Accuracy 2010 Symposium, July 20-23, Leicester, UK

Accuracy-aware web-based geoprocessing chains

Andreas Donaubauer
Geographic Information Systems Group
ETH Zurich
Zurich, Switzerland
donaubauer@geod.baug.ethz.ch

Tatjana Kutzner, and **Florian Straub**
Fachgebiet Geoinformationssysteme
Technische Universität München
München, Germany
{tatjana.kutzner, florian.straub}@bv.tum.de

Abstract—**Quality information is essential to determine the fitness for use of geospatial data for certain applications and furthermore to correctly interpret the results of spatial analyses. However, up to now the latter aspect is still not sufficiently included in geospatial applications and geospatial web services. This paper presents an approach to dynamically generate quality information for the results of spatial analyses within a Web-based chain of geoprocessing operations, hereby taking into account the quality of the input data. The approach is based on standards of the International Organization for Standardization (ISO) and the Open Geospatial Consortium (OGC). An implemented prototype together with a use case from forestry shows the practicability and relevance of this approach.**

Keywords: Metadata, Data Quality, Positional Accuracy, Geoprocessing Workflow, OGC, Web Processing Service, SOA, Error Propagation, Model Driven Architecture, ISO 19100

I. INTRODUCTION

Advances in interoperability research and standardization have led to Geospatial Web Services that allow for easy access to and combination of spatial data from distributed, heterogeneous sources. Furthermore, recently specified service interfaces like the OGC Web Processing Service (WPS) (OGC, 2005) make it even possible to not only access data but also to invoke web-service-based geoprocessing operations on them.

A common way of modeling problems from the real world in a service-oriented architecture (SOA) is to chain multiple Geospatial Web Services in order to obtain a meaningful result. However, Web-based geoprocessing chains often neither take into account the accuracy of the input data nor do they propagate this information within the chain so that a statement on the quality of the output of the chain cannot be made.

In this paper we propose a generic framework that allows for integrating and propagating the concept of accuracy/uncertainty in Web-based geoprocessing chains. In accordance with the Model Driven Architecture (MDA) paradigm, this framework consists of a platform-independent conceptual model which can be mapped to platform-specific implementations.

As a proof of concept we present a prototype of an accuracy-aware Web Processing Service representing one possible platform-specific implementation of the generic framework. Exemplified by a use case from forestry involving a rather simple model for the propagation of positional accuracy in polygon overlay (intersection), we explain the concept and show its relevance to real world applications.

II. RELATED WORK

Our framework for integrating and propagating the concept of accuracy/uncertainty in Web-based geoprocessing chains makes use of scientific work, standards and technologies from a variety of research areas.

MDA as well as SOA are common paradigms in the field of information science. The idea behind MDA is to realize an automated mapping between platform-independent models and platform-specific implementations whereas SOA is an architectural concept which relies on distributed software components, called services.

In the field of geoinformatics the International Organization for Standardization (ISO) and the Open Geospatial Consortium (OGC) have developed a range of standards and reports focusing on building geoprocessing chains, e.g. ISO 19119 and reports on OGC test beds using the Business Process Execution Language (BPEL) (OASIS, 2007). An overview of these activities can be found in (Schäffer, 2009). For a description of geoprocessing chains at conceptual level, a conceptual schema language is required. Our framework applies the language UMLT, described by (Staub et al., 2008).

For exchanging geospatial data over the World Wide Web several standards have been developed by the OGC, most notably the Geography Markup Language (GML) (ISO 19136) which is based on XML to facilitate the modeling and transfer of spatial data. Spatial analysis functionality for processing spatial data can be provided using the Web Processing Service (WPS) standard. WPS does not specify the spatial analysis functionality itself; therefore additionally formal descriptions of geoprocessing operations as defined for example by (Albrecht, 1996) are required.

