Supporting interoperable interpolation: the INTAMAP approach.

Matthew Williams[1], Dan Cornford[1], Ben Ingram[1], Lucy Bastin[1],
Tony Beaumont[1], Edzer Pebesma[2], Gregoire Dubois[3]

[1] Knowledge Engineering Group, School of Applied Science and Engineering, Aston University, Birmingham, B4 7ET, UK. {williamw, d.cornford}@aston.ac.uk

[2] Geosciences Faculty, Utrecht University, The Netherlands

[3] Institute for Environment and Sustainability, Joint Research Centre Directorate, TP 441, Via Fermi 1, 21020 Ispra (VA), Italy

Abstract

In many Environmental Information Systems the actual observations arise from a discrete monitoring network which might be rather heterogeneous in both location and types of measurements made. In this paper we describe the architecture and infrastructure for a system, developed as part of the EU FP6 funded INTAMAP project, to provide a service oriented solution that allows the construction of an interoperable, automatic, interpolation system. This system will be based on the Open Geospatial Consortium's Web Feature Service (WFS) standard. The essence of our approach is to extend the GML3.1 observation feature to include information about the sensor using SensorML, and to further extend this to incorporate observation error characteristics. Our extended WFS will accept observations, and will store them in a database. The observations will be passed to our R-based interpolation server, which will use a range of methods, including a novel sparse, sequential kriging method (only briefly described here) to produce an internal representation of the interpolated field resulting from the observations currently uploaded to the system. The extended WFS will then accept queries, such as 'What is the probability distribution of the desired variable at a given point', 'What is the mean value over a given region', or 'What is the probability of exceeding a certain threshold at a given location'. To support information-rich transfer of complex and uncertain

predictions we are developing schema to represent probabilistic results in a GML3.1 (object-property) style. The system will also offer more easily accessible Web Map Service and Web Coverage Service interfaces to allow users to access the system at the level of complexity they require for their specific application. Such a system will offer a very valuable contribution to the next generation of Environmental Information Systems in the context of real time mapping for monitoring and security, particularly for systems that employ a service oriented architecture.

1 Introduction

Knowledge of the current state of an environmental system is often critical to decision making, for example, in contexts such as disaster response, public health protection or routine environmental management. This knowledge of the state of can only be obtained from (direct or indirect) observation of the system of interest. It can be particularly important that information on environmental variables, (such as the local exposure to hazardous material), is available in real-time, especially in emergency situations. In many cases, where the temporal dynamics of a system are well known, data assimilation methods (e.g. Kalnay, 2003) are used to estimate the system's current state, given the observations. While data assimilation methods are very relevant in certain contexts, modern methods are computationally expensive, and typically will not provide answers within the range of 0-3 hours. In addition to this problem of temporal lag, current data assimilation systems cannot adapt to new observation types easily without recoding. An alternative approach to estimating the state of the system might be to use methods from spatial statistics / geostatistics; however, the recent Spatial Interpolation Comparison (SIC2004) exercise (EUR, 2005, Dubois and Galmarini, 2005) showed that automating such real-time spatial interpolation methods remains an open problem.

In this paper we describe an open architecture we are developing with the INTAMAP project[1] based on extending a range of open standards developed to enable interoperability in geospatial information systems. In particular we address:

- the definition of the architecture of such a system, including interfaces;
- the standards and formats used to communicate between the interfaces;
- a number of open questions that remain.

The paper is intended to provide a framework for discussion and represents work in progress. A wiki documenting progress, with some discussion of the issues being addressed can be consulted for the latest developments[2]. We note that the architecture we describe is one of several possible candidates being developed as part of INTAMAP, but all have a similar, service oriented character.

2 System Infrastructure

To minimise latency in the system a decision to split the system into two individual subsystems was made (Figure 1). The core component is a dedicated computational server (Interpolation server) dealing solely with the interpolation of spatial data. The interpolation process is expected

[1] http://www.intamap.org/
[2] http://wiki.intamap.org/index.php/INTAMAP_Wiki

to be the most computationally demanding part of the system and thus a separate web server will be used to handle more routine tasks such as data storage and trivial requests; this will also allow the web server to manage a primitive scheduling system.

