Automated quality evaluation

In the context of spatial data infrastructure

By:
Amin Mobasher
(mobasher23795@itc.nl)

December 2010
MOTIVATION AND PROBLEM DEFINITION

Growth of SDI in Geo-Informatics:

Important outcomes of SDI:

• Connect geo-service systems
• Interoperability
• Harmonization
• Data sharing

Problem definition:

To find different aspects of spatial data quality that can be automatically evaluated.
PROBLEM DEFINITION AND RESEARCH OBJECTIVES

Research hypothesis:
An automated quality evaluation web service for spatial data can facilitate the process of spatial data quality evaluation in SDI.

Research objectives:
• To analyze the selected quality elements suitable for automated quality evaluation.
• To design the process for evaluating spatial data quality in web services.
• To Study about available standards of spatial web services.
• To design a web service that automatically evaluates different aspects of quality in spatial data on the Internet.
• To validate the designed service.
DEFINITION OF QUALITY AND SPATIAL DATA QUALITY

Quality is:
- conditional
- Fully subjective
- Different requirements = different understanding

Spatial data quality is the degree which dataset meets the requirements of a specific user.

“Spatial data quality = fitness for use”
Spatial data quality evaluation is defined as:
matching the user requirements against the dataset itself, to check its fitness for use.

Different types of users:
• Human: expert, naive
• Services (e.g. WFS)

Different types of quality evaluation:
• Non-automated
• Semi-automated
• Automated
AUTOMATED QUALITY EVALUATION

Main characteristics:
• User un-related
• Process-based
• Computer algorithm (service)
Main components of user requirements:

• Data quality element
• Data quality sub-element
• Data quality scope
• Data quality measure
• Data quality evaluation methods
• Data quality result
Suitable data quality elements and sub-elements for automated/semi-automated check:

- Logical consistency: conceptual consistency
- Logical consistency: domain consistency
- Logical consistency: format consistency
- Logical consistency: topological consistency
- Completeness: omission
- Completeness: commission
- Temporal accuracy: temporal validity
**USER REQUIREMENTS**

Scope is defined as a smaller grouping of data located physically within the dataset sharing common characteristics

- Spatial extent: defined by combination of two latitude and longitude (e.g. every item bounded by latitudes(52.183,52.241) and longitudes(6.847,6.928))

- Object-based scope: defined by specific objects and specific attributes of dataset (e.g. roads)

- Complex scope: combination of both.
Data quality measure is defined simply as evaluation of a data quality sub-element. (e.g. the percentage of the values of an attribute that are correct)

Counting related data quality basic measures:
- Error indicator
- Correctness indicator
- Error count
- Correct items count
- Error rate
- Correct items rate
Two types of quality evaluation methods:

- **Direct:**
  - Internal: internal reference information used.
  - External: external reference information used.

- **Indirect:** methods which use external knowledge for quality evaluation
USER REQUIREMENTS

Data quality result:
The value(s) resulting from applying a data quality measure, or the outcome of comparing the obtained value against a quality conformance level.

Quality conformance level:
A threshold value for data quality result, used to determine how well a dataset meets the requirements.

Data quality result can be reported as:
• Pass/fail
• Quantitative value
Figure 1. Expanded version of “Define User Requirements” sub-process in BPMN.
Domain consistency is defined as adherence of values to the value domains.

- Field type (e.g. integer, string, etc.)

- Domain type:
  - Range domain (e.g. age: 1-100)
  - Coded domain (e.g. road types: driveway, highway, etc.)
QUALITY EVALUATION PROCEDURE

For performing coded domain check:

- **Semi-automated evaluation:**
  - Scope
  - The list of un-desired values
  - Conformance level

- **Automated evaluation:**
  - The scope is equal to the whole dataset, by default.
  - The “null” values are treated as inconsistency, by default.
  - Conformance level value is equal to zero, by default.
QUALITY EVALUATION PROCEDURE

Figure 2. Expanded version of “Domain Type Check” sub-process in BPMN.
QUALITY EVALUATION REPORTING

- Quality evaluation results can be reported as:
  
  - Metadata:
    Based on ISO19115 there exist several elements suitable for recording data quality information. The model uses appropriate elements to report the results of quality evaluation as metadata.