In the field of data quality, ISO has developed the standard ISO 19113 which defines quality elements for describing the quality of geographic information and the standard ISO 19115 for metadata, which amongst others contains these quality elements. In addition, the standard ISO 19139 was developed to provide a technical specification for encoding the metadata in XML. As an example of quality information which can be generated dynamically, this paper focuses on the positional accuracy of spatial data. Based on former work by Caspary, Scheuring and Frank, several models have been developed by (Glemser, 2000) for integrating information on the accuracy of geometry into spatial data and he describes how to treat this information in spatial analysis.

The use case presented in our paper requires quality information to be added to the spatial data not just on dataset level. The quality information is rather needed on different levels, especially on feature and on sub feature level for describing the quality of individual features and of geometries or individual geometric objects in complex geometries, respectively. In (Donaubauer et al., 2008) an approach for how to implement these requirements using ISO 19139 and GML is presented.

III. A GENERIC FRAMEWORK FOR ACCURACY-AWARE GEOPROCESSING CHAINS

As described above, there are a several ways of how to implement web-based geoprocessing chains, each of them tailored to a specific platform. In order to model geoprocessing chains in a platform-independent way, a concept in accordance with the MDA can be applied. The generic framework described here is an implementation of the MDA paradigm. The building blocks of the framework are as follows:

First, we need to have a conceptual model allowing for a formal description of geoprocessing chains. This model includes formal descriptions of atomic geoprocessing operations, a formal description of activities and their sequence as well as conceptual data models.

Second, we need to have a set of rules describing the mapping between the platform-independent conceptual model and a platform-specific implementation of the geoprocessing chain.

The following list mentions standards, technologies and research work applicable to the components of the framework described above.

[1.] formal description of atomic geoprocessing operations: e.g. (Albrecht, 1996), UMLT

[2.] formal description of activities and their sequence: e.g. UMLT

[3.] formal description of data models: e.g. ISO 19109 (for features), ISO 19107 (for geometry), ISO 19123 (for coverages)

[4.] encoding for [1.]: e.g. WPS profiles or encodings for proprietary web services

[5.] encoding for [2.]: e.g. BPEL + BPEL workflow engine or proprietary implementations of application logic

[6.] encoding for [3.]: e.g. ISO 19136, ISO 19139 and services providing access to this kind of data, e.g. WFS, WCS

The generic framework described above is applicable to geoprocessing chains in general. For introducing the notion of accuracy-awareness into this framework it must be customized in the following way.

At the platform-independent level, a formal description of an uncertainty propagation operation must be attached to each atomic geoprocessing operation. An atomic, quality-aware geoprocessing operation is then a (atomic geoprocessing operation, uncertainty propagation operation)-tuple. This allows for an explicit calculation of the accuracy at each step in the processing chain. The concept described here proposes the conceptual language UMLT for describing these operations on a platform-independent level. In addition to formal descriptions of data models, formal descriptions of metadata models must exist as well. To achieve interoperability, the framework described here makes use of (profiles and application schemas of) ISO standards to formally describe data and metadata models. Different encodings of the platform-independent framework result in different platform-specific implementations of quality-aware geoprocessing chains.

A standards-based platform-specific model could for example be implemented using the following technologies: In a WPS each geoprocessing functionality is provided as a so-called WPS process. Thus each atomic, quality-aware geoprocessing operation can be represented by a WPS process. The WPS will then on the one hand execute geoprocessing operations and on the other hand will be able to dynamically generate accuracy information for the results of spatial analyses by taking into account the accuracy of the input data. The description of activities and their sequence is encoded using the Business Process Execution Language BPEL. The GML is used for encoding the semantic model and transfer format of the spatial data exchanged in the geoprocessing chain. GML allows for embedding metadata containing information on data quality (ISO 19139), such as positional accuracy. For accessing GML data the OGC Web Feature Service (WFS) interface is applied. Details of the proposed platform independent framework and a platform specific implementation are given in the following paragraphs.