2.1 Web Server

The web server is the gateway to the system; all requests will be processed here and, if sanctioned, forwarded to the interpolation server. The heart of the web server will be an Open Geospatial Consortium (OGC) Web Feature Service (WFS) (OGC, 2005) providing a finite set of operations available to end users. In conjunction with the WFS there will be two supplementary interfaces: an OGC Web Map Service (WMS) (OGC, 2006) and an OGC Web Coverage Service (WCS) (OGC, 2003). These three distinct interfaces provide the service oriented backbone to the system, allowing users to access the system in a manner that is appropriate to their needs, as discussed below.

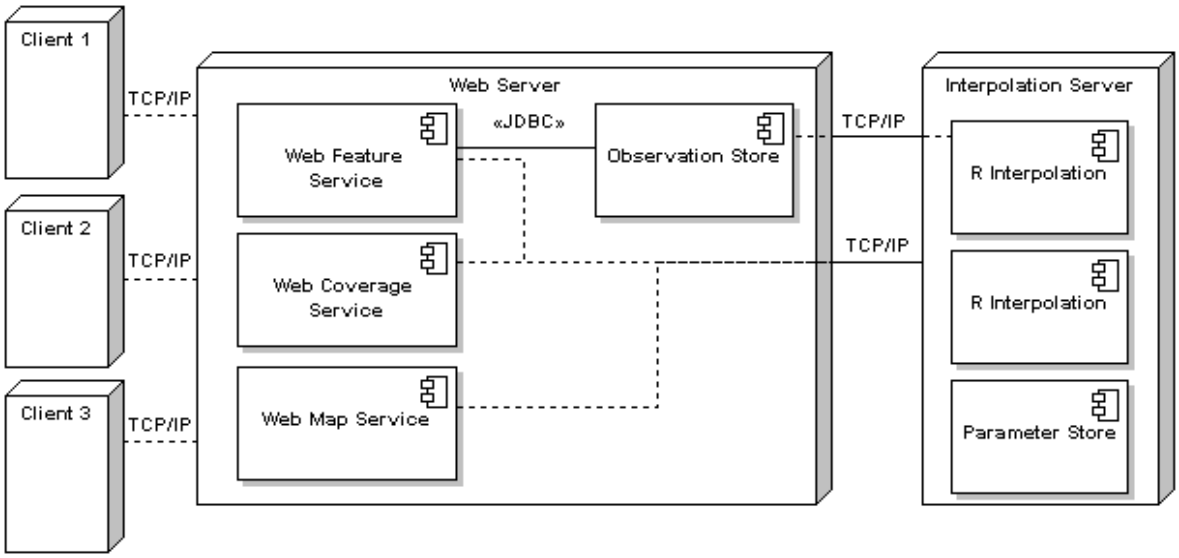

Figure 1 Overview of the architecture of the proposed system

2.2 Interpolation Server

Trivial requests are filtered out by the web server, leaving the interpolation server free to concentrate on a specific task: interpolation. The moment a significant request is received the web server will send a notification to the interpolation server, which will commence the interpolating immediately. As automatic interpolation remains an unsolved problem, there is unlikely to be a unique solution and the proposed architecture allows a range of probabilistic interpolation methods to be implemented on the interpolation server. The sparse, sequential kriging method we briefly describe in this paper is a novel approach that offers significant benefits, in the real-time context, over more traditional methods (Cornford *et al.*, 2005), but a range of other methods are possible and will be explored in the INTAMAP project.

2.3 Why separate systems?

For the majority of the time, the data received may be sporadic at best, and splitting the system into distinct parts could be perceived as an unnecessary complication. However, during emergency situations the delegation of less significant requests to the web server frees the interpolation server to process emergency data, saving vital time. The architecture also allows us to change the interpolation server, for example to a grid based cluster, or a massively parallel machine, without requiring any change in the web service interface implementation. This is an important feature as we expect our web service to be of particular value to users who are engaged in their own modelling exercises. Such users may well be interested in using the maps produced by the server for further analysis and modelling, and are likely to take an interest in issues of computational strategy and efficiency.

The separation of the systems is not without its disadvantages. As all requests must come through one of the three mentioned interfaces they will be encoded in Geography Markup Language (GML) (OGC, 2002). GML, an extension of eXtensible Markup Language (XML) (W3C, 2006), is verbose by nature and when dealing with large requests may be several Megabytes in size, although our prototype employs compression to minimise this effect. Therefore, it would be inefficient to communicate with the interpolation server internally via GML. For this reason a new protocol has been developed to allow efficient yet descriptive communication between the two subsystems, resembling a simple TCP/IP protocol.

