interoperability, to support metadata harvesting.
Integrity: It refers to the extent to which data is regarded as true and credible. It is also known as believability, i.e., the user should believe that the data comes from credible source.
Legibility: Both machines and humans should be able to catch, read, interpret and use available data. An example is the recognition of satellite imagery where resolution variations can reflect the effective bit-depth of the sensor, the optical and systemic noise, the altitude of the satellite's orbit, etc. Resolution can have the spatial, spectral, temporal, and radiometric types.
Repurposing: This is more an open access issue. However, reusability helps verify the quality of geospatial data.
Transparency: Transparency in the process of data creation and acquisition can help confirm high standards in data.
Validity: It refers to the extent to which data is reasonable and in correct format. For example, account numbers usually falls within a specific values range, numeric data are all digits, dates always have a valid day, month, and year format.
Verifiability: It refers to the extent to which data is verifiable. Usually, a data is highly verifiable if it followed some professional standards.
Visualisation: It refers to the quality of the visual presentation (e.g, colour, visibility etc.) of data in terms of geospatial maps.
Value-Added: It refers to the extent to which data is beneficial and provides advantages from its use.
Resolution: It refers to minimum size of features that are discernible.
Lineage: It refers to proper documentation of the source materials, explaining method of derivation and transformations applied to the initial data.

Overall, accuracy, completeness, and consistency of the data are the key metrics for any geospatial application in order to produce credible results. Based on the literature, Table 1 shows rank-wise spatial Data Quality metrics.

In the next section, we present CROCUS, a tool developed for measuring accuracy in context of GeoKnow project task T3.5.

Table 1: Ranked list of Spatial Data Quality Metrics.

Rank	Metric	Citation
1	Accuracy	ISO/TC 211, SDTS, [23, 19, 13, 22, 4, 5, 6, 8, 11, 2]
1	Consistency	ISO/TC 211, SDTS, [23, 19, 13, 22, 4, 5, 6, 8, 11, 2]
1	Completeness	ISO/TC 211, SDTS, [23, 19, 13, 22, 4, 5, 6, 8, 11, 2]
2	Lineage	ISO/TC 211, SDTS, [13, 4, 5, 8, 11]
3	Resolution	ISO/TC 211, [19, 2, 5]
4	Currency	[13, 4, 2]
5	Accessibility	[19, 23]
5	Discoverability	[19, 23]
5	Integrity/Believeability	[19, 23]
6	Validity	[19, 23]
7	Objectivity	[19]
7	Relevancy	[19]
7	Security	[19]
7	Sufficiency	[19]
7	Compatibility	[23]
7	Legibility	[23]
7	Repurposing	[23]
7	Transparency	[19]
7	Visualization	[23]
7	Value-Added	[19]
7	Reputation	[19]
7	Verifiability	[19]

3 CROCUS

Many applications require the generation of high-quality data in a short amount of time. A brute-force approach for quality assurance on a given or newly generated data set might be a significant number of domain experts checking the data and defining constraints. However, depending on the size of the given data set, the manual evaluation process by domain experts is time consuming and expensive.

Commonly, a data set is improved and maintained in repeated iteration cycles generally leading to the desired level of quality over time, with the exception of newly inserted or updated instances potentially tending to be error-prone. Hence, the quality of the data set is in danger of being contaminated after each re-import.

From this scenario, we derive the requirements for our data quality evaluation process.

- Our aim is to find singular faults, i.e., unique instance errors, conflicting with other similar (e.g, from same class) instances in the knowledge base.
- The data evaluation process has to be efficient. Due to the size of geospatial LOD datasets, reasoning is infeasible due to performance constraints, but graph-based statistics and clustering methods can still work efficiently.
- This process has to be agnostic of the underlying knowledge base, i.e., it should be independent of the evaluated dataset.

Often, mature ontologies that have grown over years, that have been edited and improved by a large number of processes and people and that have been created by a third party provide the basis for industrial applications (e.g., DBpedia).

Aiming at short time-to-market, industry needs scalable algorithms to detect errors in these data sets. The industry partners in the GeoKnow project, Brox and Unister, are no exception.

Furthermore, the lack of costly domain experts requires non-experts to validate the data before it goes live in a productive system. Resulting knowledge bases may still contain errors, however, they offer a fair trade-off in an iterative production cycle.

CROCUS is a semi-automatic cluster-based ontology data cleansing approach for instance-level errors detection presented in [7] and developed in the context of GeoKnow project task T3.5. It can be configured to find several types of errors in a semi-automatic way, which are afterwards validated by non-expert users called quality raters.

By applying CROCUS' methodology iteratively, resulting ontology data can be safely used in industrial environments. To the best of our knowledge, CROCUS is the first tool tackling error accuracy (intrinsic data quality), completeness (contextual data quality) and consistency (data modelling) at once in a semi-automatic manner reaching high f1-measure on real-world data. Overall, the results presented in [7] show that it is successfully able to detect outliers in synthetic and real-world data and work with different knowledge bases.

4 Method

In this section, we first describe the methodology followed in CROCUS development followed by the evaluation results. The CROCUS installation and running details are given in the next section.

4.1 Step 1: Extraction of target data

To comply with one of the main requirements, we need a standardised extraction of target data to be agnostic of the underlying knowledge base in the first step. SPARQL [20] is a W3C standard to query instance data from Linked Data knowledge bases. The `DESCRIBE` query command is a way to retrieve descriptive data of certain instances. However, this query command depends on the knowledge base vendor and its configuration. To circumvent knowledge base dependence, we use *Concise Bounded Descriptions* (CBD) [21]. Given a resource r and a certain description depth d the CBD works as follows: (1) extract all triples with r as subject and (2) resolve all blank nodes retrieved so far, i.e., for each blank node add every triple containing a blank node with the same identifier as a subject to the description. Finally, CBD repeats these steps d times. CBD configured with $d = 1$ retrieves only triples with r as subject although triples with r as object could contain useful information. Therefore, a rule is added to CBD, i.e., (3) extract all triples with r as object, which is called *Symmetric Concise Bounded Description* (SCDB) [21].

