Marine Microbial Biodiversity, Bioinformatics & Biotechnology

Grant agreement n°287589

Acronym: Micro B3

Start date of project: 01/01/2012, funded for 48 month

Deliverable 5-10

Micro B3 Information System Database Documentation

Version: 1.0

Circulated to: WP5

Approved by: Prof. Frank Oliver Glöckner, 2014-07-18

Expected Submission Date: 2013-31-12

Actual Submission Date: 2014-07-21

Lead Party for Deliverable: MPI-MM

Mail: rkottman@mpi-bremen.de **Tel.: +49 421 2028 974**

Public (PU)	X
Restricted to other programme participants (including the Commission Services) (PP)	
Restricted to a group specified by the consortium (including the Commission Services) (RE)	
Confidential, only for members of the consortium (including the Commission Services) (CO)	

The Micro B3 project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no 287589 (Joint Call OCEAN.2011-2: Marine microbial diversity – new insights into marine ecosystems functioning and its biotechnological potential.

Summary

The work on Micro B3 Information System's Database Documentation is based on the work performed in all tasks of work package 5 with main work done in task 5-3. The main tangible product is the Microbial Ecological Genomics Database (MegDb), which serves as the main data integration platform, as part of the Micro B3 Information System. MegDb includes several extensions which allow efficient integrated access to molecular sequences and environmental data. It is also used in an architecture which bridges the gap between analytical pipelines and storage of data in a scalable manner.

Table of Contents

Introduction

One central theme of Micro B3 project is to enable open access and transparent data flow for marine microbial ecosystems research and biotechnology. This data flow includes contextual data about the environment as well as sequence data from initial sampling (Task 5-1) to web based end-user access (Task 5-4, 5-5, 5-6) as well as to various data products derived from the data by the various analysis pipelines (Task 5-2).

Here we describe the Microbial Ecological Genomics Database (MegDb) underlying the Micro B3 Information System. The aim of MegDb is to guaranty consistent storage of all data needed for general marine microbial ecosystems research and biotechnology as well of for the special use case Ocean Sampling Day (WP2). The technical design and implementation also draws from conclusions and requirements of WP 3 (marine environmental data), WP4 (standards and interoperability) and WP6 and WP7 which input requirements.

Overview

This deliverable focuses on documenting the main design decisions and approaches taken in the implementation of MegDb. Section on “Availability” includes references to documents describing the technical implementation details.

MegDb as a data integration platform

The Micro B3 Information System and the role of MegDb as a data integration platform is best described from a data flow perspective (Figure 1). The data flow has three parts and is divided into seven steps:

I: Data convergence is the process of transforming a multitude of different contextual, environmental and microbial sequence data into a common data model. It comprises the following steps:

1. Discovery and generation of data
2. Harvesting data
3. Filtering data

II: The data integration part delineates the border between a data convergence and data divergence

4. Integrating data

III: Data divergence

5. Augmenting data
6. Analyzing data
7. Acting and visualizing data for stakeholders

Additionally, Figure 1 depicts which WP5 tasks produces software for which step and shows relations to other work packages.

The Microbial Ecological Genomics Database (MegDb) stores the generated data in an integrated data model. This allows further augmentation and analysis of the integrated data as well as dissemination to scientists to gain new information and knowledge. Moreover, MegDb is the platform from which newly generated data gets transferred to other infrastructures like ENA and SeaDataNet for long term storage and archiving.

