Malariapedia Maps: A Semantic Wiki Query Result Visualization and Annotation Tool for Bulk Geocoded Data

Günther Zehetner
Max-Planck-Institute for Molecular Genetics, Berlin, Germany

Abstract. Malariapedia Maps serves as query result display format for bulk geocoded data sets stored as semantic multi-value properties in the Malariapedia semantic wiki. Those sets usually present results from on-field research projects and contain besides experimental result also precise location information. Results can consist of several hundreds or thousand single data points each represented by a semantic property in the wiki. The maps are generated by executing semantic queries. Key values of a data set can translate to the size and colour of map markers and interactive facet browsing allows filtering the displayed data points by various criteria. Additional relevant information can be added as overlays and via links to external position specific data. The goal is to aggregate a wide spectrum of information relevant to research results at a specific area by providing a tool which allows to easily mash-up own experimental data with published data.

1 Introduction

Malariapedia is an experimental semantic wiki for genomic data with the purpose to act as research and annotation tool for local research data by combining them with a pool of publicly available data from major domain databases in a common and searchable format. The prototype site uses Malaria related information from the species Anopheles gambiae and Plasmodium falciparum, which play a major role in the transmission of the disease. A subset of the wiki data originate from published results of various research projects in different locations around the world producing large sets of georeferenced. Malaria Maps[1] uses a specially adapted version of the SIMILE Exhibit application (David F. Huynh, 2007) as query result printer for the wiki and provides an intuitive way to not only display such large data collections but to also gain additional insights by adding various location relevant information from a wide range of sources to generate a profile of the area around a data point.

Figure 1. Data and information workflow using Malariapedia Maps system

[1] http://maps.malariapedia.info/

Interactive tools are made available within the map display to achieve this. They allow to filter data by individual values, to produce a mash-up display with data from other geocoded sets stored as semantic properties, to add different overlays to the displayed map regions and to link to location specific external information.

2 Semantic Wiki and Geocoded Data

The wiki is build using the open source software Mediawiki (Daniel J. Barret, 2008) to which new functionality can easily be added by installing extensions. The Semantic MediaWiki (SMW) extension (Markus Krötsch et al., 2007) is used for storing semantic information as properties. A new extension SemanticQuery-FormTool (SQFT)[2] has been written for the project and a number of other extensions have been adapted including the Exhibit programme. The extended Exhibit application (SQFT_Exhibit) implements a new result format called 'sexhibit' which serves as result format printer for the semantic queries used in the Malariapedia (see Table 3). Significant modifications of the original Exhibit code were required to deal with the large number of data values used in this system, especially if more data sets are combined in a query.

Table 1. Geocoded data sets implemented in the wiki. The data can either be static (result sets from data files accompanying published papers) or dynamic like the actual malaria net distribution information from Against Malaria which is automatically retrieved as KML files on a daily basis and converted to semantic properties.

Data set	**Data values** (semantic properties)	**Project**
Distribution Records of Anopheles gambiae	30360 (2530)	Mapping Malaria Risk in Africa / Atlas du Risque de la Malaria en Afrique (MARA/AMRA) http://www.mara.org.za
Entomological Inoculation Rates	2895 (193)	
Karyotype and Microsatellite Data	102100 (5105)	Individual-level Population Genomics Database for Arthropod Disease Vectors (PopI) https://grassi2.ucdavis.edu
Malaria Parasites Rates	36552 (4569)	Malaria Atlas Project (MAP) http://www.map.ox.ac.uk
Malaria Net Distribution	6084 (507)	Against Malaria (Initiative of the Against Malaria Foundation) http://www.againstmalaria.com

SQFT_Exhibit uses a different method to interpret geocoded data than the already existing map display extensions for semantic properties (e.g. Semantic Maps) and significantly reduces the work required to create, store and maintain large data sets. A number of geocoded data sets have been retrieved from publicly accessible

[2] http://mbw.molgen.mpg.de/wiki/Help:Sask

data files, converted into SMW multi-value property syntax and stored as semantic properties which contain experimental result values as well as location information like country, area, latitude and longitude on experiment specific pages of the wiki. Table 1 lists the data sets which have been added to the wiki. Examples of semantic properties from two different data sets are shown in Table 2.