4.2 Step 2: Numeric representation

Second, CROCUS needs to calculate a numeric representation of an instance to facilitate further clustering steps. Metrics are split into three categories:

(1) The simplest metric counts each property (*count*). For example, if a company policy is that a person should have only one registered telephone number. Now, this metric will declare each instance (employee) as an outlier if the telephone number for that instance is greater than one.

(2) For each instance, the range of the resource at a certain property is counted (*range count*). The values assigned to each proerty of the instance should follow the underlying organisation policy. For example, if a university policy is that an undergraduate student can only take undergraduate courses, i.e., he/she is not allowed to be registered in master level course. Now, If there is an undergraduate student who is registered in master level course, this metric should be able to detect such policy violation.

(3) The most general metric transforms each instance into a numeric vector and normalizes it (*numeric*). Since instances created by the SCDB consist of properties with multiple ranges, CROCUS defines the following metrics: (a) numeric properties are taken as is, (b) properties based on strings are converted to a metric by using string length although more sophisticated measures could be used (e.g., n-gram similarities) and (c) object properties are discarded for this metric.

4.3 Step 3: Clustering

As a third step, we apply the *density-based spatial clustering of applications with noise* (DBSCAN) algorithm [12] since it is an efficient algorithm and the order of instances has no influence on the clustering result. DBSCAN clusters instances based on the size of a cluster and the distance between those instances. Thus, DBSCAN has two parameters: ϵ, the distance between two instances, here calculated by the metrics above and $MinPts$, the minimum number of instances needed to form a cluster. If a cluster has less than $MinPts$ instances, they are regarded as outliers.

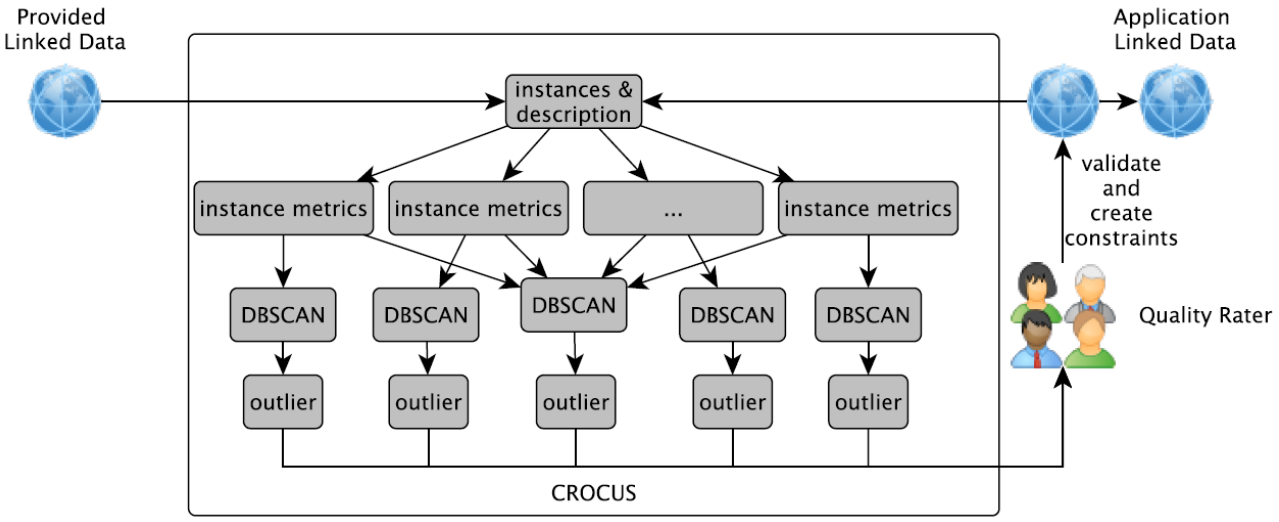

Figure 1: Overview of CROCUS.

4.4 Step 4: Outlier examination

Finally, potentially identified outliers are extracted and given to human quality judges. Based on the revised set of outliers, the algorithm can be adjusted and constraints can be added to the Linked Data knowledge base to prevent repeating discovered errors.

4.5 General Evaluation

We subjected the CROCUS algorithm to a general evaluation described in this section.

LUBM benchmark. First, we used the LUBM benchmark [14] to create a perfectly modelled dataset. This benchmark allows to generate arbitrary knowledge bases themed as university ontology. Our dataset consists of exactly one university and can be downloaded from our project homepage[14].

The LUBM benchmark generates random but error free data. Thus, we add different errors and error types manually for evaluation purposes:

- *completeness of properties (count)* has been tested with CROCUS by adding a second phone number to 20 of 1874 graduate students in the dataset. The edited instances are denoted as I_{count}.
- *semantic correctness of properties (range count)* has been evaluated by adding graduate courses (`Course`) to the course list of 20 non-graduate students ($I_{rangecount}$).
- *numeric correctness of properties (numeric)* was injected by defining that a graduate student has to be younger than a certain age. To test this, 20 graduate students' ($I_{numeric}$) age was replaced with a value bigger than the arbitrary maximum age of any other graduate.

