A process model for preparation and analysis of cetacean sighting data off the coast of La Gomera

Jochen Wittmann
Environmental Informatics
HTW Berlin
D-12459 Berlin, Germany
wittmann@htw-berlin.de

Aljoscha Marcel Everding
Environmental Informatics
HTW Berlin
D-12459 Berlin, Germany
aljoscha.everding@student.htw-berlin.de

Fabian Ritter
M.E.E.R. e.V.
Bundesallee 123, D-12161 Berlin, Germany
ritter@m-e-e-r.de

***Abstract*—**
Off the coast of La Gomera sightings of whales and dolphins in the context of tourism whale watching tours are recorded continuously since 1995. The tour operators work together with the German registered association M.E.E.R. e.V., which has set itself the target to evaluate the data scientifically. The aim of this thesis is to develop requirements on a process model for acquisition and analysis of this data material. The recording of the data is performed under difficult environmental conditions in the open sea. The data material should be stored persistently in a database that is accessible to both the employees (full access) and the tourists (limited access) on La Gomera, but on the other hand is also the basis for the scientific evaluation. With appropriate role and rights concepts it shall be prevented, that data is distributed too widely and the animals in the observation area are disturbed. Otherwise the requirements for simple access for all interested parties will be fulfilled. This paper proposes a web-access on the protected database and adds a data export for statistical analysis and a geographically based interface via an adequate layer structure in the GIS ArcMap.

I. Setting: Dolphin and whale watching off La Gomera and the M.E.E.R. e.V.

In many ways, the Canary Islands are hot spots, not only as volcanoes. As for marine diversity of whales and dolphins (cetacean) and the concentration of whale watching tourism, the Canary archipelago ranks in a top position. With 28 out of 85 cetacean species it is home to a diversity that knows no comparison worldwide [1].

The Canary Islands are volcanic islands surrounded by waters up to 3,000 m depth, which favors the presence of some purely pelagic cetacean species (eg, sperm whales, pilot whales and beaked whales) in relative close to the coast. Several species are permanent residents, many more are found regularly over the course of a year.

At the same time, the islands are a magnet for millions of tourists. For example, about half a million tourists alone per year take off from Tenerife for whale watching more than anywhere else in Europe. On the island of La Gomera, however, whale watching is practiced in a sustainable way for many years. M.E.E.R. e.V. contributes significantly to this with its best-practice project MEER La Gomera [2]. Data sets that cover the long periods of time which are for the evaluation, are comparatively rare. The sighting data of the cetacean in the La Gomera area, that are collected by M.E.E.R. e.V. since 1995 represent one of the largest data collections in Europe. The sighting database of M.E.E.R. e.V. currently holds over 9,000 sightings collected from 1995 to 2014. The data is collected in cooperation with whale watching trips that are offered twice a day. The fact, that the data is continuously for many years practically without interruption collected, makes the database unique in a further aspect: It is not only possible to analyze differences in the occurence and distribution (for each species) between the years respectively seasonal, but also long-term trends.

The relationship between long-term sighting data and the corresponding environmental parameters could not yet be adequately studied. To enable investigations to this valuable information assets, this article sets the goal to investigate the process from data collection to the representation and analysis and to support it with an integrated software solution in all its phases.

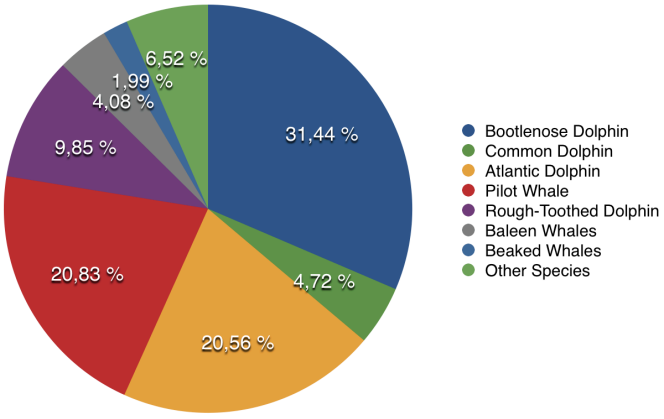


Fig. 1. Relative frequency distribution of cetaceans off La Gomera (1995 - 2007) [1]

