



# Text Mining for Marine Ecological Genomics – A First Approach



JACOBS UNIVERSITY



Renzo Kottmann<sup>1,2</sup>, Thierry Lombardot<sup>1</sup>, Jacek Błażewicz<sup>3</sup> and Frank Oliver Glöckner<sup>1,2</sup>

<sup>1</sup> Max Planck Institute for Marine Microbiology, Germany

<sup>2</sup> Jacobs University, Germany

<sup>3</sup> Poznan University of Technology, Institute of Computing Science, Poland

## Introduction

"**Marine ecological genomics**" can be defined as the application of genomic sciences to understanding the structure and function of marine ecosystems. Nowadays, numerous complete genomes of marine prokaryotes (*Bacteria* and *Archaea*) are available in public databases, and more metagenomics-derived DNA sequences are produced in many ecological studies.

Understanding the function of microorganisms in the environment can be achieved if geospatial and ecological parameters are integrated with DNA sequence data. This allows investigating microorganisms in their natural context.

The information about the natural context of genomic data can only be found within scientific publications and hence have to be mined from literature. In a diploma work an algorithm for the extraction of geographic location descriptions of metagenomic samples was developed. It achieves a balanced F-Measure of 69%. Currently, the European research project **MetaFunctions** is developing a combined Information Retrieval and Information Extraction System to gather **contextual data** of genomic sequences of environmental relevance in the marine ecosystem. The objective of this Text Mining System named **Poseidon** is to extract more environmental entities in a framework including expert curation. The extracted data will be integrated into Genomes Mapserver (Figure 2). See <http://www.megx.net/gms>

## Information Extraction

Geographic location of a genomic sample is the most important information to be extracted.

- Diploma work: Extraction of the geographic location from metagenomic studies
- Manual inspection revealed that almost all geographic location descriptions are given in single sentences (Sample Location Sentence = SLS)
- Question: Which words occur with a significant frequency only in SLS?
- Condition: Take words that are not themselves part of a geographic description
- Method: **Pearson's  $\chi^2$  test**

## Method Outline

Preprocess:

- Calculate  $\chi^2$  for each word that occurs  $\geq 5$  times in SLS

Main Process:

- Tag each  $\chi^2$  significant word **w** in document
- Calculate average **a** of  $\chi^2$  values for each sentence
- Weight **wg** of sentence = **a** \* num(**w**) \* DSW
- SLS = sentence with **wg**  $\geq$  threshold **t**