For each set of instances holds: $|I_{count}| = |I_{rangecount}| = |I_{numeric}| = 20$ and additionally $|I_{count} \cap I_{rangecount} \cap I_{numeric}| = 3$. The second equation overcomes a biased evaluation and introduces some realistic noise into the dataset. One of those 3 instances is shown in the listing below:

[14] https://github.com/AKSW/CROCUS

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix ns2: <http://example.org/#> .
@prefix ns3: <http://www.Department6.University0.edu/> .

ns3:GraduateStudent75 a ns2:GraduateStudent ;
   ns2:name "GraduateStudent75" ;
   ns2:undergraduateDegreeFrom <http://www.University467.edu> ;
   ns2:emailAddress "GraduateStudent75@Department6.University0.edu" ;
   ns2:telephone "yyyy-yyyy-yyyy" , "xxx-xxx-xxxx" ;
   ns2:memberOf <http://www.Department6.University0.edu> ;
   ns2:age "63" ;
   ns2:takesCourse ns3:GraduateCourse21 , ns3:Course39 , ns3:
      GraduateCourse26 ;
   ns2:advisor ns3:AssociateProfessor8 .
```

Listing 1: Example of an instance with manually added errors (*in red*).

Results. To evaluate the performance of CROCUS, we used each error type individually on the adjusted LUBM benchmark datasets as well as a combination of all error types on LUBM[15] and the real-world DBpedia subset.

	LUBM								
	count			***range count***			***numeric***		
$MinPts$	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**
2	—	—	—	—	—	—	—	—	—
4	—	—	—	0.49	1.00	0.33	—	—	—
8	—	—	—	0.67	1.00	0.5	—	—	—
10	0.52	1.00	0.35	1.00	1.00	1.00	—	—	—
20	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
30	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
50	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
100	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Table 2: Results of the LUBM benchmark for all three error types.

Table 2 shows the f1-measure (F1), precision (P) and recall (R) for each error type. For some values of $MinPts$ it is infeasible to calculate cluster since DBSCAN generates only clusters but is unable to detect outlier. CROCUS is able to detect the outliers with a 1.00 f1-measure as soon as the correct size of $MinPts$ is found.

[15]The datasets can also be found on our project homepage.

5 CROCUS Tutorial

The previous section 4 describes the CROCUS algorithm in general and its application to any linked data set. This section gives a hands-on tutorial on how to use CROCUS on data with geospatial annotations, namely, LinkedGeoData [16], which uses the information collected by the OpenStreetMap[17] project and makes it available as geospatial RDF knowledge base.

We have post-processed CROCUS results (in XML RDF) into RDF DataCubes[18] which can be visualised with CubeViz[19]. We first explain installing and running CROCUS followed by the explanation of the DataCubes results generation and their CubeViz visualization.

5.1 CROCUS Services and Setup

CROCUS is exposed as a web-service configurable by a JSON object. The results of the algorithm are provided in RDF.

CROCUS setup involves the following steps:

1. Install Apache Tomcat[20] server (tested with version 6 and 7) on your computer.
2. Download CROCUS War archive from http://goo.gl/j1FqRs and paste it into "/Tomcat/webapps" folder of Apace Tomcat installation directory.
3. Check out the CROCUS DataCube generation Java project from `https://github.com/GeoKnow/GeoQuality` into any Java execution environment such as Eclipse.
4. Go to <tomcat-url>/ontologymetrics/ and get a session key to be used later in when running CROCUS.

5.2 Running CROCUS

Open the Client.java file of the CROCUS DataCube generation Java project and provide the following parameters as input to the CROCUS configuration:

- **Session Key:** The key obtained from <tomcat-url>/ontologymetrics/ as explained in the previous sub-section
- **EndPoint Url:** The URL of the SPARQL endpoint hosting the RDF data for which data quality needs to be measured.
- **Class Name:** The name of the class whose instance quality needs to be measured.

[16] http://linkedgeodata.org/
[17] http://www.openstreetmap.org/
[18] `http://www.w3.org/TR/vocab-data-cube/`
[19] `http://cubeviz.aksw.org/`
[20] `http://tomcat.apache.org/`

Table 3: DataCube 1 sample results. Prefix gdo = http://linkedgeodata.org/ontology/

Type	Class	TimeStamp	InstanceCount
Outlier	gdo:ReceptionArea	Sat Jun 28 19:12:19 CEST 2014	10
Normal	gdo:ReceptionArea	Sat Jun 28 19:12:19 CEST 2014	94

```
1 prefix gk-dim: <http://www.geoknow.eu/properties/>
2 prefix sdmx-measure: <http://purl.org/linked-data/sdmx/2009/measure#>
3 prefix qb: <http://purl.org/linked-data/cube#>
4 SELECT ?Type ?class ?TimeStamp ?InstanceCount
5 WHERE
6 {
7 ?obsrv qb:dataSet <http://www.geoknow.eu/dataset/ds1> ;
8     gk-dim:InstanceType ?Type;
9     gk-dim:Class ?class ;
10    gk-dim:TimeStamp ?TimeStamp ;
11    sdmx-measure:InstanceCount ?InstanceCount ;
12    a qb:Observation .
13 }
```

Listing 2: SPARQL query to print DataCube 1

- **Properties Set:** The set of properties of the the class for which instance-level quality is desired. This parameter can be null, indicating all the class properties needs to be considered.

After the execution is complete, a DataCube file named "DataCubeResults.n3" will be generated which contains the the required quality results in the form of DataCubes (explained in next section).

5.3 DataCube Results

The DataCube results file consists of three DataCubes:

1. DataCube 1: This DataCube has *IsOutlier, Class, TimeStamp, InstanceCount* as observations which show how many of the instances of a class are outliers and how many of them are normal instances along with the time stamp. A sample resulting DataCube is shown in Table 3. The corresponding query to print this table is shown in Listing 2. This DataCube always contains only two rows, i.e., one shows the total number of normal instances and other shows the total number of outlier instances.

2. DataCube 2: This DataCube includes *Instance, IsOutlier, Class, TimeStamp, PropsCount* as observations and show whether a particular instance of a class is outlier or not. A sample resulting DataCube is shown in Table 4. The corresponding query to print this table is shown in Listing 3.

3. DataCube 3: This DataCube has *Property, Class, TimeStamp, ObjectsCount, SubjectsCount* as observations and show various statistics about each distinct properties. A sample resulting DataCube is shown in Table 5. The corresponding query to print this table is shown in Listing 4.