Table 2. Examples of semantic multi-value properties defining geocoded data. Field values within the properties are separated by semi-colons and correspond to the fields in the definition line. Multiple values within any field can be separated by the system if requested in the semantic query.

Distribution Records of Anopheles gambiae data set	
Example property	[[DisR::-1;6.4000,2.5167;COTONOU AREA(GOD+KET);BENIN;01-Jul-1968;31-Jul-1968;House Resting Daytime; GAMBIAE; chromosomal banding sequences; COZ 1973; MARA;It is not known what 'a' (aka -1) for 'Number of specimens' stands for in George Davidson's original database.|]]
Karyotype and Microsatellite Data set	
Example property	[[KaryotypeData::Cameroon;Bamessing;5.63330,10.23330;1474; 07/08/2004;002,007,016,056,063; An. gambiae; Savanah;S form; Gambiae hybrid; inversion is present on only one of the paired chromosomes; no inversion;inversion is present on only one of the paired chromosomes; no inversion; no inversion;no inversion;b/+;b/+ (hybrid-For/Sav); Bamessing/2004-08-07/2/multiple_a.png, Bamessing/2004-08-07/2/multiple_b.png;PopI|]]

3 Semantic Query and Result Display

Executing a Semantic MediaWiki #ask query with the 'format' parameter set to the value 'sexhibit', results in the display of a Malariapedia map similar to the wiki page shown in Figure 2 provided that the SQFT extension is installed and the wiki contains the queried semantic multi-value properties.

Query parameters can be specified to indicate which data should be included, which data values should be used as facets to filter the data, which overlays should be made available, how data conglomerates within a data field should be separated, which data values should be used to calculate the size and colour of the map markers and so forth. Table 3 shows an example query which would produce the map shown in Figure 2.

Some of the original web sites which describe the used data sets provide already the possibility to display data points on simple maps. However, they lack any additional integrated tools to analyze results further by using information outside of the data set. Examples are the PopI database map display[3] (web based) and the MARA Lite software[4] for data from the MARA/AMRA project (requires installation on a local computer).

[3] https://grassi2.ucdavis.edu/PopulationData/GeoInfo/GMaps/

[4] http://www.mara.org.za/lite.htm

Table 3. Semantic query to display a Malariapedia Map. Query parameters define which data are included and certain aspects of the map display. '?AM-Nets' specifies the name of the property included in the map and the 'format' parameter defines which format is used to display the query results. 'sexhibit' is the name of the result format used for the Malariapedia Maps display. 'codegeo' indicates the fields whose values might need special conversion to be unified over all data sets. 'views' specifies which map views should be made available (in this case two map and the tiles view). The 'customtiles' and 'countryinfos' parameters determine which overlays are made available. Their values are the results of SQFT extension specific #sask queries. 'countries' allows to restrict the display to a list of specified countries.'sizekey' and 'colorkey' define the fields used for map marker size and color.

```
{{#ask:[[Category:Africa]][[AM-Nets-table::+]]
| ?AM-Nets
| format=sexhibit
| title=Template:AM-Nets_header
| split=",:Years"
| codegeo=Country
| views=map:Map - colour code "Status",map:Map - colour code "Years",tiles
| customtiles={{#sask:
              [[Maps]]|?customoverlay|format=text|nospace=|sep=###|lastsep=|
                                                    namespace=all|nocache=}}
| countryinfos={{#sask:
                [[Maps]]|?countryinfo|format=text|nospace=|sep=###|lastsep=|
                                                    namespace=all|nocache=}}
| countries={{{countries|}}}
| shape=square
| sizelegendlabel=<i><u>Size of icons indicates number of distributed nets:
| sizegradientcoder="0,15;{{{maxnets|26000}}},40"
| sizekey=Nets
| colorkey=Status#Years
| facets=search,Country:cloud,District,Area,Nets,Distributor,Status,
                                                  Start_date,End_date,Years
| facetsleft=4
| lens=AM-Netslens
| latlng=latitudelongitude}}
```

Map Markers: A clustering feature was added to the SQFT_Exhibit code which combines close markers to one cluster marker displayed as a green box showing the number of aggregated markers. Clicking on a cluster marker will automatically zoom the map to a level which is sufficient to show the individual markers without overlap. If several results are stored under the exact same position a normal marker symbol labelled with the number of available results is shown.