II. The development goals

In cooperation between the M.E.E.R. e.V. and the university of applied sciences (HTW) in Berlin a project has been set up to develop a state-of-the art data collection, data storage and data evaluation platform with the following basic development goals:

A. Ensure quality of the basic data

The whole project depends on the datasets with daily sighting data, which where collected during the touristy boat

 ISBN: 978-87-7903-712-0

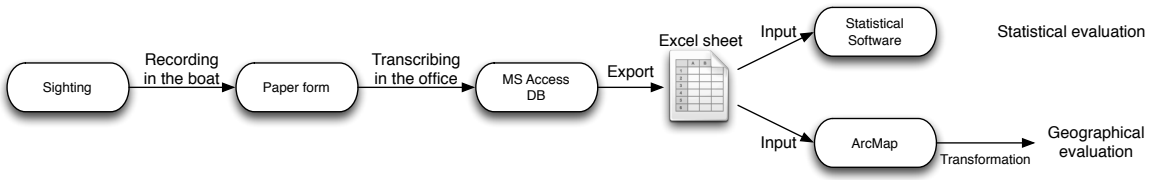


Fig. 2. The current structure of the processing and analysis process

trips. Although the records are just byproducts of the service provided to the tourists, they have to be correct, complete and consistent. Otherwise they would not be useful for analysis.

A view on the data material stored so far on paper-based forms shows with evidence the need of a strict data preprocessing such as

- standardizing names
- consistency check for the position of the sightings
- given possible values for attributes like seastate and interactional patterns etc.
- verification of mutually exclusive details

B. Enable spatial evaluation

In addition to conventional statistical analysis spatial studies are of great interest: for instance the influence of water depth, water temperature, distance to coast etc. as spatial parameters that affect the occurences of each species. In all cases, the geographical location of sightings plays a decisive role.

If such analysis is intended, it would appropriate to have an interface to a geographic information system (GIS), that provides numberless tools for a spatial data analysis. Also, such studies should be supported. Using a geographic information system (GIS) is perfect for this purpose.

C. Enable statistical evaluation

Based on this basic data set it shall be possible to accomplish statistical evaluations. The data is a comfortable selection of attributes, which are interesting for the analysis. It is necessary to provide a technical interface to the data set for established analysis tools (i.e. MATLAB, SPSS etc.)

D. Use of the data for public relations

To support this research, good public relations are essential. The problems and the research objectives have to be vividly depicted. Otherwise there won't be support ideational and financial by society. The creation of information material which is used for this purpose should be possible by the target system. The focus of this material is the sighting data. This data is processed differently than the processings for scientific purposes. The use of data for public relations is important for advertising for new tour attendees. After all the touristy tours are necessary for continious observation and collection of data.

The data should demonstrate in the web the continuous work of the M.E.E.R. e.V. on the one hand and should provide the tour organizers and the tourists with reliable data on current sighting possibilities and may be even probabilities on the other hand1.

E. Consideration of privacy issues

The aforementioned aspects are targeted on an easy and comfortable access to the sighting data. Nonetheless it is important to remember that this data deserve protection. Too many tourist boats lead to harassment of the animals. A simple inquest where which animal was sighted recently could lead to similar effects as in the African National Parks. Very interesting sightings resulted in jams of safari vehicles, which have affected the behaviour of the animals significantly. Fishery and hunting interets are opposed to a complete disclosure of data material. This should be considered during designing the system.

III. The current state

Figure 2 shows an overview of the current structure of the processing and analysis process. The current process begins with the collection of data of each individual observation/sighting on the boat. During the observation a paper form (figure 3) is filled out. This happens for each sighted species. Every paper form is transcribed by hand to an Access-database onshore in La Gomera. On demand the content of the database is manually exported to an Excel-sheet. This file is the starting point for the statistical evaluation with special software applications (SPSS, R, MATLAB etc.).