```
prefix gk-dim: <http://www.geoknow.eu/properties/>
prefix sdmx-measure: <http://purl.org/linked-data/sdmx/2009/measure#>
prefix qb: <http://purl.org/linked-data/cube#>
SELECT ?Instance ?Type ?Class ?TimeStamp ?PropsCount
WHERE
{
?obsr qb:dataSet <http://www.geoknow.eu/dataset/ds2> ;
        gk-dim:Instance ?Instance;
    gk-dim:InstanceType ?Type ;
    gk-dim:Class ?Class ;
    gk-dim:TimeStamp ?TimeStamp ;
    sdmx-measure:PropsCount ?PropsCount ;
    a qb:Observation .
}
```

Listing 3: SPARQL query to print DataCube 2

```
prefix gk-dim: <http://www.geoknow.eu/properties/>
prefix sdmx-measure: <http://purl.org/linked-data/sdmx/2009/measure#>
prefix qb: <http://purl.org/linked-data/cube#>
SELECT DISTINCT ?property ?class ?timeStamp ?objectsCount ?subjectsCount
WHERE
{
?obsr qb:dataSet <http://www.geoknow.eu/dataset/ds3> ;
        gk-dim:Property ?property ;
    gk-dim:Class ?class ;
    gk-dim:TimeStamp ?timeStamp ;
    sdmx-measure:SubjectsCount ?objectsCount ;
    sdmx-measure:ObjectsCount ?subjectsCount ;
    a qb:Observation .
}
```

Listing 4: SPARQL query to print DataCube 3

Table 4: DataCube 2 sample results. Prefix gdo = http://linkedgeodata.org/ontology/, Prefix gdt = http://linkedgeodata.org/triplify/

Instance	Type	Class	TimeStamp	PropsCount
gdt:node1039036534	Normal	gdo:ReceptionArea	Sat Jun 28 19:12:19 CEST 2014	13
gdt:node1368677765	Normal	gdo:ReceptionArea	Sat Jun 28 19:12:19 CEST 2014	12
gdt:node539387437	Outlier	gdo:ReceptionArea	Sat Jun 28 19:12:19 CEST 2014	12
gdt:node539388589	Outlier	gdo:ReceptionArea	Sat Jun 28 19:12:19 CEST 2014	11
gdt:node583233731	Normal	gdo:ReceptionArea	Sat Jun 28 19:12:19 CEST 2014	13

Table 5: DataCube 3 sample results. Prefix Prefix gdo = http://linkedgeodata.org/ontology/

Property	Class	TimeStamp	ObjectsCount	SubjectsCount
gdo:building	gdo:ReceptionArea	Sat Jun 28 19:12:19 CEST 2014	1	1
gdo:layer	gdo:ReceptionArea	Sat Jun 28 19:12:19 CEST 2014	1	1
gdo:localName	gdo:ReceptionArea	Sat Jun 28 19:12:19 CEST 2014	59	60
gdo:changeset	gdo:ReceptionArea	Sat Jun 28 19:12:19 CEST 2014	34	104
gdo:information	gdo:ReceptionArea	Sat Jun 28 19:12:19 CEST 2014	1	1

5.4 DataCubes Transformation Process

According to RDF DataCube vocabulary[21], a well-formed representation of the statistical data comprise:

Datasets Definition: Listing 5 shows the datasets used in CROCUS results to DataCubes transformation. We refer each DataCube (explained above) as separate dataset. Therefore, we define 3 datasets as shown in Listing 5. Each dataset has a data structure definition and is explained next.

Data Structure Definition: Listing 6 shows the structure of each of the data sets. Our first DataCube contains 4 components while the second and third DataCube contains 5 components each. Each dataset component can be a dimension, property or measure and is explained next.

Component Specification: Listing 7 and Listing 8 shows the specification of each of the dataset component defined in Listing 6. Our first DataCube contains 3 dimensions, i.e., *InstanceType, Class, TimeStamp* and one measure, i.e., *InstanceCount*. Our second DataCube contains 4 dimensions, i.e., *Instance, InstanceType, Class, TimeStamp* and one measure, i.e., *PropsCount*. While our third DataCube contains 3 dimensions, i.e., *Property, Class, TimeStamp* and 2 measures, i.e., *ObjectsCount, SubjectsCount*.

Writting Observations: Since CROCUS results are in RDF, the required observations can be obtained by simple SPARQL query execution. It is important to note that the observations for third DataCube is directly obtained from the SPARQL endpoint for which the Data Quality needs to be observed.

[21] http://www.w3.org/TR/vocab-data-cube/

```
<http://www.geoknow.eu/dataset/ds1> a qb:DataSet ;
    dcterms:publisher "AKSW, GeoKnow" ;
     rdfs:label "DataCube1 Results" ;
     qb:structure <http://www.geoknow.eu/data-cube/dsd1> ;
     dcterms:date "Sat Jun 28 20:13:42 CEST 2014".

<http://www.geoknow.eu/dataset/ds2> a qb:DataSet ;
     dcterms:publisher "AKSW, GeoKnow" ;
    rdfs:label "DataCube2 Results" ;
    qb:structure <http://www.geoknow.eu/data-cube/dsd2> ;
    dcterms:date "Sat Jun 28 20:13:42 CEST 2014".

<http://www.geoknow.eu/dataset/ds3> a qb:DataSet ;
    dcterms:publisher "AKSW, GeoKnow" ;
    rdfs:label "DataCube3 Results" ;
    qb:structure <http://www.geoknow.eu/data-cube/dsd3> ;
    dcterms:date "Sat Jun 28 20:13:42 CEST 2014".
```

Listing 5: DataCubes generation: Datasets definitions

```
<http://www.geoknow.eu/data-cube/dsd1> a qb:DataStructureDefinition ;
     rdfs:label    "A Data Structure Definition"@en ;
     rdfs:comment  "Definition for DataCube1" ;
qb:component <http://www.geoknow.eu/data-cube/dsd1/c1>,
     <http://www.geoknow.eu/data-cube/dsd1/c2>,
     <http://www.geoknow.eu/data-cube/dsd1/c3>,
     <http://www.geoknow.eu/data-cube/dsd1/c4> .

