Open Data Sources for the Development of Mobile Applications and Forecast of Microbial Contamination in Bathing Waters

Gianluca Correndo, Zoheir Sabeur

Open Data Sources for the Development of Mobile Applications and Forecast of Microbial Contamination in Bathing Waters

Gianluca Correndo and Zoheir A. Sabeur

{gc, zas}@it-innovation.soton.ac.uk
University of Southampton IT Innovation Centre, Electronics and Computer Science, Faculty of Physical Sciences and Engineering, Southampton, United Kingdom

Abstract. This paper describes a service oriented architecture for mobile and web applications and the enablement of participatory observations of the environment. The architecture hosts generic microbial risk forecast models in bathing zones, which are trained by heterogeneous input data. Open observation data sources, specializing in water quality indicators and environmental processes are used for the construction of such applications. Nevertheless, the encountered integration of the open data sources was challenging due to the various incompatibilities found in data samples. These included gaps in the data with diverse temporal and spatial coverage as well as conflicting collection policies.

Keywords: service oriented architecture · bathing water quality directives · open data · mobile and web applications

1 Introduction

The state of bathing water quality in European Member States and beyond is important for a number of industries. These span from fisheries, aquaculture to retail business and tourism industries. With the Water Framework Directive and the later simplified Bathing Waters Directive in place, the EU set up the mandatory standards for the quality of Bathing Water throughout the European Union[1]. As a result, the respective delegated organizations by each Member State Environmental Department have the statutory mandate to monitor bathing zones and regularly report on the state of water quality at all designated bathing zones of each Member States to the European Commission (EC). Hence with the affordability of ad-hoc sensing using mobile devices and accessibility to open data from UK and international sources, it has become a reality to build new and low-cost web and mobile applications. The applications will aid bathing water quality managers, with the support of volunteers, achieve regular reporting on water quality to the EC.

In this study, data-driven models for microbial risks predictions in bathing zones have been put in context of a service oriented architecture. The software service infrastructure provides mobile clients specializing in microbial risk alerts, which are generated by forecast models. The service infrastructure is also enabled for crowd sourcing and the collection of environmental parameters. The risk models have been trained and tested using open data sources prior to their operations.

2 Open Data Collection and Processing

In order to train and test the microbial risk models in bathing zones, it is important to access to open data about environmental observations and measurements such as precipitations (rainfall), river flows, hours of sunshine and so on.

2.1 Open Data for Bathing Water Quality Forecast

The relevance of environmental parameters which explain microbial contamination in bathing waters, is important to construct reliable causal models of microbial risks forecast. Access to open data in order to rapidly develop such models is paramount. The Environmental Agency (EA) of England and Wales has been collecting bathing water quality data since 1988 [2]. The EA is the official organization for England and Wales, which is delegated by the UK Department of the Environment, Fisheries and Rural Affairs (DEFRA) to collect water quality data and report back to the EC. The EA provides both an API to access to the dataset, a Linked Data interface to link to data observations; and dump files with all collected data samples. The data is also associated with UK Open Government Licensing. It includes 530 sampling locations, which cover all the bathing zones of England and Wales. The EA water quality samplings span from 1988 up until 2013. The samplings are taken within the bathing season period, starting from the end of April until the month of October each year.

2.2 Explanatory Processes and Parameters

Various environmental processes may contribute to microbial contamination in bathing zones. However, their level of influence is very complex to predict from first-principles as it will greatly depend on understanding water transport processes and land-sea morphologies near the coast [3]. In this case, models can be efficiently constructed using observation data time series to predict the causal effects of microbial contamination in bathing zones [4]. The data time series may include measurements on precipitation, river runoffs, water salinity, sea surface temperature, wind-induced currents and so on. All these measurements relate to water transport processes and water contamination processes near the bathing zones. They can be retrieved from many open sources currently. For example, the following open data was collected in this study:

- Precipitation (related to preceded 24hrs, 48hrs, 72hrs rainfalls)

- Sea surface temperature
- Wind fields(offshore and onshore)
- Hours of sunshine(also related to cloud cover)

Precipitation

This parameter can be accessed from the American National Center for Environmental Prediction (NCEP) and the National Center for Atmospheric Research (NCAR) [5]. Additionally *atmospheric pressure*, *temperature* and *cloud cover* can be retrieved. The dataset, named DOE Reanalysis 2, is recorded under a 2.5x2.5 degrees grid resolution for daily averages

Sea Surface Temperature

The American National Climatic Data Center provides historical daily measurements of sea surface temperature, sea surface temperatures anomalies, sea ice concentration, and estimated error standard deviations worldwide [6]. The dataset, named Optimum Interpolation Sea Surface Temperature (OISST), reports the measurements on ¼ degree grid and has been collected by two satellites: The American Advanced Very High Resolution Radiometer (AVHRR) and the Japanese Advanced Microwave Scanning Radiometer-EOS (AMSR-E).