3 Web Server Implementation

The web server has numerous roles within the system, which include providing an interface for the client and storing observation data in a spatial database. The decision to employ a Service Oriented Architecture (SOA), and more specifically web services, emphasises valuable characteristics such as reusability, autonomy and discoverability. Introducing a WFS, WMS & WCS into an abstract service integration layer generates a loosely coupled system which is ideal for an interoperable environment (Erl, 2004).

Producing an interoperable system using a WFS relies on GML which, as previously mentioned, is verbose. This poses two problems when working with large datasets. Firstly, the GML instance documents containing the data can reach large sizes. Secondly, parsing of these documents, (which must occur on both the client and server) can be slow. Operationally it is likely that numerous requests regarding the same dataset will be made. Given the aforementioned problems, it would be inefficient to transmit the unchanged data repeatedly. A simple solution is to provide persistent data using a spatial database such as PostGIS[3], which offers many advantages over traditional databases when handling geospatial data. PostGIS is an open source spatial extension to the popular PostgreSQL database, the main benefit of which is to provide efficient spatial queries and coordinate transformation capabilities via the Proj4[4] libraries.

[3] Available from http://postgis.refractions.net/
[4] Available from http://proj.maptools.org/

3.1 Web Feature Service

The flexibility and extensibility of XML has led to the development of a number of application schemata including several relevant ones such as AGSML[5], CityGML[6] & MathML[7]. Interoperability, however, relies on a clearly defined set of standards and an abundance of application schemata describing similar phenomena is a threat to this. Thus wherever possible we attempt to employ existing schema for input and output data, mainly GML3.1. The OGC-defined specification for Web Feature Services was designed to enable sharing of geographic features over the Internet, using a predefined set of standard operations, and employing GML rather than XML to describe the geographic data. There are four operations epitomising simple database operations which provide a functional system. A further two methods, GetCapabilities & DescribeFeatureType, are used to gather information about the WFS and the features it serves. The first returns the capabilities of the WFS in GML and provides a replacement to the WSDL file found in traditional web services. The latter method returns the GML Schema of the requested feature.

There are various levels of Web Feature Services available, all of which support the basic GetCapabilities, DescribeFeatureType and GetFeature (equivalent to an SQL Select statement) methods. For a more complete solution the Transaction operation can be implemented enabling Insert, Update and Delete capabilities. This WFS is commonly known as a 'Transaction WFS' or WFS-T and is implemented in this project. Each adaptation of a WFS returns geographic features in the form of GML encoded instance documents. The meta-data provided by such a service will be critical when the web service is used as part of a processing chain in, for example, a disaster response system. However, if the features that are served are only going to be displayed, for example, as a raster image, then a Web Map Service is better suited to the task. To further complement the system a Web Coverage Service, with its simpler interface, provides a streamlined method for obtaining grid data. Providing a WMS and WCS in conjunction with a WFS offers users a choice of data formats depending on their individual requirements.

3.2 Using the WFS

GML 3.1 introduced an Observation schema which may be used for describing the act and results of observing or measuring some quantity. The schema is quite simple, being composed of four properties. The properties describe the information about the instrument or sensor used to obtain the observation, optional target of observation, the time of the observation and the result of the observing process. A property describing the location of the observation platform is inherited from the Feature schema. A simplified example of how observations encoded in GML3.1 can be sent to a WFS is shown in Figure 2.

Interpolation is in essence a process of prediction and whenever an estimate is made there is always uncertainty. Making decisions based on interpolated data without knowledge of the associated uncertainty can be dangerous. The INTAMAP project will provide the user with an option of choosing the predictive model used, as well as including uncertainty estimates.