<http://www.geoknow.eu/data-cube/dsd2> a qb:DataStructureDefinition ;
     rdfs:label    "A Data Structure Definition"@en ;
     rdfs:comment  "Definition for DataCube2" ;
qb:component <http://www.geoknow.eu/data-cube/dsd2/c1>,
     <http://www.geoknow.eu/data-cube/dsd2/c2>,
     <http://www.geoknow.eu/data-cube/dsd2/c3>,
     <http://www.geoknow.eu/data-cube/dsd2/c4>,
     <http://www.geoknow.eu/data-cube/dsd2/c5> .

<http://www.geoknow.eu/data-cube/dsd3> a qb:DataStructureDefinition ;
     rdfs:label    "A Data Structure Definition"@en ;
     rdfs:comment  "Definition for DataCube3" ;
qb:component <http://www.geoknow.eu/data-cube/dsd3/c1>,
     <http://www.geoknow.eu/data-cube/dsd3/c2>,
     <http://www.geoknow.eu/data-cube/dsd3/c3>,
     <http://www.geoknow.eu/data-cube/dsd3/c4>,
     <http://www.geoknow.eu/data-cube/dsd3/c5> .
```

Listing 6: DataCubes generation: Data structure definitions

```
1  |
2  |<http://www.geoknow.eu/data-cube/dsd1/c1> a qb:ComponentSpecification ;
3  |     qb:dimension gk-dim:InstanceType .
4  |<http://www.geoknow.eu/data-cube/dsd1/c2> a qb:ComponentSpecification ;
5  |     qb:dimension gk-dim:Class .
6  |<http://www.geoknow.eu/data-cube/dsd1/c3> a qb:ComponentSpecification ;
7  |     qb:dimension gk-dim:TimeStamp .
8  |<http://www.geoknow.eu/data-cube/dsd1/c4> a qb:ComponentSpecification ;
9  |     qb:measure sdmx-measure:InstanceCount .
10 |
11 |<http://www.geoknow.eu/data-cube/dsd2/c1> a qb:ComponentSpecification ;
12 |     qb:dimension gk-dim:Instance .
13 |<http://www.geoknow.eu/data-cube/dsd2/c2> a qb:ComponentSpecification ;
14 |     qb:dimension gk-dim:InstanceType .
15 |<http://www.geoknow.eu/data-cube/dsd2/c3> a qb:ComponentSpecification ;
16 |     qb:dimension gk-dim:Class .
17 |<http://www.geoknow.eu/data-cube/dsd2/c4> a qb:ComponentSpecification ;
18 |     qb:dimension gk-dim:TimeStamp .
19 |<http://www.geoknow.eu/data-cube/dsd2/c5> a qb:ComponentSpecification ;
20 |     qb:measure sdmx-measure:PropsCount .
21 |
22 |<http://www.geoknow.eu/data-cube/dsd3/c1> a qb:ComponentSpecification ;
23 |     qb:dimension gk-dim:Property .
24 |<http://www.geoknow.eu/data-cube/dsd3/c2> a qb:ComponentSpecification ;
25 |     qb:dimension gk-dim:Class .
26 |<http://www.geoknow.eu/data-cube/dsd3/c3> a qb:ComponentSpecification ;
27 |     qb:dimension gk-dim:TimeStamp .
28 |<http://www.geoknow.eu/data-cube/dsd3/c4> a qb:ComponentSpecification ;
29 |     qb:measure sdmx-measure:ObjectsCount .
30 |<http://www.geoknow.eu/data-cube/dsd3/c5> a qb:ComponentSpecification ;
31 |     qb:measure sdmx-measure:SubjectsCount .
```

Listing 7: DataCubes generation: Component specifications