The problem with this approach is the data consistency. Many different database-versions are in circulation, which are edited and processed in parallel. There is no controlled workflow for synchronisation and update of the database. The Excel-export is used as input for spatial evaluation with ArcMap. In ArcMap the sightings are processed by hand for each version of the data set into thematic maps. The resulting maps allow spatio-temporal analysis by species, season, sea depth, distance to coast etc.

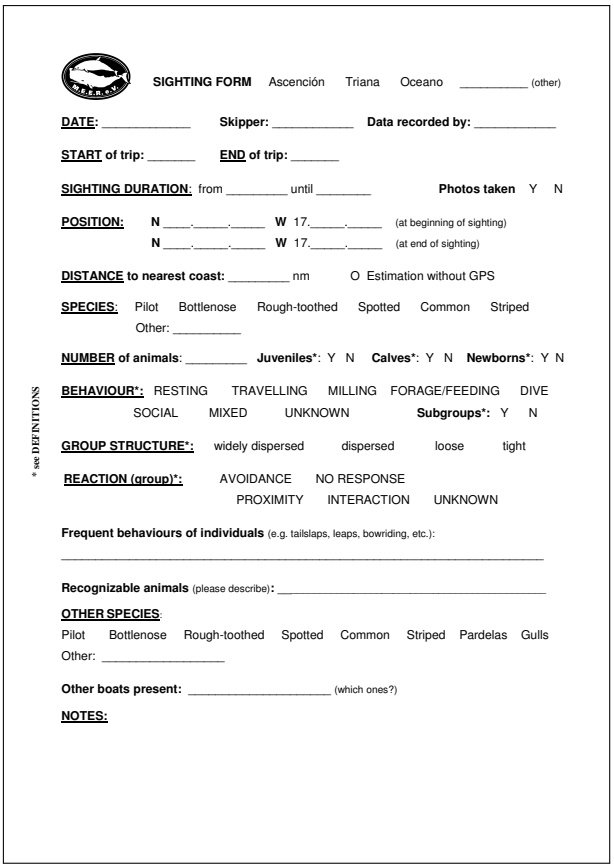
IV. Specification of the target system

As remarked under II the intention of the project was a complete redesign of the data handling process with regard the objectives mentioned and under respect on using adequate software and hardware environment. Thus, the following points will describe the main characteristics of the target system.

A. Replace MS Access/Excel by central database solution

In order to achieve the objectives set out in Chapter II, it is necessary to modify the system described in Chapter III and the currently practiced workflow. Figure 4 provides an overview of the target system's architecture.

The current system's weak spot is the usage of a software solution based on Microsoft Access and Excel. This approach



SIGHTING FORM Ascención Triana Oceano _____ (other)

DATE: _____ **Skipper:** _____ **Data recorded by:** _____

START of trip: _____ **END of trip:** _____

SIGHTING DURATION: from _____ until _____ **Photos taken** Y N

POSITION: **N** ___.____.____ **W** 17.____.____ (at beginning of sighting)
N ___.____.____ **W** 17.____.____ (at end of sighting)

DISTANCE to nearest coast: _____ nm O Estimation without GPS

SPECIES: Pilot Bottlenose Rough-toothed Spotted Common Striped
Other: _____

NUMBER of animals: _____ **Juveniles***: Y N **Calves***: Y N **Newborns***: Y N

BEHAVIOUR*: RESTING TRAVELLING MILLING FORAGE/FEEDING DIVE
SOCIAL MIXED UNKNOWN **Subgroups*:** Y N

GROUP STRUCTURE*: widely dispersed dispersed loose tight

REACTION (group)*: AVOIDANCE NO RESPONSE
PROXIMITY INTERACTION UNKNOWN

Frequent behaviours of individuals (e.g. tailslaps, leaps, bowriding, etc.):
_____

Recognizable animals (please describe): _____

OTHER SPECIES:
Pilot Bottlenose Rough-toothed Spotted Common Striped Pardelas Gulls
Other: _____

Other boats present: _____ (which ones?)