IV. A PLATFORM INDEPENDENT LANGUAGE FOR QUALITY-AWARE GEOPROCESSING CHAINS

As mentioned above, UMLT is proposed as a platform independent language for describing quality-aware geoprocessing chains at the conceptual level. UMLT has been successfully used for describing schema translation processes involving a limited amount of geoprocessing functions (e.g. calculate the area of a polygon) (Donaubauer et al., 2007). The following example shows how to describe a quality-aware geoprocessing chain using UMLT.

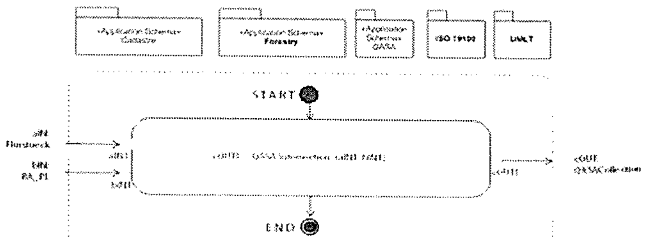

Figure 1. UMLT fragment

First, several UML packages are imported. The packages "Cadastre" and "Forestry" represent the data models of the input data used throughout the geoprocessing chain. The "ISO 19100" package references schemas from the ISO 191xx series of standards which for example contains the basic GML classes and data types being used in the data models and the ISO 19115 metadata schema allowing for modeling accuracy metadata. The package "UMLT" contains the UMLT metamodel which formally defines all the elements needed for describing geoprocessing chains apart from the geoprocessing operations, which are contained in the "QASA" (quality aware spatial analysis) package. This package defines atomic, quality-aware geoprocessing operations, as for example intersection, which is represented by the operation "QASA.Intersection" in figure 1 as well as the output data model of the operation, called "QASACollection". In the example, this operation forms part of a so-called UMLT transformation activity. A transformation activity encapsulates all the transformation rules needed to create the desired output. Here, only one transformation rule is employed. Objects of the classes "Flurstueck" (aIN1) and "BA_PL" (bIN1) represent the input and are routed through the transformation activity where the quality-aware geoprocessing operation "QASA.Intersection" is performed. The operation outputs objects (cOUT1) the structure of which is determined by the output data model QASACollection. Using further schema mapping functionality provided by UMLT, it is possible to transform the output data according to any user defined data model.

The operation QASA.Intersection represents an intersection and is quality-aware. This means, in order to generate accuracy-enhanced output we first need to have a consistent and commonly understood quality description (semantics) for the input. This semantics is provided by the metadata standard ISO 19115. Among the metadata elements defined for describing spatial information and services there are several elements for describing the quality of spatial data. These elements as for example the positional accuracy are ideally suited for our purposes.

A second requirement for producing quality-enhanced output is to take into account the level of the spatial data to which the quality information refers to (compare section VI for a detailed explanation on this). Quality information can either describe the whole dataset (dataset level), individual features (feature level), feature properties, geometry or individual geometric objects in complex geometries (sub feature level). The standard ISO 19115 allows for adding quality information at these levels. In the standard these levels are correspondingly defined as hierarchy levels dataset, feature and attribute.

V. A QUALITY-AWARE WEB PROCESSING SERVICE

We now describe a prototype of an accuracy-aware Web Processing Service representing one possible platform specific implementation of the generic framework.

The OGC Web Processing Service (WPS) interface allows for accessing any kind of functionality needed for processing spatial data. Each geoprocessing functionality is provided as a so-called WPS process. The WPS required in our approach has to be able to not only perform spatial analyses but also to generate quality information for the output data based on the quality of the input data and the geoprocessing operation(s) applied. With regard to section III this means that the WPS has to be capable of processing (atomic geoprocessing operation, uncertainty propagation operation)-tuples. This challenge is met perfectly by the WPS process design by implementing each tuple as separate WPS process.