---

[5] Available from http://www.ags.org.uk/agsml/downloads.cfm
[6] Available from http://www.citygml.org/
[7] Available from http://www.w3.org/TR/MathML2/

```
<wfs:Transaction>
 <wfs:Insert>
  <gml:Observation gml:id="1">
   <gml:location>
    <gml:Point><gml:pos>-45600 234500</gml:pos></gml:Point>
   </gml:location>
   <gml:resultOf><map:Radiation>34.5</map:Radiation></gml:resultOf>
  </gml:Observation>
  ......................
  <gml:Observation gml:id="10000">
   <gml:location>
    <gml:Point><gml:pos>34998 699765</gml:pos></gml:Point>
   </gml:location>
   <gml:resultOf><map:Radiation>26.8</map:Radiation></gml:resultOf>
  </gml:Observation>
 </wfs:Insert>
</wfs:Transaction>
```

Figure 2. GML example of a WFS Insert operation

```
<wfs:GetFeature>
 <wfs:Query typeName=”gml:RectifiedGridCoverage”>
  <ogc:Filter>
   <ogc:And>
    <ogc:PropertyIsEqualTo>
     <ogc:PropertyName>
gml:RectifiedGridCoverage/gml:rectifiedGridDomain/gml:RectifiedGrid/gml:orig
in/gml:Point/gml:pos
     </ogc:PropertyName>
     <ogc:Literal>-143722 -55000</ogc:Literal>
    </ogc:PropertyIsEqualTo>
    <ogc:PropertyIsEqualTo>
     <ogc:PropertyName>
gml:RectifiedGridCoverage/gml:rectifiedGridDomain/gml:RectifiedGrid/gml:offs
etVector
     </ogc:PropertyName>
     <ogc:Literal>7000 0</ogc:Literal>
    </ogc:PropertyIsEqualTo>
    <ogc:PropertyIsEqualTo>
     <ogc:PropertyName>
gml:RectifiedGridCoverage/gml:rectifiedGridDomain/gml:RectifiedGrid/gml:offs
etVector
     </ogc:PropertyName>
     <ogc:Literal>0 7000</ogc:Literal>
    </ogc:PropertyIsEqualTo>
   </ogc:And>
  </ogc:Filter>
 </wfs:Query>
</wfs:GetFeature>
```

Figure 3. Example GML GetFeature request

Currently our system provides a limited subset of queries; users can request predictions of the mean and prediction variance for a data set for a single point or list of points, or over a grid with

specified spatial bounds and resolution. Figure 3 demonstrates a request for a prediction over a grid with specified bounds and offset values.

The mean and variance are currently generated by simple kriging (Cressie, 1991) and maximum likelihood estimation of the variogram parameters. The data is returned to the user encoded as a 'DataBlock' or array of comma separated values within a GML Coverage Feature (Figure 4). Currently within the GML and WFS schemata there is no way to describe uncertainty in the returned values or grids; therefore a fundamental aim of this project is to develop extensions to GML schema which will provide the user with a means of obtaining the uncertainty of our results. At present the INTAMAP application schema is only able to describe the mean and variance for each prediction.

```
<wfs:FeatureCollection>
 <gml:featureMember>
  <gml:RectifiedGridCoverage gml:id="SIC1">
   <gml:rectifiedGridDomain>
    <gml:RectifiedGrid dimension="2">
     <gml:limits>
      <gml:GridEnvelope>
       <gml:low>0 0</gml:low>
       <gml:high>100 100</gml:high>
      </gml:GridEnvelope>
     </gml:limits>
     <gml:origin>
      <gml:Point>
       <gml:pos>-143722 -55000</gml:pos>
      </gml:Point>
     </gml:origin>
     <gml:offsetVector>7000 0</gml:offsetVector>
     <gml:offsetVector>0 7000</gml:offsetVector>
    </gml:RectifiedGrid>
   </gml:rectifiedGridDomain>
   <gml:rangeSet>
    <gml:DataBlock>
     <gml:rangeParameters>
      <gml:ValueArray>
       <gml:valueComponents><map:Mean/><map:Variance/></gml:valueComponents>
      </gml:ValueArray>
     </gml:rangeParameters>
     <gml:tupleList>
      109.63,112.85 110.71,112.27 111.79,111.62 112.83,110.91
      ..................
      113.25,98.73 95.34,98.02 95.32,97.99 95.38,97.97
     </gml:tupleList>
    </gml:DataBlock>
   </gml:rangeSet>
  </gml:RectifiedGridCoverage>
 </gml:featureMember>
</wfs:FeatureCollection>
```

Figure 4. Example response from the request shown in Figure 3.