NOTES:

* see DEFINITIONS

Fig. 3. Sighting paper form

has it's strengths in a local office environment. However a local use no longer meets the requirements. The database has to be writeable and retrievable from different locations. With MS Access this is not easily achievable. It is neccessary to use a relational database management system on a centralized server. This allows consistent versioning of the database, which was not possible with the previous system and procedures.

The data structure represented in the Access database is a grown system which does not satisfy current standards of data modelling. Redundancies and anomalies can be prevented by a remodelling and normalization of the existing database design. Data is summarized and arranged in a more reasonable way (figure 5). This ensures consistency of the stored data, which is important for data security and continuity without data loss during creation of new entries and modifications of existing ones.

The database on the server is the heart of the new system. In addition to the database, there is a server-side backend. This backend makes it possible to implement an API (Application Programming Interface), which can be addressed by different client systems (web, mobile etc.). This API allows data input by different clients as well as output of the data via web interface plus the export into an Excel file. Data security has a very high priority. Current standards of authentication guarantee, that only authorized clients have access to the backend and consequently the database. Data will be transferred encrypted only.

By replacing the current solution by a web-based application, it is possible that multiple users can work with the database with the same state of data at the same time. In addition, the new solution is usable plattform independent. To use the new system, just a computer with internet access and a modern internet browser are necessary. Therefore the usage does not require a computer with an operating system from the Microsoft Windows family. It is also runnable on devices with operating systems such as Linux or Apple Mac OS X. Also, since the entire data can always be accessed in an up-to-date state via internet, the user does not rely on a specific computer.

B. Optimize Excel Interface

The system architecture provides that sighting data will be kept in a central database. In addition, it must still be possible to proceed statistical and spatial-temporal evaluation in specialized software systems like SPSS or ArcMap.

The new architecture still uses an Excel-Interface as exchange format. It turned out that importing this format into the specialized software is quite easy. Furthermore, it is a transparent and human readable form of data transfer. For statistical evaluation the entire dataset will be exported into a single table. This simplifies the import into the statistical applications.

For spatial evaluation in GIS the table will be preprocessed during export. Data will be combined into several sheets, so it won't be necessary to use additional selection methods to transfer the data into previously generated layer structures. Thus the data access will be still pretty easy, though it creates several new versions of unsynchronized and unmaintained dataset. This approach is pragmatic and reasonable, because the basic data isn't modified, but just read. This regards both, working in GIS and statistical analysis.

In addition, the concept supports an update to the latest version of the data set by a semi-automatic export-import mechanism via the Excel-file interface. Editing of the basic data should be allowed only via the database interface.

C. Web output

There are three layers on which the web output of the data shall be possible:

1) Easy and comfortable database access for project staff:
   So far, the system comes with an Access view and further processing in Excel. With the new system, it shall be possible for project staff to maintain a comfortable view (figure 8) and the possibility of subsequent data processing in the browser.
   Due to the order records are stored in the database, it will be possible to visualize them in a fast and simple manner, whether as a report, graph or table.
2) For tourists:
   The data set makes it possible to provide special processed data and information about their tour, so the tourists can experience the tour again and have additional benefit and a nice souvenir of their trip. For this view critical information will be hide.