Wind fields

The DOE Reanalysis 2 source also provides daily means for wind fields components worldwide [5].

Hours of sunshine

The UK Meteorological Office provides hours of sunshine estimates from the Spinning Enhanced Visible and Infrared Imager (SEVIRI), which is mounted on the Meteosat Second Generation (MSG) satellite. The estimates are based on the fraction of cloud cover per day [7]. The dataset spatial coverage include the whole of Europe for the period [2009, 2012]. The geographical grid is composed of 204 x 367 (longitude/latitude cells).

2.3 Open Data Access

The above mentioned open data sources are heterogeneous. The geographical coverage ranges from the national boundaries to the globe. The gridded datasets use different spatial resolutions under various coordinate systems (i.e. Latitude/Longitude or Easting/Northing). The temporal coverage is also diverse, ranging from decades to just few years. This inevitably challenges the long term analysis of the integrated

open data. Such heterogeneity in open data is due to the fact that they have been collected by different organizations over time with somehow conflicting data sampling and collection policies. Hence the need for the explicit semantic descriptions of the datasets in order to support their discovery and integration will be essential. Alternatively, this can be delegated to users for discovering and inspecting such datasets and manually implementing them under data wrappers.

It is also worth mentioning that there are other datasets on relevant environmental observation for water quality than those introduced earlier. However, their licensing would not allow their direct usage for integration in this study. Redundancy is therefore an aspect to take into account when searching for open and useful data for integration (see for example, Microsoft FetchClimate[1]).

3 Open Data-Driven Models

Microbial risk forecast modelling is often validated at specific marine coastal regions. In this study, the work is on the generic deployment of the risk models into an open service oriented architecture that supports mobile and web applications. The models are efficiently trained using a common data pre-processing and modelling approach which is based on standard multi-linear regressions [4].

For every monitored bathing zone (group of beaches), a number of time series of the environmental parameters were retrieved. The data series were then harmonized and cleansed, prior to training the regressions. The strength of correlations between all selected explanatory parameters and the targeted variable Log10 (Total coliform) was analyzed. In order to find a set of regression variables for a particular beach, a backward stepwise regression is applied. The analyzed data covered several bathing zones and their respective beaches for the period 2009-2010. This was achieved using all collected open data sets with overlapping temporal coverage. For each sampling point contained in the studied region, all the relevant data (explanatory and target data) have been collected, combined and temporally harmonized to daily observations and measurements. Once the data have been quality pre-processed, a linear regression model was deployed. The linear regression based model was then simplified using a stepwise backward procedure which eliminates the statistically less relevant variables. Fig. 1 below illustrates an exemplar analysis of the open data tested for the construction of the models prior to their training, then deployment.

[1] http://research.microsoft.com/en-us/projects/fetchclimate/

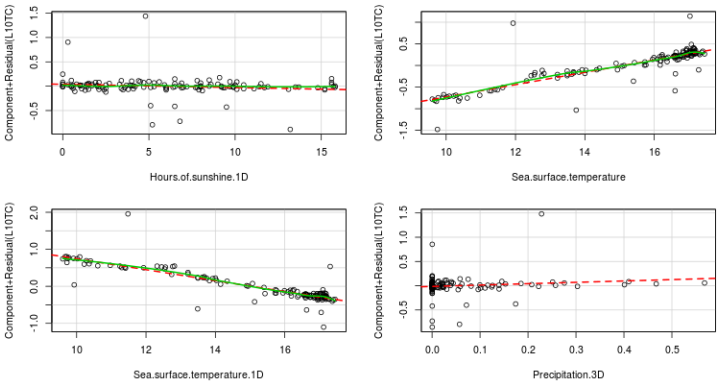

Fig. 1. Data Analysis for model deployment at Bournemouth Pier Beach (South of England)

4 Mobile Application Framework

In recent years, several service oriented architectures based and web-accessible systems have been developed for environmental risk management. The users, mostly specialists, could access to web sensor observations and invoke data processing services for various environmental risk forecasts with great efficiency [8] and [9]. But with the recent advancement in mobile phone communication it is now possible for more user communities to participate in environmental monitoring at local scales. They will need however the support of new open platforms to connect their mobile applications for the ingestion of local environmental observation and measurements. In this paper, a mobile application framework for crowd sourcing environmental observations in bathing zones is considered.

The mobile application framework enables the dissemination of microbial risk alerts in of bathing zones. It is driven by open environmental data sources and made useful for the specific operational needs of environmental regulators, local authorities, industries and volunteers. The framework provides an entry point for providing information about relevant environmental parameters that may affect bathing water quality. On-site users can contribute as volunteers to collect first hand observations for local authorities (e.g. cloud coverage, presence of bird nesting near the coast, number of bathers etc.). The various type of users are described in **Table 1** below.