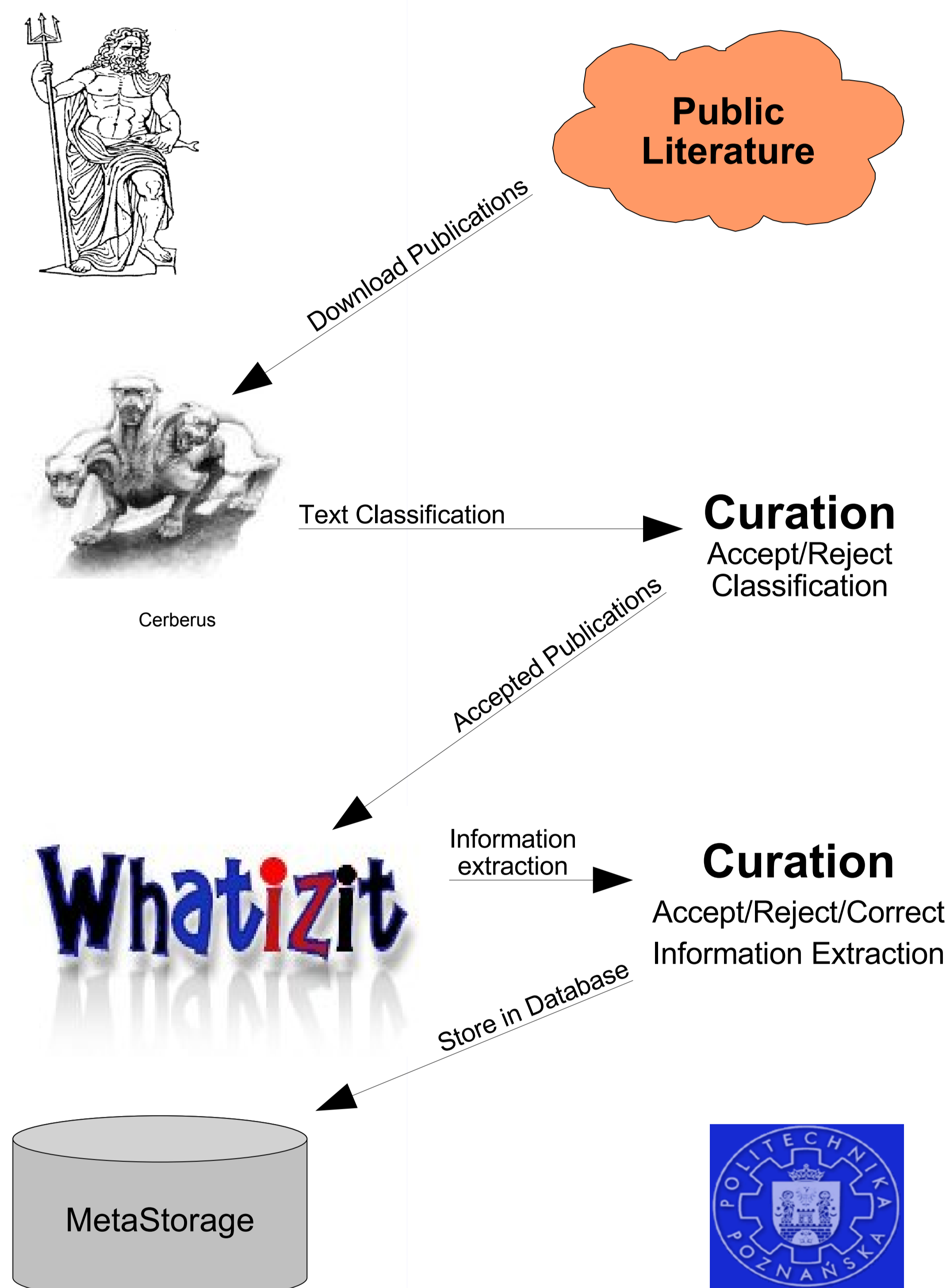
## Result

- Precision 0.72
- Recall 0.67
- Balanced F-measure 0.69

## References

Lombardot T, Kottmann R et al. Megx.net -- database resources for marine ecological genomics. Nucleic Acids Res. 2006 Jan 1;34 D390-393.

## Poseidon Text Mining System



**Figure 1** Simplified workflow of the Poseidon Text Mining System. The classification and Information Extraction step involves expert curation in order to avoid entering false positive- and false negative results into the MetaStorage database.

## Text Mining

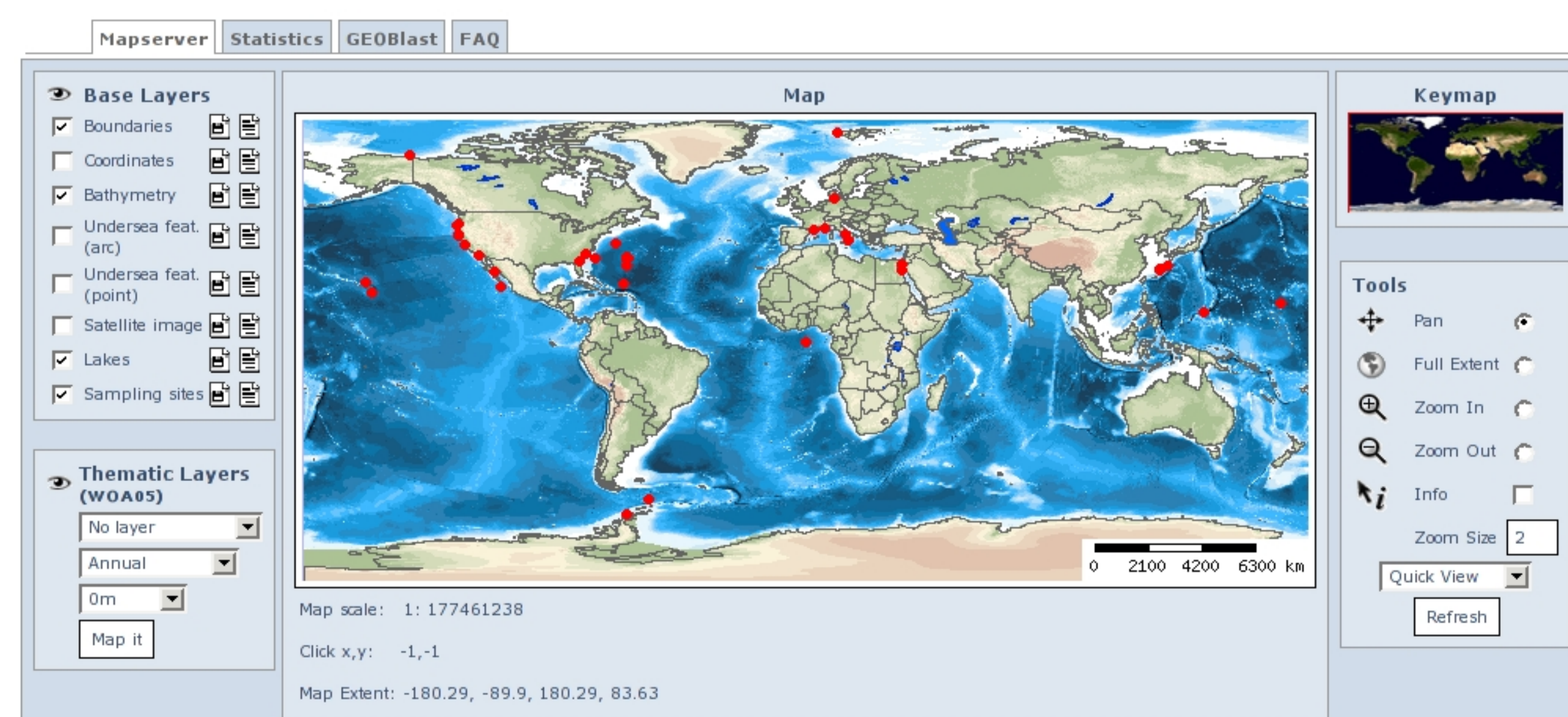
The Poseidon Text Mining System combines Information Retrieval and Information Extraction of environmentally relevant data from scientific publications. It is actively developed at the Poznan University of Technology. The system consists of three parts:

- 1) Remote access to publicly available, open access full-text articles
- 2) Classification of texts into categories:
  - Metagenome / non-metagenome articles
  - Marine / non-marine biological study sites
  - Based on simple keyword lists
- 3) Information Extraction
  - Use of the Whatizit system
  - Improved geographic location extraction
  - Extraction of environmental parameters such as depth, temperature, oxygen concentration

## Genomes Mapserver

The Genomes Mapserver is currently developed within the European research project **MetaFunctions**: "Environmental- and meta-genomics – a bioinformatics system to detect and assign functions to habitat-specific gene patterns". This project aims to elucidate the role of genes with yet unknown function by analysing the sequences within their environmental context.

It is the first **Geographic Information System (GIS)** that systematically integrates (meta)genomic sequence data with contextual information extracted from heterogeneous sources like publications, web pages, and global data streams from *in situ* measurements and remote sensing.



**Figure 2** Screenshot of the Genomes Mapserver main page. Red dots represent all currently available marine genomic and metagenomic sampling sites.

## Acknowledgements:

The MetaFunctions project is supported by the European Commission within FP6, under the contract number NEST-511784

Contact: [rkottman@mpi-bremen.de](mailto:rkottman@mpi-bremen.de)



### Services

#### Genomes Mapserver:

- Allows to zoom from world map view to sequence view
- Allows to map environmental parameters to sequences in different time scales
- Supplements map view with thematic layers like chlorophyll
- Interpolate environmental parameters for any sampling site

#### Geographic BLAST:

- Shows distribution of genes on the world map based on sequence similarity searches with BLAST
- Reports basic statistics for the hits and highlights potential outliers see (Fig. 2)
  - Gives information on the environmental context for each hit

**Table 1** Statistics of environmental parameters as reported in the scientific literature for a sample. Only reports with geographic coordinates were considered (100%).

Parameter Name	Reported [% of samples]
Latitude/Longitude	100
Date of sampling	51.61
Depth	66.13
Volume of sample	29.03
Temperature	16.13
pH	4.84
Oxygen	8.06
Salinity	9.68

### Status

Genomic sequences of 60 samples from 52 distinct sampling sites are available.

#### • On site data:

- Semi-automatic extraction from the scientific literature and public sequence databases
- 36 samples are from ocean waters
- 24 samples from sediments.
- Not all environmental parameters are reported:
  - Less than 70% of the samples have depth (see Table 1)
  - 10% of the samples have salinity
  - No single sample has more than 5 parameters.
- The highest percentage of samples has 2 parameters.

### Conclusions

- The Genomes Mapserver provides services for genomic sequence analysis in an environmental context
  - The integration strategy allows to tackle ecological questions like:
    - ➔ What is the environmental distribution of genes?
    - ➔ How do they influence the global cycling of matters?
    - ➔ Do habitat specific genes exist?
- The publicly available environmental data is sparse and not consistent:
  - ➔ Need to set up a consistent data retrieval and storage system for on site usage