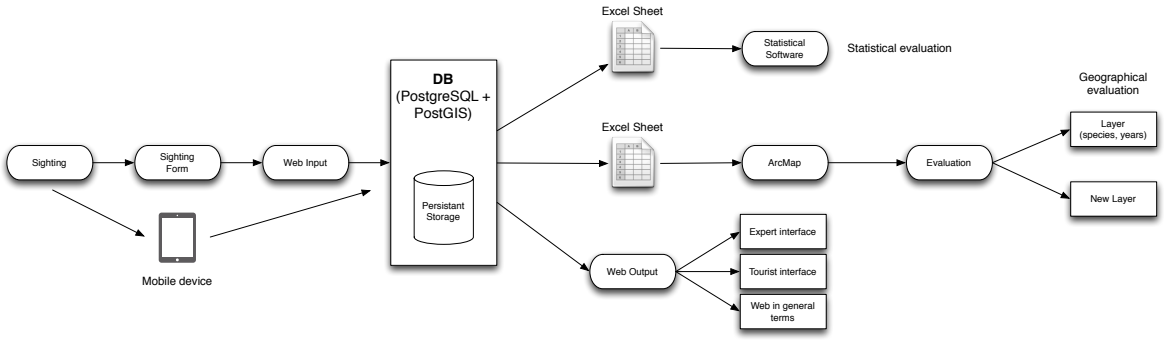


Fig. 4. The processing and analysis process in the target system

Another use of the sighting data would be a forecast of future sightings. Prior to booking of a whale watching boat trip, tourists could get information, which sightings might be possible under current conditions.

3) For general info site:
With the API, it is easy to make sighting data accessible in processed form on the website of M.E.E.R. e.V.

D. Optimize data entry

An input mask in the browser replaces the input form in MS Access. Via mobile devices, such as smartphones and tablets, it is possible to enter sighting data into the database digitally and directly.

This renders filling out sighting paper forms and their subsequent transfer to the database unnecessary. Thus, a media break, which is not only labor intensive but also error prone, can be avoided. By managing the data input in direct connection to the database in behind all restrictions, rules, and interdependencies between the data entries can be controlled online and the typical transcription errors due to illegible handwriting or typing errors, can be reduced and the consistency and the completeness of data can be enhanced significantly. The result is a notably improved quality of data.

E. Provide GIS structures for analysis

Recurring steps can be identified for data processing and analysis in GIS:

a) Import of the data table (database connection)
b) Adjustment of the geographical coordinates of the data format used in GIS
c) Processing of the data set into individual standard layer with sightings by species, observation period, location etc. including pleasant symbologies.
d) Determination of the sea depth for the positions of sightings
e) Determination of the distance to the coast for the positions of sightings
f) Providing spatiotemporal maps for other environmental parameters such as surface temperature of the sea, chlorophyll level, etc.

The latter point is solved through a suitable research and the integration of the corresponding thematic maps in the GIS.

The items a) to e) represent steps, that

1) concern the import of existing data (step a) to c)). After the basic structures are created in the GIS, step a) to c) can be automatically executed by an update tool.
2) complement existing information about spatial attributes (step d) and e)). This feature should be triggered under user control.

The Python scripting interface of ArcMap [3] is perfect for both types of actions. A more user-friendly alternative would be an additional visualization by so called Modelers [4].

As Jermann demonstrate in [5] most of these steps can be executed automatically and produce maps with sighting visualizations even on a daily basis ”just at the touch of a button”.

V. The achieved subgoals

A. Processing of the data material

First of all, for preparation numerous Excel sheets in circulation had to be collected and merged. This work was performed manually. The present database contains more than 9,000 data sets of sightings continuously from the year 1995 to March 2014.

In addition, based on nautical maps, the Access database was completed by adding the missing attributes sea depth and distance to coast. The correctness of the data was checked using the sighting paper forms (figure 3) and corrected if necessary. This provides a complete and consistent data basis for the integration of the new system architecture.

B. Implementation of the central database component

To implement the web application the programming language Ruby and the Ruby on Rails web framework has been used. Ruby on Rails is considered a very safe and stable framework that is being actively developed and maintained. It innately comes with many features required in the target system.

As database management system, PostgreSQL with PostGIS is used. PostGIS is ideal for storing spatial data like sighting positions and their distance to coast.

The API is based on the REST (Representational State Transfer) paradigm by Roy Thomas Fielding [6] and uses JSON (JavaScript Object Notation) [7] as exchange format.