Table 1. User typologies

User typology	Functionality
Bather	Get: daily notifications of bathing water quality
Casual data provider (Volunteer)	Provide: qualitative information (i.e. presence/absence of bathers), estimates (e.g. bathers density), mobile phone's readings (e.g. temperature, light intensity), and multimedia (e.g. pictures for

	estimating cloud coverage).
Environmental Officer	Provide: sensor readings (e.g. daily river flows) Get: weekly forecasts of water quality for planning samplings
Local authority Officer	Provide: Bathing location's profile (e.g. presence of waste water treatment plants in the catchment area) Get: weekly forecasts of water quality for planning bathing water quality notifications
Industrial	Get: weekly forecasts of water quality for planning operational activities

The functionalities for the target users entail a one way communication from the server to clients and also a back-channel from some users. Smartphones and tablets can be deployed with different types of sensors, some of which can measure relevant environmental properties to water quality foreacsts.

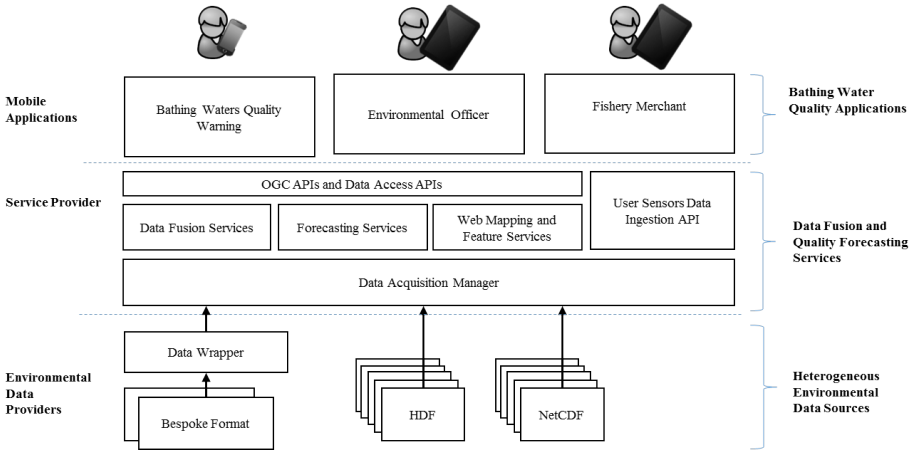

Fig. 2. Bathing water quality mobile application framework architecture

The above architecture (**Fig. 2**) supports a range of applications that are fed with environmental data. A Data Acquisition Manager is included to semantically enrich the ingested open data with relevant formats. Semantic enrichment will disambiguate the data dimensions such as time resolution and extent, geographical granularity and coverage, properties measured etc. For this particular study, a number of environmental data formats were identified. These include: NetCDF and HDF. Other third party data formats need software wrappers in order to be integrated in the framework. The framework maintains a semantic index of the datasets, which is managed along with the represented physical dimensions, temporal and geographical extents and resolutions. This additional metadata is used to harmonise the differences in data representations and allowing the use of ingested data for the forecasting models. The Data Fu-

sion Services exploit the semantic description to perform the data pre-processing functionalities. These are needed in the rapid construction of the forecasting models. Specifically, it entails resampling, interpolation of missing values, temporal series composition and so forth. The Forecasting Services employ the data-driven models, while selecting the relevant parameters to be used for a given beach of interest at a bathing zone. Then risks of microbial contamination are predicted and pushed to the mobile application via an OGC compliant API. This is specifically done by mapping layers that depict the risk values via a browser accessible real-time communication API (**WebRTC**[2] protocol is considered).

WebRTC is an open project that is supported by many Internet technology providers (Google, Mozilla and Opera among others). It provides real-time communication (RTC henceforth) capabilities to web browsers. Mobile applications (depicted in the top part of Fig. 2.) access those APIs via javascript.. The project is undergoing also a phase of standardization within the World Wide Web Consortium community. Hence, it benefits from a wide adoption base with the most used browsers around (Internet Explorer supported via third party plugin).

WebRTC includes the possibility to create an RTCDataChannel, which is a bi-directional data channel between peers that is not based on HTTP protocol. HTTP protocols would be extremely inefficient and slow for high data traffic mobile applications. However, Clients use HTTPS protocol only to initiate a session, while the rest of the communication switches to alternative protocols.