For the sake of simplicity, the UMLT fragment in figure 1 just shows one atomic, quality-aware geoprocessing operation (intersection). In addition, the prototype implements a second operation (area calculation). Thus there are two tuples which are mapped to WPS processes:

- Tuple 1: (intersection, positional accuracy)
- Tuple 2: (area calculation, accuracy of the area)

The data sent to the WPS processes need to be encoded as GML FeatureCollections consisting of features of types "Flurstueck" and "BA_PL" which both have surface properties. To be able to determine the accuracy of the output data each surface polygon of the input data is represented as MultiCurve aggregate and provided with quality information as described in (Donaubauer et al., 2008). The quality information is read from the input data, tuple 1 or 2 or both are then executed and a new FeatureCollection with the resulting data and its quality information, attached to the same level as the quality information of the input data, is finally returned.

VI. PROOF OF CONCEPT

Based on a use case from forestry (Straub et al., 2004) the feasibility of the above introduced generic framework for accuracy-aware geoprocessing chains as well as its relevance to the real world are demonstrated.

The use case identifies the most suitable types of tree for a given land parcel X as well as the available surface area in hectares. Two kinds of data are needed: 1. Data about forest stands and the suitability of certain types of trees for these areas (feature type "BA_PL"). 2. A digital land register with information about individual land parcels (feature type "Flurstueck"). The data was accessed using a WFS, the quality-aware geoprocessing operations were provided by a WPS and the service chaining was realized by a so-called aggregate service.

The operation QASA.Intersection is performed on the digital land register, which contains land parcel X, and the forest stands, which cover the area of land parcel X. The result will be new surfaces showing which part of land parcel X is suitable for planting which types of trees. Furthermore the surface area of the new surfaces is calculated.

If quality information is integrated into the intersection, it is even possible to indicate the quality of the resulting surfaces with regard to the accuracies of the boundaries and the surface areas. For example, a certain part of land parcel X is well suited for growing spruce. By applying the quality-aware WPS on the use case, it's not only possible to calculate the size of that part in ha, but also the possible standard deviation for that area can be determined.

According to (Lother, 2003) this information is in forestry especially important for economic reasons since the value of a forest stand is calculated by its area and the information also contributes to salary calculation. Tests carried out by Lother using "Forst-GIS", the Bavarian GIS for forestry, revealed for example that a forest stand with an area of 5.8 ha, the boundary of which has a positional accuracy of 5 m, produces a standard deviation for the calculated area of ca. 2%.

The data used shows, why it is important to be able to store the quality information and especially the positional accuracy directly with the geometry of the data. Forest stands have natural, uncertain boundaries which derive from vegetation. These boundaries cause difficulty in exactly locating their vertices, so that the forest stand can exhibit varying values of positional accuracy within its boundary. To know all accuracy values one boundary can have, it is necessary to store each value directly with the corresponding geometry section of that boundary. The land parcels in contrast have artificial boundaries with less uncertain and homogeneous vertice-co-ordinates. Thus it is sufficient to store the quality information once for the whole dataset.

Storing the quality at geometry level is done using the line-by-line approach described in (Glemser, 2000). Each polygon is separated into line segments and the quality information, here the positional accuracy, is added directly to the corresponding line segment. When applying QASA.Intersection on these surfaces, the propagation model for the positional accuracy behaves in the following way: the quality information of the line segments is simply transferred to the line segments of the newly created polygons.

Figure 2 shows a polygon resulting from an intersection carried out on the land parcel and forest stand data sets. The polygon contains information about the most suitable types of tree and the surface area. By using the quality information from the input data, the resulting polygon can additionally be provided with line segment accuracies describing the border of the polygon.