Facet browsing and filtering: The displayed data can be filtered by selecting values or value ranges on lists or sliders positioned around the map. Which data fields are displayed as facets can be specified in the query parameters.

Overlays: The implemented overlays show examples of information originating from shape files, image files or coded country specific information which is added to an internal table (the system knows about the border coordinates of each country and can use the code to colour a country shape accordingly). If an overlay is selected, the description and a legend are automatically added to the map display (see Figures 2 and 3). The following overlays have been implemented as examples: **a) Biome**: WWF-Terrestial Ecoregions of the World [ArcView Shapefile] (Olson, D. M, et al., 2001).[5] **b) Cross-blended Hypsometric Tints**: Natural Earth

[5] http://www.worldwildlife.org/science/

with Shaded Relief, Water and Drainages (land colouring based on elevation) [Shape file].[6] **c) Biomes with a `Human Footprint Index<=10`:** from SEDAC [Shape file] (Last of the Wild, 2005)[7] **d) Malaria infos** from Malaria Atlas Project [JPG images].[8] **e) Malaria infos** from Malaria Information System (MIS) by Health-mapping (an initiative of IHPH Institute for Hygiene and Public Health University of Bonn / WHO Collaborating Centre for Health Promoting Water Management and Risk Communication) [KML file with country specific data].[9]