```
1  |gk-dim:InstanceType a qb:DimensionProperty ;
2  |        rdfs:label    "Normal or Outlier instance"@en .
3  |gk-dim:Class a qb:DimensionProperty ;
4  |        rdfs:label    "class of instance"@en .
5  |gk-dim:TimeStamp a qb:DimensionProperty ;
6  |        rdfs:label    "Time Stamp"@en .
7  |sdmx-measure:InstaceCount  a qb:MeasureProperty ;
8  |        rdfs:label    "Instance Count"@en .
9  |gk-dim:Instance a qb:DimensionProperty ;
10 |        rdfs:label    "Instance"@en .
11 |sdmx-measure:PropsCount  a qb:MeasureProperty ;
12 |        rdfs:label    "Properties Count"@en .
13 |gk-dim:Property a qb:DimensionProperty ;
14 |        rdfs:label    "Property name"@en .
15 |sdmx-measure:ObjectsCount  a qb:DimensionProperty ;
16 |        rdfs:label    "Distinct Objects Count"@en .
17 |sdmx-measure:SubjectsCount a qb:DimensionProperty ;
18 |        rdfs:label    "Distinct Subjects Count"@en .
```

Listing 8: DataCubes generation: Dimensions Units and Measures

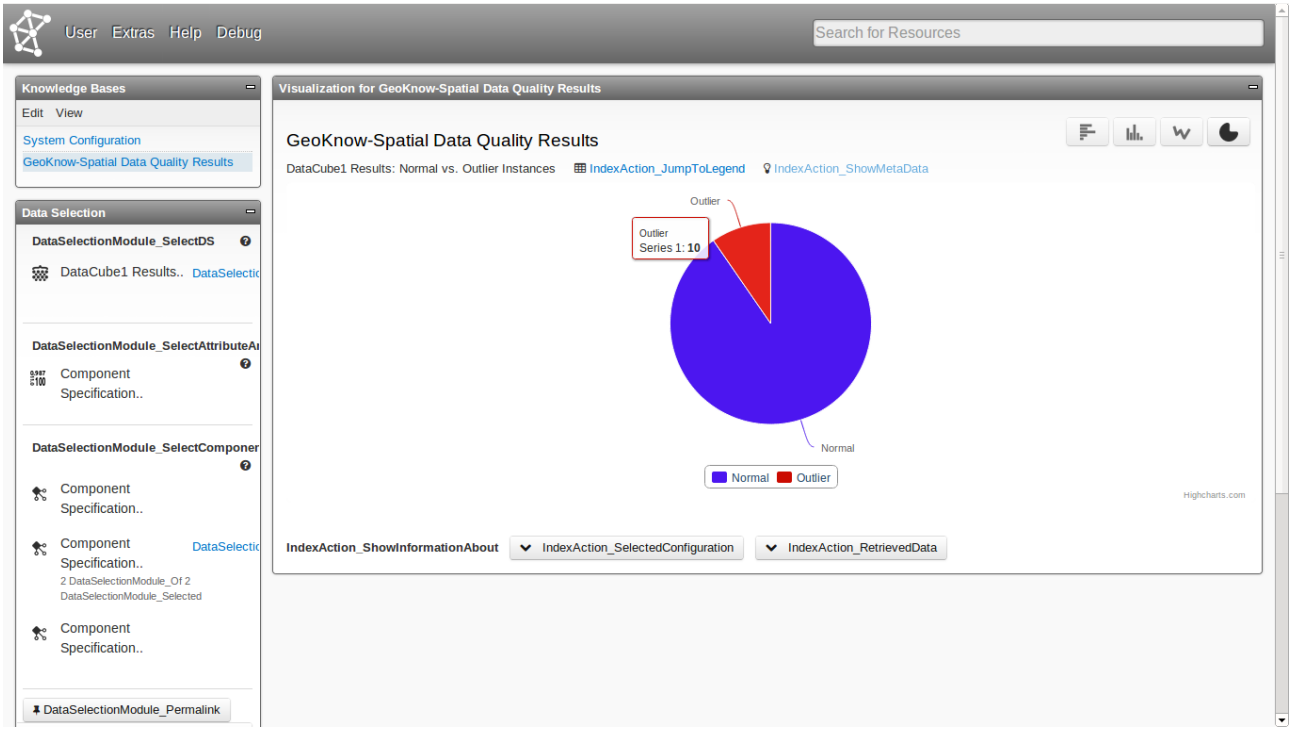

Figure 2: DataCube 1 CubeViz Visualization Linked Geo Data Class ReceptionArea

5.5 CubeViz Visualization

CubeViz[22] represents the statistical dataset to be visualized as a facetted based browsing component (located on the left side of the Figures 2-5). This component enables the users to select interesting parts of the dataset. After selection the user can proceed while clicking on the button Update Selection / Update Chart. One can instantly start using CubeViz while clicking on the button Start CubeViz above. As a result CubeViz processes a chart according to user's selection. The current version of CubeViz processes basic chart types - such as line, bar and pie chart facilitating the exploration of up to two statistical dimensions in a data structure. Further usage information can be found at http://cubeviz.aksw.org/page/howto.

Figure 2 shows the CubeViz visualization of our first DataCube for class http://linkedgeodata.org/ontology/ReceptionArea of SPARQL endpoint http://linkedgeodata.org/sparql. This class contains a total of 104 instances out of which 10 instances are outliers. Figure 3 and Figure 4 show the CubeViz visualization of our second DataCube for the same class name and SPARQL endpoint. These visualizations show whether a particular selected instance is normal or outlier. Further, it also shows the number of distinct properties of the instance. It is important to note that we have only selected 5 out of 104 instances to be visualized. CubeViz gives the flexibility to select any number of instances. Figure 5 shows the number of distinct objects for each distinct property of the class.