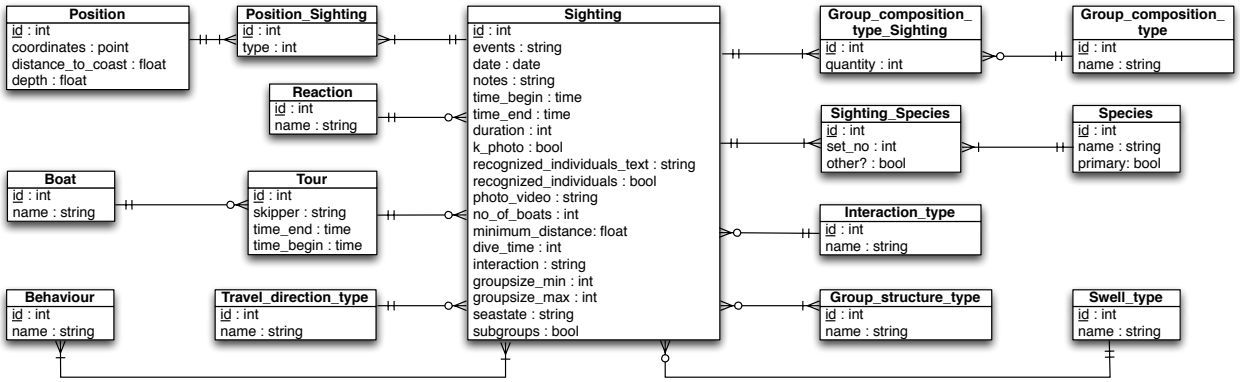


Fig. 5. ER model of the target system's backend

Because of the usage of HTTPS, the data transfer is generally encrypted. The view layer of the webapplication is implemented with a mixture of Twitter Bootstrap and the JavaScript HTML5 framework Ember.js.

For the mobile application view Sencha Touch is used. With a framework like Sencha Touch based on modern web technologies, it is possible to create a plattform independent offline-enabled mobile application without using plattform-specific APIs [8]. To get a native/hybrid mobile application, the mobile web view is wrapped in a PhoneGap container [9].

All technologies used are Open Source and are free of licence fees, which meets the limited budget of a non-profit project.

C. Composition of data structures within the GIS

The data structures outlined in Chapter IV-E are created in ArcMap. They already allow visualization of the first reports as sighting maps which can be generated according to the attributes that are interesting for the evaluation.

For example (all figures from [10]):

- map of sea depths of all sightings (figure 6)
- map of sightings by species in April 1995-2011 (figure 7)
- map of sightings of one species (Bryde's whale) in september for all observed years (figure 9).

Based on this map material, it is now possible to analyse the data under several aspects like correlations between the occurrence of the species among themselves, neighborhoods, observation period, sea depth, inshore and the additional parameters which are mentioned in Chapter IV-E. The (geographical) sighting position and the temporal dimension via sorting by year are visualizable and allow an extended view of the base data that outreach a purely statistical analysis.

VI. Future Prospect

In the steps described here, a software environment could be created that allows domain experts to proceed spatial and time-related analyzes on the valuable record of the cetacean sightings. The consistency of the data management is ensured by the central database server solution. The different user profiles are represented by a corresponding role concept. The implementation of a web-based solution provides comfort

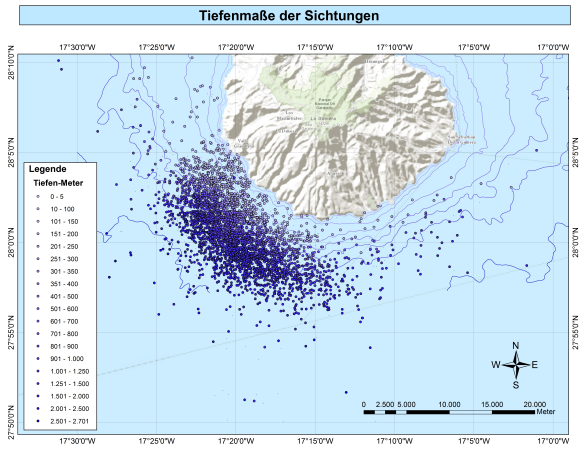


Fig. 6. Sea depth at the sighting positions (from [10])