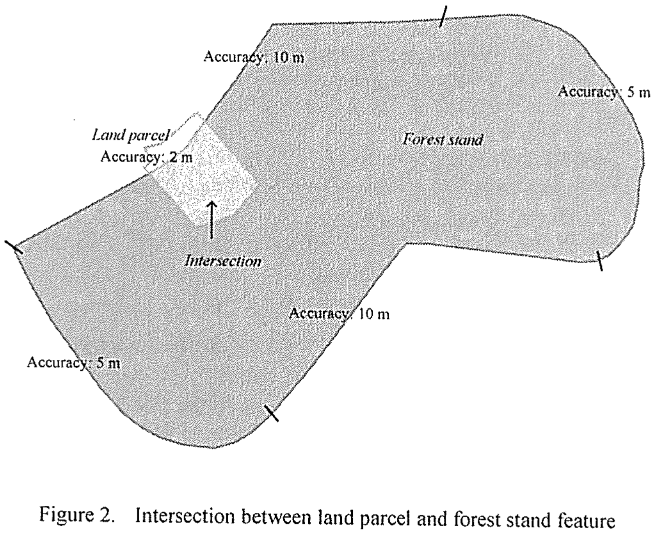

Figure 2. Intersection between land parcel and forest stand feature

Besides knowing the positional accuracy of the polygon boundaries it is important as well for this use case to be able to calculate the standard deviation of the surface area. This corresponds to tuple 2 defined in section V. To be able to roughly determine the area accuracies of natural surfaces an error propagation formula described by (Magnussen, 1996) is implemented in the prototype.

VII. CONCLUSIONS AND FUTURE WORK

Quality information is essential to determine the fitness for use of geospatial data for certain applications and furthermore to correctly interpret the results of spatial analyses. However, up to now the latter aspect is still not sufficiently included in geospatial applications and geospatial web services. This paper presents a generic framework for accuracy-aware geoprocessing chains. Geoprocessing chains described at the conceptual level using this framework can be mapped to various platform-specific implementations. As an example we describe an implementation of the framework based on OGC standards. The use case employed for demonstrating our approach focused on the intersection operation and the quality element "positional accuracy". However, in practice there are a broad range of other geoprocessing operations and corresponding quality elements which should be accounted for in future research. Furthermore, current developments in research, standardization and technology, like the Uncertainty Markup Language (OGC, 2009) should be taken into consideration. The visualization of metadata and especially quality information is another important facet for future research.

REFERENCES

Albrecht, J. (1996). *Universal GIS operators*. Dissertation. University of Osnabrück.

Donaubauer, A., Schilcher, M., and Straub, F. (2007). mdWFS. A concept of web-enabling semantic transformation. In: *Proceedings of the tenth AGILE conference on GI science*, Aalborg, Denmark.

Donaubauer, A., Kutzner, T., and Straub, F. (2008). Towards a quality aware web processing service. In: *GI-Days 2008. Proceedings of the 6th Geographic Information Days*, Münster, Germany.

Glemser, M. (2000). *Zur Berücksichtigung der geometrischen Objektunsicherheit in der Geoinformatik*. Dissertation. Universität Stuttgart.

Lother, G. (2003). *Konzeptionelle Aspekte eines landesweiten Fachgeoinformationssystems für die Bestandsdokumentation forstlicher Geodaten*. Dissertation. Technische Universität München.

Magnussen, S. (1996). A coordinate-free area variance estimator for forest stands with a fuzzy outline. *Forest Science*. 42 1, 76-85.

OASIS (2007). *Web Services Business Process Execution Language* Version 2.0, Oasis Standard, 11 April 2007.

OGC (2005). *OpenGIS Web Processing Service*. OGC Document OGC 05-007r7.

OGC (2009). *Uncertainty Markup Language* (UncertML): Version 0.6. OGC Document OGC 08-122r2.

Schäffer, B. (2009). OWS-6 *Geoprocessing Workflow Architecture* Engineering Report. OGC Document OGC 09-053r3.

Staub, P., Gnägi, H.R., and Morf, A. (2008). Semantic interoperability through the definition of conceptual model transformations. *Transactions in GIS*. 12(2), 193-207.

Straub, F., Donaubauer, A., and Löwis of Menar, O. von (2004). Webbasierte verschneidung verteilter geodaten für die forstliche standortserkundung. *GeoBit - Geoinformatik für die Praxis* 11, 30-31.