[22] http://cubeviz.aksw.org/

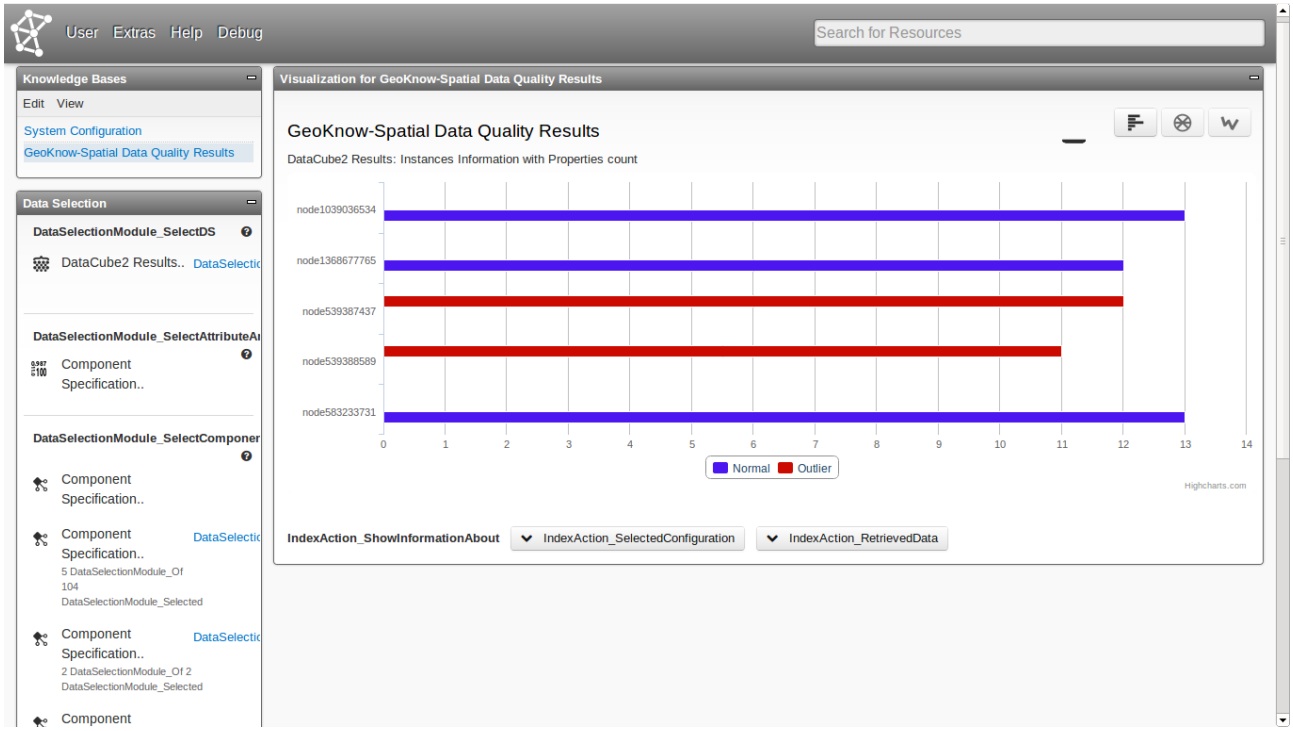

Figure 3: DataCube 2 CubeViz Visualization of the selected instances of Linked Geo Data Class ReceptionArea

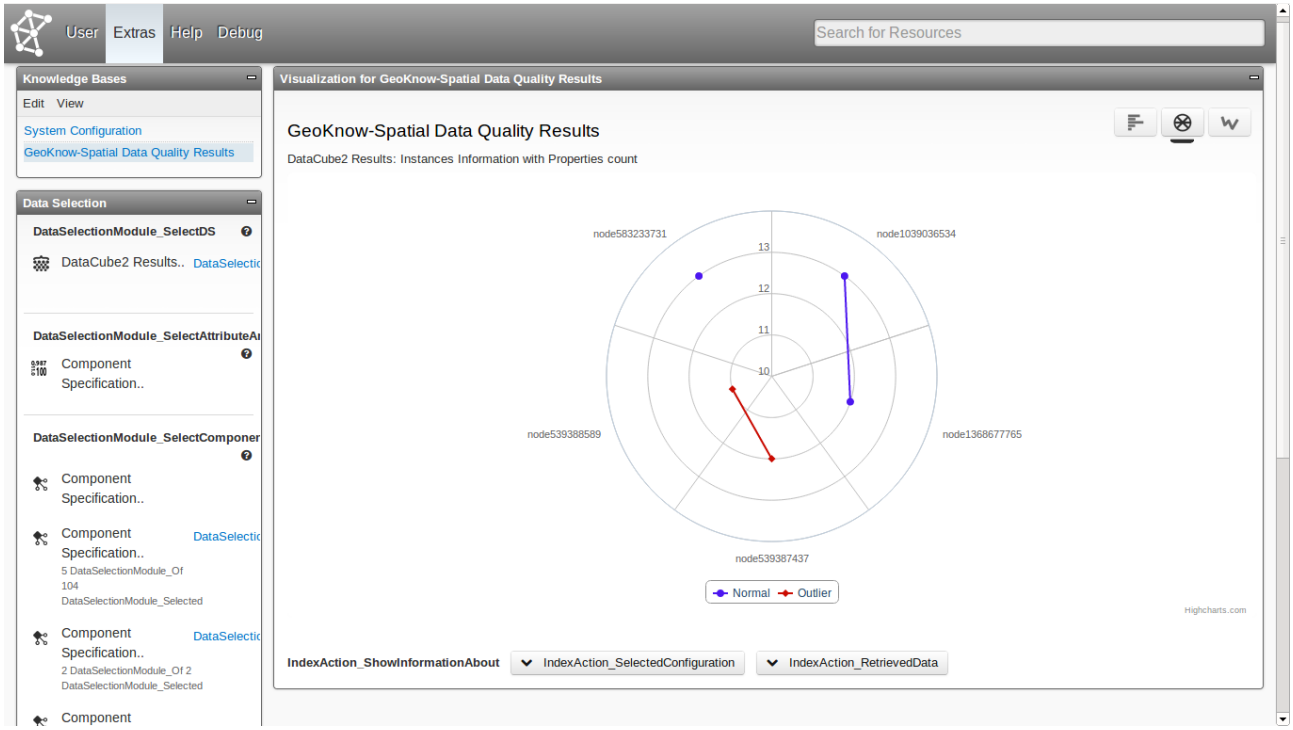

Figure 4: DataCube 2 CubeViz Visualization of the selected instances of Linked Geo Data Class ReceptionArea

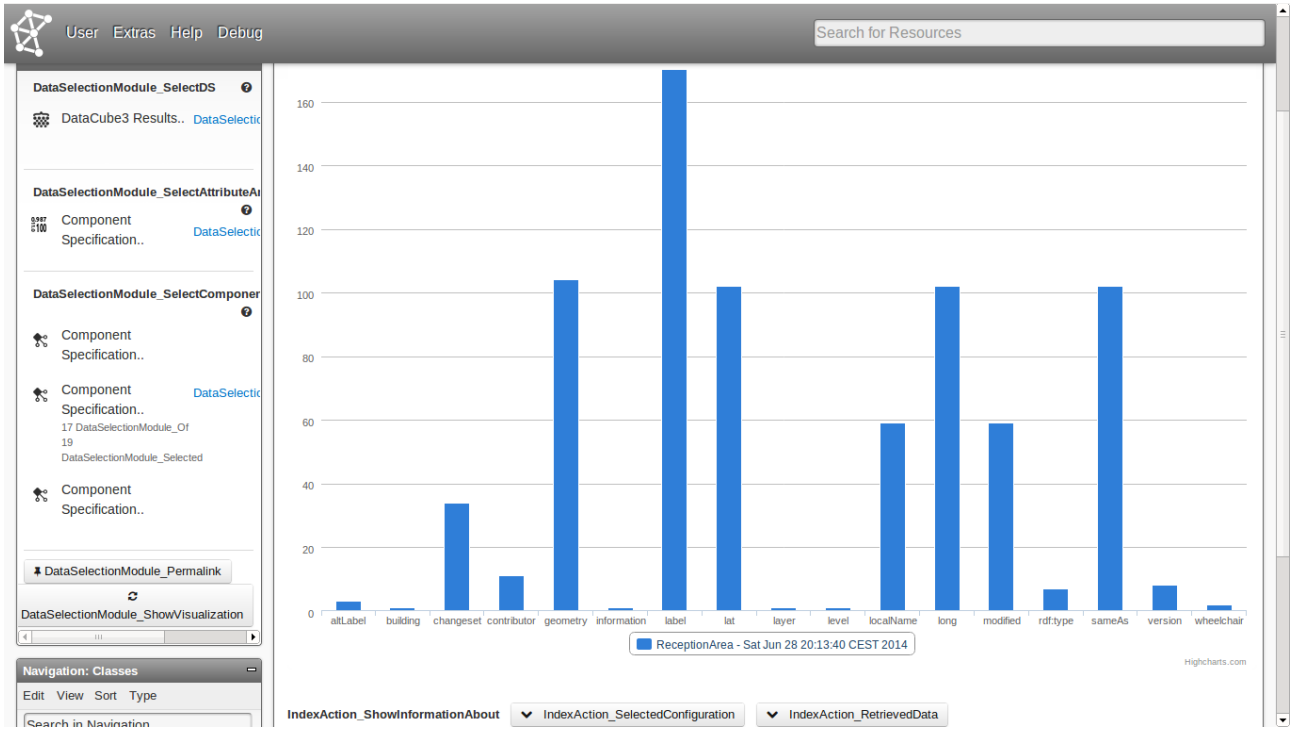

Figure 5: DataCube 3 CubeViz Visualization of the number of distinct objects of different properties of Linked Geo Data Class ReceptionArea

6 Conclusion and Future Work

In this deliverable we presented an initial report on Geospatial Data Quality assessment. We first presented a survey of Geopspatial Data Quality standards and scientific contributions followed by the ranked list of different metrics used for measuring spatial Data Quality. The ranked list shows that *Accuracy*, *Consistency*, and *Completeness* are top ranked metrics. CROCUS is presented in the next section which produces three types of DataCubes, where the first DataCube refers to the *Accuracy*, second and third DataCube addresses the *Completeness* and *Consistency* of spatial data.

In the final deliverable, CROCUS will be extended to address other linked Geo Data related metrics such as *Lineage, Currency, Accessibility, Validity* etc. Further, CROCUS DataCubes results will be automatically pipellined to CubeViz interface in order to enable on-the-fly visualization of generated results.

References

[1] James Abello, Panos M Pardalos, and Mauricio GC Resende. *Handbook of massive data sets*, volume 4. Springer, 2002.

[2] Parmenter Barbara. Data qaulity for gis. In *Report Tuft University*, 2007.

[3] Enrico Bertini, Andrada Tatu, and Daniel Keim. Quality metrics in high-dimensional data visualization: an overview and systematization. *Visualization and Computer Graphics, IEEE Transactions on*, 17(12):2203–2212, 2011.

[4] Anna T Boin and Gary J Hunter. What communicates quality to the spatial data consumer. In *Proceedings of the 7th international symposium on spatial data quality (ISSDQ 2007), Enschede, The Netherlands*, 2007.

[5] M Caprioli, A Scognamiglio, G Strisciuglio, and E Tarantino. Rules and standards for spatial data quality in gis environments. In *Proc. 21st Int. Cartographic Conf. Durban, South Africa 10–16 August 2003*, 2003.

[6] M Caprioli and E Tarantino. Standards and quality in gis contexts. *Multidimensional Appraoches and new Concepts in SIM, Paris*, 2003.