Once a data communication channel is established, the peers can exchange data using proprietary applicative protocols. These are implemented over a real-time carrier. RTC API is employed to communicate with the User Sensor Data Ingestion API and the submission of sensor observation and measurements. The crowd sourced observation is therefore stored using Data Access API. It will subsequently be enhanced with richer meta-information for usage by the risk models. The client layer is composed of a number of Android applications that will access the framework APIs and fetch model prediction and historic observation data. In particular, the mobile application layer supports data intensive applications which render map based visualizations and data plots. It will also sustain data backchannels to the server for the submission of user based observations with streams of sensor measurements.

Most current Android handsets have a range of sensors that fall broadly into one of these following categories:

- Motion sensors: Include accelerometers, gravity, gyroscopes and rotational vector sensors along 3D spatial axes.
- Environmental sensors: Include sensors that measure ambient temperature, pressure, illumination and humidity.
- Position sensors: Include the device physical position with orientation sensors and magnetometers.
- Camera: Include a CCD image sensor (or a CMOS sensor), ranging from 2 to 8 megapixels in resolution.

[2] http://www.w3.org/TR/webrtc/

Sensor readings are usually supplied to the javascript engine. For those sensors that are not yet accessible via javascript API, Android allows to register Java methods as javascript functions. The readings are made available within the mobile application for the visualization and further processing. They are also made available in the server via **WebRTC** along with user's identification and application parameters.

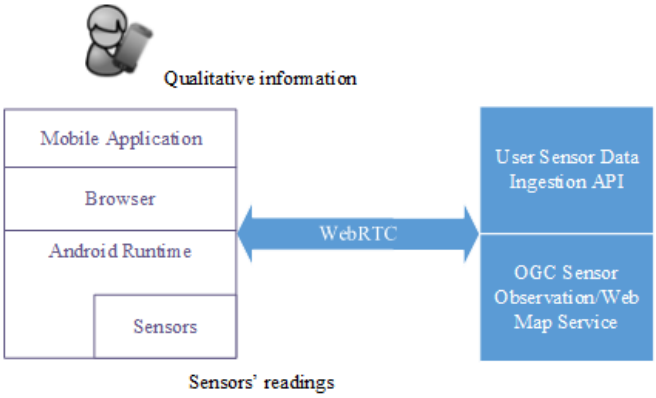

Fig. 3. Mobile Data Provisioning Interface

Users can provide the framework with qualitative observations which may be relevant to the microbial risk forecast application. Number of bathers in a beach of interest, or the presence of birds' nests along the coastline, or cloud cover levels for example are useful qualitative local observations. Cloud cover particularly can be submitted by simply providing digital pictures of the sky at the beach location of interest. The picture can be then processed to estimate the cloud cover levels (in %).

The choice of adopting Android as a target platform is attractive enough in this study. This is due to of its open source development tools and high rate of penetration in the market. This is clearly useful for maximizing crowd sourcing qualitative observation and measurement but also reducing the cost of building further mobile environmental applications and services..

5 Conclusion

This paper describes a service oriented architecture with a common approach for the deployment of generic models and mobile applications. The applications specialize in microbial risks forecast in bathing waters. The models are generically built on standard multi-linear regressions which are trained by heterogeneous open environmental data. Nevertheless, meta-information enhancements were needed to make use of the open data. The development of the mobile application using Android open platforms was also discussed. The application enables crowd sourcing qualitative observations and measurement at bathing zones. Further, it will assist environmental managers and

authorities perform their statutory bathing water quality reports to the EC more effectively and at low cost.

References

[1] Directive 2006/7/EC of the European Parliament and of the Council of 15 February 2006 concerning the management of bathing water quality and repealing Directive 76/160/EEC, Official Journal of the European Union. L64/67, 4.3.2006

[2] http://www.geostore.com/environment-agency/WebStore

[3] B. Lin, M. Syed and R. A. Falconer, "Predicting faecal indicator levels in estuarine receiving waters – An integrated hydrodynamic and ANN modelling approach," Environmental Modelling and Software, vol. 23, no. 6, pp. 729–740, June 2008.

[4] Z. Sabeur, J. Williams, N. Dewey, A. Kozakiewicz and M. Piwowarska. "Development of environmental information tools for the prediction of water quality risks in bathing waters," 2006. ICREW Final Technical Report. pp128. BMT Limited.

[5] http://www.esrl.noaa.gov/psd/data/gridded/data.ncep.reanalysis2.html

[6] http://www.ncdc.noaa.gov/sst/

[7] E. Good, "Estimating daily sunshine duration over the UK from geostationary satellite data," *Weather*, vol. 65, no. 12, pp. 324–328, 2010.

[8] Orchestra. An Open Service Architecture for Risk Management. ISBN: 978-3-00-024284-7. 2008

[9] SANY. An Open Service Architecture for Sensor Networks. ISBN: 978-3-00-028571-4. 2009