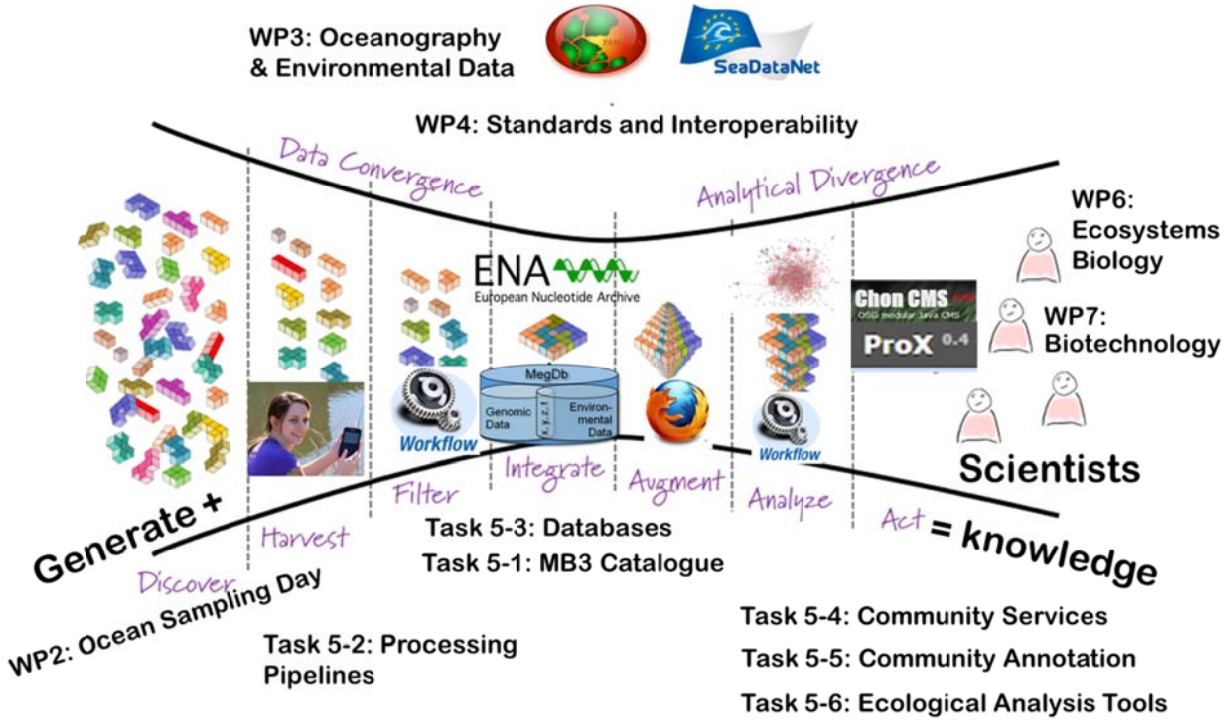

Figure 1: MegDb as the integration platform in the context of Micro B3 Information System depicted from a data flow perspective. Shown are the WP5 tasks producing software for the data flows. Additional relations to products, deliverables, and outcomes from other Micro B3 work packages are depicted (from left to right). MegDb (middle) is the main software for the data integration step.

High level data model

MegDb currently comprises 16 Schemas[1] from which the "core" schema is the schema in which external data gets integrated. The "core" schema itself has 33 tables with numerous relationships. Figure 2 describes the main entities and their relationship reflecting the main features of MegDb on a more abstract level[2]. The main entity is the "Samples" entity which holds all data of an 'Environmental Sample' which is defined as:

> ***The material extracted from nature which either constitutes a biophysical environment or is an organism.***

In the case of Ocean Sampling Day, the environmental sample is simply a certain amount of water collected. Having the concept of environmental sample allows modeling the cascade of sample transformations and/ or subsampling, until final DNA sequences and

[1] Here schema is used the same way as in PostgreSQL: http://www.postgresql.org/docs/9.3/interactive/ddl-schemas.html

[2] The detail documentation of the database implementation can be found http://resources.mpgx.net/megdb-doc/index.html

environmental measurements are obtained, as entities derived from the 'samples' entity. E.g. the OSD Sterivex filter (Figure 3) are modeled as one own entity with relationship to 'samples' (not shown here). For ecological and legal considerations the most important relationship is the final relationship of 'samples' and 'DNA sequences'. Figure 2 shows cases: a) metagenome sequences (bottom) and b) genome sequences from cultured organisms (top). The example of genome sequences illustrates that if needed several entities can be put in-between 'samples' and 'DNA sequences' modeling the cascade of sample transformations derived from 'samples'. Therefore, this data model is extensible and can evolve over time.

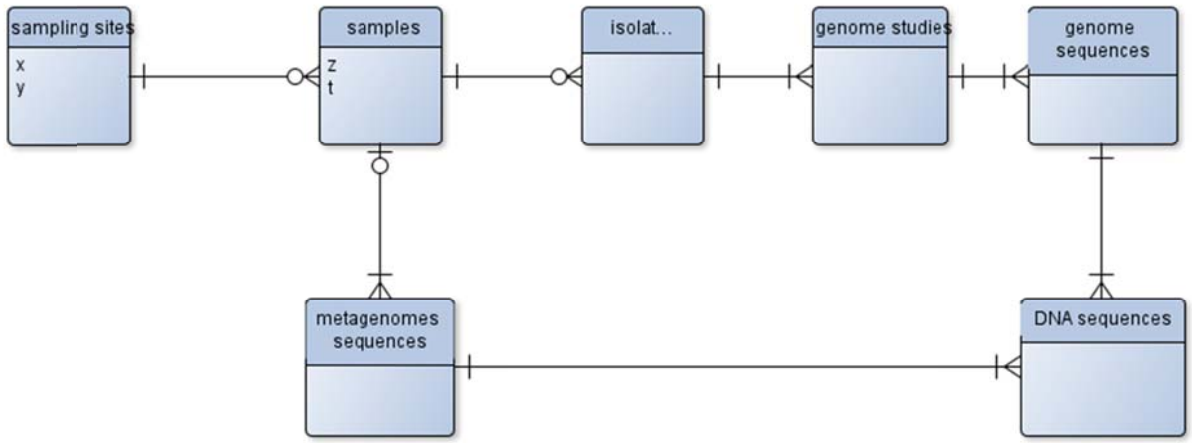

Figure 2: High level Entity-Relationship Diagram showing the main entities and their relations in the data model of MegDb.