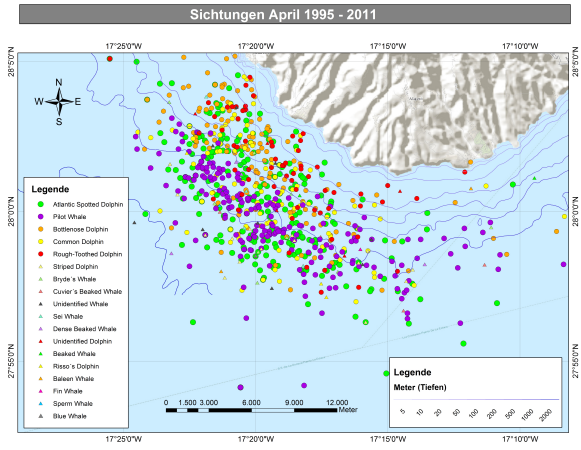


Fig. 7. All sightings in April by species (from [10])

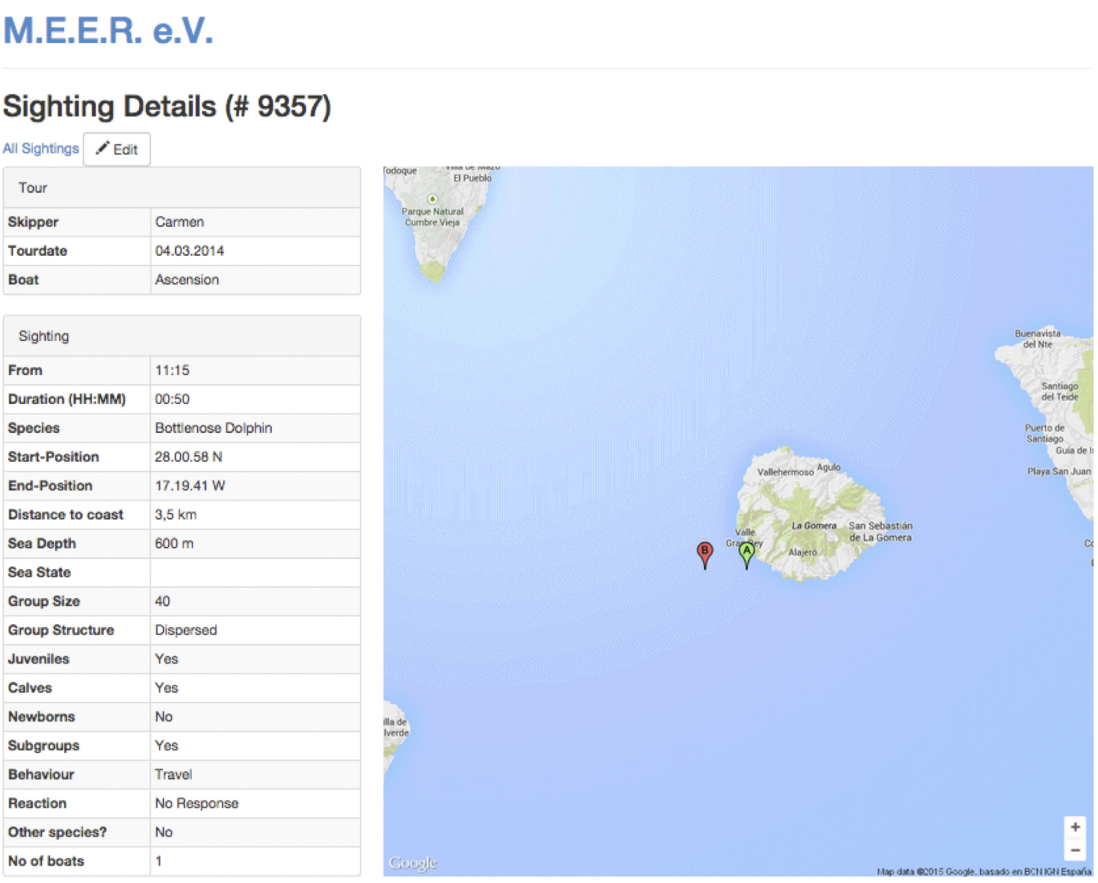


Fig. 8. Screenshot of sighting details in the web application

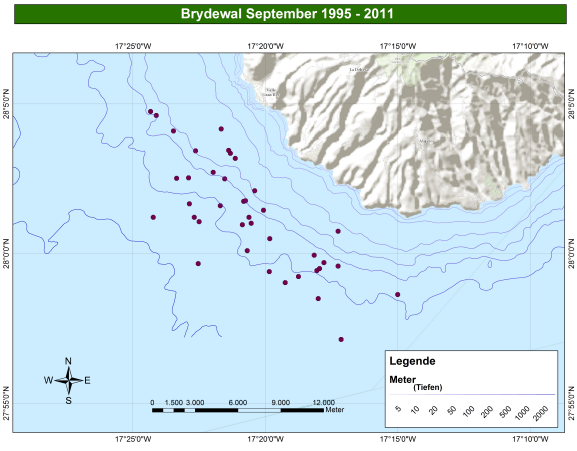


Fig. 9. All Bryde's whale sightings in the month of September (from [10])

using the latest software components. The GIS connection simplifies data transfer for spatial analysis.

The next steps are now the actual investigation of correlations between observational data and environmental parameters. Here the GIS provides powerful tools. First considerations suggest that the analyzing workflow will be a mixture of conventional statistical methods and GIS tools. This leads to further considerations and concepts allowing an intuitive problem-oriented operation and spatial-temporal evaluations.

References

[1] F. Ritter, *Delfine und Wale vor La Gomera - Artenvielfalt im Wandel*, Brochure by M.E.E.R. e.V., Berlin 2008.

[2] F. Ritter, *Model for a Marine Protected Area designed for sustainable Whale Watching Tourism off the oceanic Island of La Gomera (Canary Islands)*, Report by M.E.E.R. e.V., Berlin 2012.