The prototype provides a proof of the interoperable concept and can return predictions to a client in seconds. With a sparse interpolation method (as discussed below) and extended uncertainty schemata the accuracy and speed of the predictions will improve. In future implementations it is envisaged that the system will allow a user to either use a default automatic method, or select from a range of novel interpolation methods being developed, often using R, in INTAMAP.

4 Interpolation server implementation

Spatial interpolation encompasses a large number of techniques that are used for prediction at spatial locations where data has not been observed. Kriging is a very popular interpolation technique, also known as the best unbiased linear predictor (BULP). By benefiting from the information provided by a model of the spatial correlation of the analysed process, kriging can frequently generate maps from incomplete or noisy datasets that are better than those obtained by means of simpler deterministic methods. As noted previously, to reduce latency in the INTAMAP architecture, interpolation is performed on a separate system. To further exploit this architecture, a variety of alternative interpolation methods can be made available, and individually specified in the request to the WFS. One of these methods will be Projected Process Kriging (PPK) (Ingram *et al.*, 2007), an extension to existing kriging algorithms.

4.1 Projected Process Kriging

PPK has a number of properties that make it particularly attractive for use in this architecture. PPK is a model based approach similar in spirit to the model based geostatistics proposed in Diggle *et al.* (1998). Firstly, the algorithm employs a sequential method whereby observations are processed individually. This means that the interpolation algorithm can begin computation before all observations have been made available to the interpolation server. During the iterative process an updated approximate posterior distribution is computed after each observation is considered, making it possible for the web service to request intermediate results before all the observations have been processed. A maximum likelihood type II based approach is used to automatically estimate the covariance function parameters.

The basis of the PPK method is to select a representative subset of the observations and project the effect of the remaining observations onto this representative subset in an iterative fashion and without any significant loss of information. The complexity of typical kriging algorithms grows cubically with the size of the data set, or imposes a neighbourhood which introduces artificial discontinuities, whereas the complexity of the PPK algorithm grows quadratically with the size of the representative subset and linearly in the size of the data set. Since the size of the representative subset can be selected, the time-complexity of the interpolation algorithm can be controlled; a feature which has particular relevance in the context of real-time mapping. After processing the data, the model parameters are stored as a compact representation of the posterior distribution for later retrieval. Also, should the need arise, further observations can be added to the model at a later date without having to re-compute the entire model. As noted previously, the intention is to provide a range of interpolation methods that are fast, accurate and provide reliable estimates of uncertainty. Thus while PPK has certain advantages in some mapping contexts, this will be one of many methods deployed on the web service.

5 Discussion

The key benefit of the system we propose is that other users can readily exploit the system using the web service interface, which employs accepted, open standards. This approach is critical to producing interoperable solutions and means that the INTAMAP system can integrate with other service oriented architectures. This integration will be of particular value in the case of SOA-

based projects related to environmental monitoring and disaster management such as ORCHESTRA[8], OASIS[9] and WIN[10]. Future work will require the closer integration of the INTAMAP WFS with standards being developed in these projects. We envisage that the developed system will be suitable for most interpolation tasks and could be employed both in the context of specialised monitoring and risk management, such as air pollution monitoring, or as a generic interpolation service that can be easily accessed via the web, for example to provide real-time maps of geotechnical information to hand held sampling devices employed in the field. The last scenario is typical of emergency situations in which data is collected on site by mobile units.

Rational decision making, in such emergency settings, requires us to quantify the uncertainty which is inevitably present in measurement and prediction. The explicit characterisation of uncertainty is optimally accomplished in a probabilistic framework. As the prototype stands, supplying regular grids of means and variances restricts the user to estimating the marginal probabilities at given points. Therefore more complex models will be developed, ranging from histogram based representations to parametric models, including a range of probability distribution functions, mixture models and samples from posterior distributions. From such results the user can derive, for example, exceedance probabilities over points, areas and grids, or other more context specific uncertainties, often going on to use the results in their own complex models. For example, in the emergency response context the user might require real-time information on radiation levels in an area. Radiological and radioecological models may use complex non-linear representations of exposure and ingestion and provide the responses required for the management of an emergency. The INTAMAP web service could be seamlessly linked to these models and, for example, samples from the posterior distribution of the interpolated dose rates could be requested from the INTAMAP web service. These could be used in the exposure model to provide Monte Carlo estimates of the risk and allow optimal decisions to be taken.