[7] Didier Cherix, Ricardo Usbeck, Andreas Both, and Jens Lehmann. Crocus: Cluster-based ontology data cleansing. In *Proceedings of the 2nd International Workshop on Semantic Web Enterprise Adoption and Best Practice*, 2014.

[8] Rodolphe Devillers, Yvan Bédard, and Robert Jeansoulin. Multidimensional management of geospatial data quality information for its dynamic use within gis. *Photogrammetric Engineering & Remote Sensing*, 71(2):205–215, 2005.

[9] Rodolphe Devillers, Marc Gervais, Yvan Bédard, and Robert Jeansoulin. Spatial data quality: from metadata to quality indicators and contextual end-user manual. In *OEEPE/ISPRS Joint Workshop on Spatial Data Quality Management*, pages 21–22, 2002.

[10] Edward. Spatial data quality and transportation applications. In *5th FIG Regional Conference*. 2006.

[11] GIS ESRI's. Implementing iso data quality standards using esri's gis data reviewer. 2004.

[12] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, volume 96, pages 226–231, 1996.

[13] Michael F Goodchild and Keith C Clarke. Data quality in massive data sets. In *Handbook of massive data sets*, pages 643–659. Springer, 2002.

[14] Yuanbo Guo, Zhengxiang Pan, and Jeff Heflin. LUBM: A benchmark for OWL knowledge base systems. *Web Semantics: Science, Services and Agents on the World Wide Web*, 3(2–3):158 – 182, 2005.

[15] Shirlee Knight and Janice Burn. Developing a framework for assessing information quality on the world wide web. *Informing Science*, 8, 2005.

[16] Weisi Lin and C-C Jay Kuo. Perceptual visual quality metrics: A survey. *Journal of Visual Communication and Image Representation*, 22(4):297–312, 2011.

[17] Hantao Liu and Ingrid Heynderickx. Visual attention in objective image quality assessment: based on eye-tracking data. *Circuits and Systems for Video Technology, IEEE Transactions on*, 21(7):971–982, 2011.

[18] Peter Mooney, Padraig Corcoran, and Adam C Winstanley. Towards quality metrics for openstreetmap. In *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 514–517. ACM, 2010.

[19] Leo L Pipino, Yang W Lee, and Richard Y Wang. Data quality assessment. *Communications of the ACM*, 45(4):211–218, 2002.

[20] Bastian Quilitz and Ulf Leser. 5021:524–538, 2008.

[21] Patrick Stickler. Cbd-concise bounded description. *W3C Member Submission*, 3, 2005.

[22] Howard Veregin. Data quality parameters. *Geographical information systems*, 1:177–189, 1999.

[23] Jingfeng Xia. Metrics to measure open geospatial data quality in. *Issues Sci. Technol. Librar*, 2012.