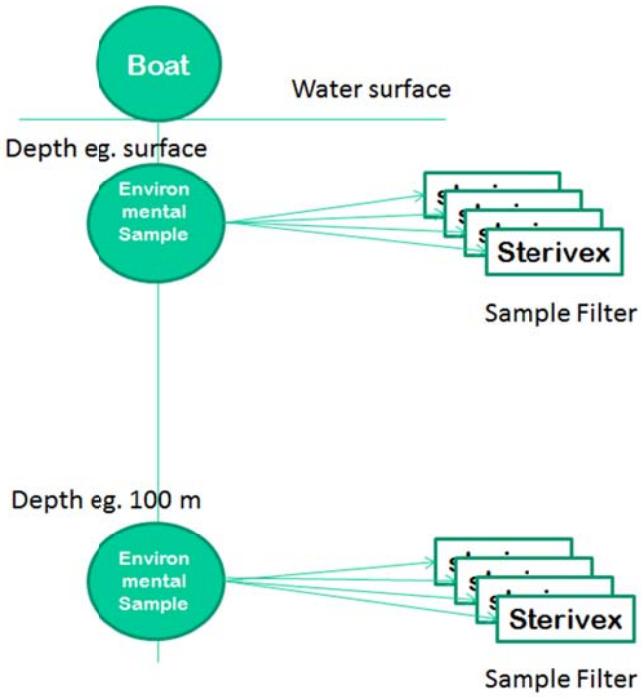

Figure 3: Schematic view of an OSD sampling event. Participants on a boat or at a beach deploy sampling devices in the water and collect 'Environmental Sample'. After sample collection the water gets filtered in 'Sample filters'.

Geospatial Integration

The concept of 'Environmental Sample' is also key to geo-referencing DNA sequences and relating them to environmental data. There are many valid ways to store geographical co-ordinates. MegDb uses the 2D geometries of PostgreSQL's PostGIS extension. Therefore, MegDb has the 'sampling sites' entity geo-spatially referenced by longitude = x and latitude = y in WGS 84. There is a one to many relationship to 'samples' entity which has the attributes z = sampling depth and t = sampling time. In other words many samples can be extracted from a sampling site at different depths and times (Figure 3).

This way it is a matter of using the geo-spatial functions of PostGIS to query e.g. the annual temperature of all sampling sites where metagenomes were derived from. This can then be visualized by e.g. the Genes Mapserver of the Micro B3 Information System (see Figure 4).