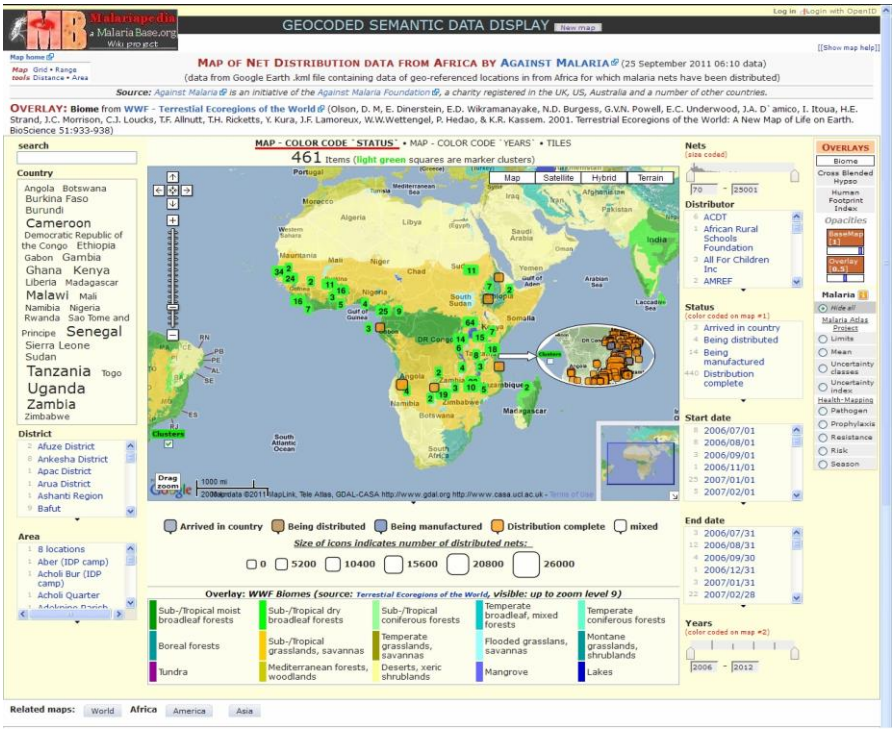

Figure 2. Map display of the semantic query shown in Table 3 (plus the selected Biome overlay). Marker clustering is switched on by default (see inset for an example where clustering was switched off by deselecting the 'Clusters' box). The top region of the display consists of the map title and source. In case an overlay is selected its description is also shown. Left of the title the set of map tools is displayed. The map itself has a right and left row of facets for browsing and selecting data. Directly above the map are selectors for different display types (in this example two map displays with different colour coding and the tiles view). Underneath the map the colour and size coding of selected values is explained. In case an overlay is selected its legend is also. If a data set contains locations from more than one continent than buttons to related maps on the other continents are presented.

[6] http://www.naturalearthdata.com/
[7] http://sedac.ciesin.columbia.edu/wildareas/
[8] http://www.map.ox.ac.uk/
[9] http://health-mapping.com/projects_3.html

Map tools: Several map tools have been integrated to assist the analysis of the data: 1) Rectangle areas can be selected on the map (using the mouse cursor) to define a zoom level showing only this area ('Drag Zoom' button. 2) Cluster markers can be switched off to force the display of single marker symbols even if they are in close proximity (see Figure 3 for a comparative display). 3) A set of tools is accessible left to the map title (see Figure 2) which either allow to measure the distance between two map points, the area defined by several map points or which display a grid over the map or show distance circles around a selected centre point (see Figure 3 C for an example).

External links: For each data point on the map a number of web links are automatically generated containing the latitude and longitude values. They are listed in the header section of the data information window. Right now three links are implemented to explore the possibility of integrating diverse data sources: 1) link to a WWARN[10] Explorer map display centred on the location of the selected data point with medical, genomic, treatment and environmental aspects (Figure 3 A). 2) link to an interactive map showing malaria related Twitter messages originating from a selectable distance around the coordinates of the point of interest (Figure 3 B). 3) link to a Malariapedia map which shows data from all stored data sets in a circle of 10 km, 100 km or 500 km distance around the point of interest (Figure 3 C).

In addition to normal text data values from research results can also represent web links to external information (Figure 3 D and E shows examples of links to more detailed or summary information on external web sites) or represent multimedia content (like photos, graphics or videos).

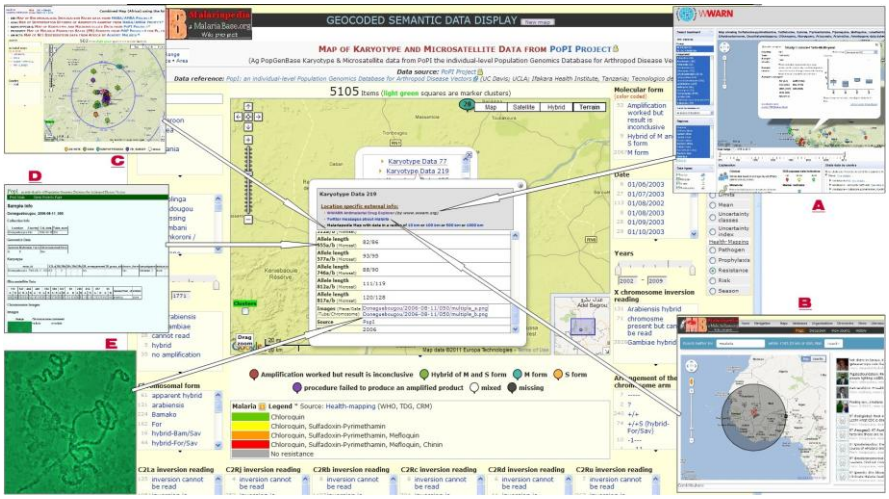

Figure 3. Location and data specific external links Experimental results from one data point of the Karyotype data set (see Table 1) map are shown with screenshots of linked information: A) WWARN Explorer window centred at the data point coordinates. B) Malaria related Twitter messages from around the data point. C) Malariapedia map showing data from all data sets within a radius of 100 km of the data point. D) Result summery page from data website for country, area, tube or allele. E) Chromosome images from data website.