[3] ESRI: *Introduction to Geoprocessing Scripts Using Python®*, http://downloads2.esri.com/campus/PRDpdfTOC/50127508_10.X.pdf [2014, July 14].

[4] ArcGIS-Tutor: *An overview of ModelBuilder*, http://webhelp.esri.com/arcgisdesktop/9.2/index.cfm?TopicName=An_overview_of_ModelBuilder, [2014, July 14].

[5] L. Jermann, *Entwicklung einer GIS-Komponente zur automatisierten Darstellung und Verarbeitung von georeferenzierten Sichtungsdaten*, Bachelor thesis, Hochschule für Technik und Wirtschaft Berlin, FB2, Umweltinformatik, Berlin 2015.

[6] R. T. Fielding, *Architectural styles and the design of network-based software architectures*, PhD Dissertation, University of California, Irvine, 2000.

[7] T. Bray, *The JavaScript Object Notation (JSON) Data Interchange Format*, RFC 7159, https://tools.ietf.org/html/rfc7159, 2014, [2015, March 10].

[8] F. Franke, J. Ippen, *Apps mit HTML5 und CSS3*, Galileo Press, Bonn 2012.

[9] R. Steyer, *Apps mit PhoneGap entwickeln*, Hanser, Berlin 2013.

[10] L. Heuer, *Darstellung von Sichtungsdaten von Walen und Delfinen vor La Gomera mithilfe eines Geographischen Informationssystems (GIS)*, Bachelor thesis, Hochschule für Technik und Wirtschaft Berlin, FB2, Umweltinformatik, Berlin 2013.