  - Report:
    - Quality evaluation measure
    - Quality evaluation method
    - Quality evaluation parameters
QUALITY EVALUATION REPORTING

<xml version="1.0" ?>
<!-- This is the Metadata file created and used by the automated quality evaluation web service -->
- <Metadata>
  - <Data>
    <Dataset_Name>sample.gml</Dataset_Name>
  </Data>
- <DataQuality>
  - <dqScope>
    <scpType>1</scpType>
    <scpExtent>Whole dataset</scpExtent>
    <exDesc>The whole features and their attributes are evaluated</exDesc>
  </dqScope>
  - <dqReport>
    <eleTypeCode>002</eleTypeCode>
    <subEleCode>002</subEleCode>
    <addSubEle>Domain Consistency</addSubEle>
    <addDesc>Domain consistency of the dataset</addDesc>
  </dqReport>
  - <dqResult>
    <measName>Number of errors</measName>
    <dateTime>12-10-2010</dateTime>
  - <measResult>
    - <Result>
      <resTitle>The conformance of Dataset for XYZ Project</resTitle>
      <conExp>Conformance to user requirements</conExp>
      <pass>0</pass>
    </Result>
    - <quanResult>
      <quanValDomain>Number</quanValDomain>
      <quanRes>141</quanRes>
    </quanResult>
  </dqResult>
</DataQuality>
</Metadata>

Figure 3. A sample of metadata file produced by web service
QUALITY EVALUATION REPORTING

<table>
<thead>
<tr>
<th>Report Identification</th>
<th>Quality evaluation report for the XYZ dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>Report scope</td>
<td>Scope defined in metadata (See dqScope)</td>
</tr>
<tr>
<td>Data Quality Measure</td>
<td></td>
</tr>
<tr>
<td>Measure ID</td>
<td>12</td>
</tr>
<tr>
<td>Measure Description</td>
<td>Number of domain inconsistencies</td>
</tr>
<tr>
<td>CompMeasure Value</td>
<td>141</td>
</tr>
<tr>
<td>Value type</td>
<td>Number</td>
</tr>
<tr>
<td>Conform reliability</td>
<td></td>
</tr>
<tr>
<td>Conform reliability values</td>
<td>Conformance level value: 0</td>
</tr>
<tr>
<td>Conform reliability domain</td>
<td>Number</td>
</tr>
<tr>
<td>Data quality evaluation method type info</td>
<td></td>
</tr>
<tr>
<td>Data quality evaluation method type</td>
<td>2 (direct internal)</td>
</tr>
<tr>
<td>Data quality evaluation sampling applied</td>
<td>0 (no sampling)</td>
</tr>
<tr>
<td>Data quality evaluation method info</td>
<td></td>
</tr>
<tr>
<td>Data quality evaluation method description</td>
<td>Compare attributes of items within scope against acceptable attribute domain (acceptable values) and determine if any are outside the domain</td>
</tr>
<tr>
<td>Data quality evaluation parameter information</td>
<td></td>
</tr>
<tr>
<td>Data quality evaluation parameter definition</td>
<td>Acceptable domain values</td>
</tr>
<tr>
<td>Data quality evaluation parameter values</td>
<td></td>
</tr>
<tr>
<td>Data quality evaluation sample method</td>
<td></td>
</tr>
<tr>
<td>Data quality evaluation sample method description</td>
<td>All items within scope in dataset</td>
</tr>
<tr>
<td>Data quality evaluation result</td>
<td></td>
</tr>
<tr>
<td>Data quality evaluation value type</td>
<td>Number</td>
</tr>
<tr>
<td>Data quality evaluation value</td>
<td>141</td>
</tr>
<tr>
<td>Data quality evaluation value unit</td>
<td>Attribute domain violations</td>
</tr>
<tr>
<td>Data quality evaluation date</td>
<td>12-10-2010</td>
</tr>
<tr>
<td>Data quality evaluation result description</td>
<td>Dataset fails. More than 141 items had attributes that violated the attribute domain</td>
</tr>
</tbody>
</table>

Figure 4. A sample of quality evaluation report produced by web service
IMPLEMENTATION

Characteristics of implementation environment:

• XML-based datasets that obey a standard model schema.

• Microsoft visual studio .Net 2005:
  – ASP.NET 2.0 development environment
  – HTML, ASP, and Javascript (User Interface design)
  – C#.net 2.0 (application design)

• User requirement file, and Result file in XML format.
TEST AND VALIDATION

• Two different CityGML datasets were used for testing the web service prototype.

• Both scenarios of automated/semi-automated quality evaluation were carried out.

• The data quality results were reported as metadata and quality evaluation report.

• Some problems:
  – Problem in transferring datasets larger than 4-5 megabytes.
  – For large datasets the evaluation procedure is time-consuming.
CONCLUSION

This research presented a process flow framework to enhance the evaluation and implementation of the geographic information quality standards.

Different aspects of spatial data quality that have the ability to be evaluated automatically were reviewed.

Model was designed in BPMN. Using BPMN brings benefits such as:
- Consistent, and standardized processes
- Sharable and re-usable
- Easy to understand

Limitation of the model is that it does not provide quality information about all data quality elements.
RECOMMENDATION

- For web service implementation:
  - Business Process Execution Language (BPEL)
  - Simple Object Access Protocol (SOAP)
  - Web Services Description Language (WSDL)

- Support other data types
- To Complete the model:
  - Use metadata to provide quality information regarding the data quality elements which were not covered in this model.
THANK YOU FOR YOUR KIND ATTENTION.