[10] http://www.wwarn.com/

References

Barret, J. Daniel. 2008. MediaWiki. O'Reilly

Huynh, F. David, Karger, R. David and Miller, C. Robert: Exhibit: Lightweight Structured Data Publishing, WWW2007 Conference (2007)

Last of the Wild Data Version 2, 2005 (LWP-2): Global Human Footprint data set (HF). Wildlife Conservation (WCS) and Center for International Earth Science Information Network (CIESIN).

Markus Krötzsch, Markus, Vrandečić, Denny, Völkel, Max, Haller, Heiko and Rudi Studer, Rudi: Semantic wikipedia. Journal of Web Semantics, 5:251-261 (2007)

Olson, D. M, E. Dinerstein, E.D. Wikramanayake, N.D. Burgess, G.V.N. Powell, E.C. Underwood, J.A. D`amico, I. Itoua, H.E. Strand, J.C. Morrison, C.J. Loucks, T.F. Allnutt, T.H. Ricketts, Y. Kura, J.F. Lamoreux, W.W.Wettengel, P. Hedao, & K.R. Kassem. Terrestrial Ecoregions of the World: A New Map of Life on Earth. BioScience 51:933-938 (2001)

Acknowledgements

Work presented has been funded by the Max-Planck-Institut for Molecular Genetics, Berlin. Semantic MediaWiki extensions are made available under an open source license by a large community of programmers. We are grateful for the permission by Against Malaria to use their data sets and to WWARN for their permission to link to the WWARN Explorer.

Appendix

Minimal requirements

1. End-user application. The wiki (and the maps result display) is an end-user application as it can be freely accessed and queried via the internet. After the development phase registered users will also be able to edit. It serves as reference page for general users and as research tool for domain experts.

2. Information sources. Data are retrieved from a variety of sources in different formats under diverse ownership and control. Geocoded data come from support data of published research projects and additional data sets can be entered by users. Overlay data can be added from any suitable data file or even suitable images and country specific data collections. Linked data can also be highly heterogeneous. Geocoded data sets entered into the prototype consist of more than ten thousand multi-value semantic properties from published scientific papers.

3. Meaning of data. The meaning of the data from the different sources is essential for the conversion to semantic properties and the access via the semantic query language of Semantic MediaWiki which is based on OWL2EL. Data are processed

to unify common fields with identical meaning and to provide an interactive map display for presentation, filtering, inter-linking, and analysing.

Additional Desirable Features

User interface. The application features a user-friendly and interactive interface to deal with potentially large data sets and provides tools for the user to inspect specific data items and combine them with other internal or external data. It allows to access a variety of data from within the interface.

Scalability. Semantic Mediawiki can store and handle millions of properties. A number of software modification have been implemented to enable the map display system to work with sets of up to ten-thousand data items with reasonable speed and without Javascript time-out warnings.

Novelty. The systems provides through special modifications and extensions of existing semantic wiki extensions a novel way for sets of geocoded data stored as semantic multi-value properties to be displayed and analysed and to easily enrich profiles of specific geographical areas from an extendible list of sources.

Functionality. The main aim of the system is to connect related data from a large number of diverse sources to provide a more complete view of specific areas and research results.

Contextual information is used for ratings or rankings. The display of map markers in terms of size or colour can depend on dat values and directly reflect stored information. Contextual information is also used for filtering.

Multimedia documents. Data items can include images and videos which are integrated in the data display besides normal text and hyperlinks (in case of videos they can be played in-line).

Dynamic data. While most of the used data contain static information downloaded and stored on-site, data sets can also be automatically updated via a regular data download or retrieved via web services.

Accuracy. The data are usually either from peer-reviewed scientific publications or curated scientific databases. For all data used in a map display, their source and if available a reference is provided.

Range of devices. The system is regularly tested on a number of available web browsers on Linux, Windows and Mac OS computers.