The basic GML Observation schema is limited in that it cannot convey information about the observing process, such as the sensor equations, or information about the accuracy of the observation. As all observations arise from some sensor we are currently investigating the possible use of a different collection of schemata called SensorML. By integrating SensorML with the Observation schema we aim to provide a more flexible solution that allows us to incorporate knowledge of the observation system and thus automate the use of non-standard observations. This work should benefit from, and possibly contribute to, similar work being undertaken by the OGC. Additionally we are developing a new uncertainty schema which we will propose to OGC for consideration as a standard for transmitting uncertain information in XML / GML. This will probably utilise MathML extensively and where possible will link with existing schema. This will allow us to characterise uncertainty in two places; the observation process which introduces uncertainties through the data likelihood; and in providing a summary of our final (posterior) uncertainty given our (prior) model and observations. The implementation of this standard will allow the transmission of information which will be essential for effectively distinguishing predicted ‘false positives’ from real events. This is of particular importance for

[8] http://www.eu-orchestra.org/index.shtml
[9] http://www.oasis-fp6.org/
[10] http://www.win-eu.org/

environmental early-warning procedures, and is a vital element in any operational system for the real-time mapping of safety critical variables. In summary, this paper presents a preliminary look at the INTAMAP architecture and aims, but much work remains to be done to achieve interoperable automatic interpolation.

Acknowledgements

This work is funded by the European Commission, under the Sixth Framework Programme, by the Contract N. 033811 with the DG INFSO, action Line IST-2005-2.5.12 ICT for Environmental Risk Management. The views expressed herein are those of the authors and are not necessarily those of the European Commission.

References

Cornford, D., Csato, L., and Opper, M. 2005. Sequential, Bayesian geostatistics: A principled method for large data sets, *Geographical Analysis*, **37**, 183–199.

Cressie, N. 1991. Statistics for spatial data. John Wiley, New York.

Diggle, P.J., Tawn, J.A. and Moyeed, R.A. (1998). Model-based Geostatistics (with discussion). *Applied Statistics*, **47**, pp 299-350.

Dubois, G. and Galmarini, S., 2005. Introduction to the spatial interpolation comparison (SIC) 2004 exercise and presentation of the datasets. *Applied GIS*, **1**, p. 1–11.

Erl, T. 2004. Service-Oriented Architecture: A Field Guide to Integrating XML and Web Services. Prentice Hall PTR, New Jersey.

EUR, 2005. Automatic mapping algorithms for routine and emergency monitoring data. G. Dubois (Ed.); Luxembourg: Office for Official Publications of the European Communities. EUR 21595 EN; ISBN 92-894-9400-X, 148 p.

Ingram, B., Cornford, D., and Evans, D. 2007. Fast algorithms for automatic mapping with space-limited covariance functions. *Stochastic Environmental Research and Risk Assessment*, accepted.

Kalnay, E., 2003. Atmospheric Modelling, Data Assimilation and Predictability, Cambridge University Press, Cambridge.

Lake, R. Burggraf, D. Trninic, M. Rae, L. 2004. Geography Mark-Up Language: Foundation for the Geo-Web. John Wiley & Sons, New York.

OGC 2002. Geography Markup Language (GML) Encoding Specification, Version: 3.00, OGC 02-023r4, Open Geospatial Consortium Inc. <https://portal.opengeospatial.org/files/?artifact_id=7174> 2007-02-23

OGC 2003. Web Coverage Service (WCS), Version: 1.0.0, OGC 03-065r6, Open Geospatial Consortium Inc. <https://portal.opengeospatial.org/files/?artifact_id=3837> 2007-02-23

OGC 2005. Web Feature Service Implementation Specification, Version: 1.1.0, OGC 04-094, Open Geospatial Consortium Inc. <https://portal.opengeospatial.org/files/?artifact_id=8339> 2007-02-23

OGC 2006. Web Map Server Implementation Specification, Version: 1.3.0, OGC 06-042, Open Geospatial Consortium Inc. <http://portal.opengeospatial.org/files/?artifact_id=14416> 2007-02-23

W3C 2006. Extensible Markup Language (XML) 1.0 (Fourth Edition), World Wide Web Consortium <http://www.w3.org/TR/2006/REC-xml11-20060816> 2007-02-23