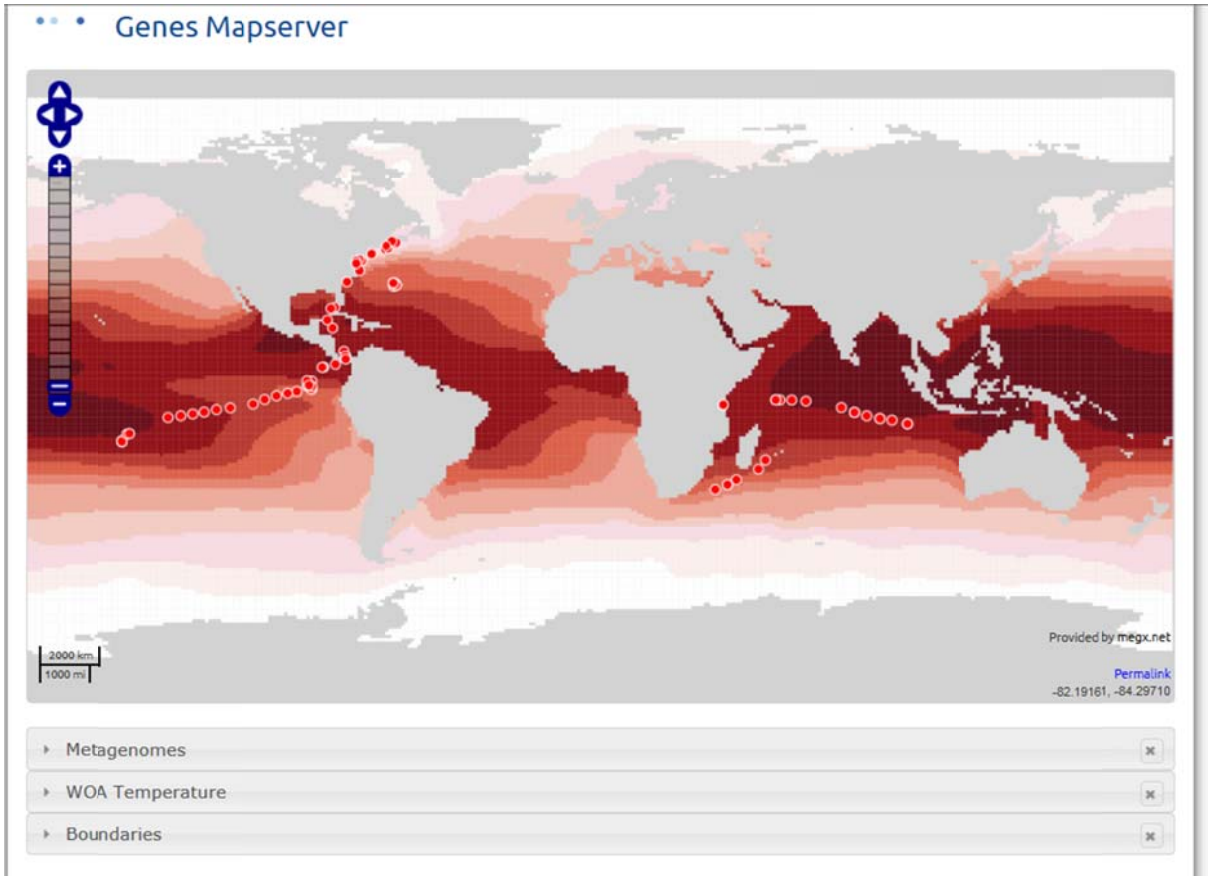

Figure 4: Screenshot from Genes Mapserver (http://mb3is.megx.net/gms). Shown is annual sea surface temperature overlaid with metagenomes sampling site (red dots).

Combining data storage with the analytical pipeline

In software architecture there is often a gap between the storage of data and the analytical tools/pipelines which make use of the data or generate new data on existing data.

We developed a message queuing architecture bridging this gap efficiently, which helps to solve asynchronous batch processing of live transactions (Figure 5). We use this architecture for the Prokaryote analysis pipeline (see deliverable D5.8), Geographic-BLAST and the Microbial Metagenomic Trait workflow (http://mb3is.megx.net/mg-traits). The whole

system is based on PgQ, which is a queuing system based on PostgreSQL developed by Skype and usually used for processing thousands of jobs per second[3].

Using this architecture with the example of a BLAST job, we can simply have an entity 'blast_run' in MegDb. When a user starts a BLAST job, the web service simply performs an insert in 'blast_run' including all information needed for running BLAST. PgQ polls the database periodically checking for new entries in 'blast_run' if one or more entries appeared since last check, it will take all information and start one or more BLAST jobs. In case of the Micro B3 Information System installation at MPIMM the BLAST job gets distributed on the compute cluster of MPIMM. Once BLAST is finished the results get written into MegDb's 'blast_results' entity. During this whole process the web interface periodically polls 'blast_results'. If results are available, it will render the results on a web page including also the case where a BLAST job was unsuccessful due to errors.

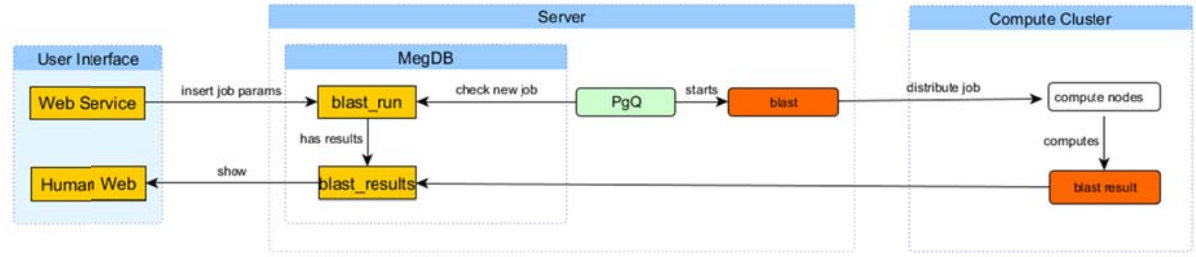

Figure 5: Simplified job queuing architecture with BLAST as an example.

This architecture has the main advantages that the web, database and computing components are decoupled which allows e.g. to change the compute cluster by changing the configuration of the BLAST job in a live system. It allows to use MegDb e.g. for auditing the use of compute resources. More importantly, it allows integrated analysis of e.g. BLAST results by means of querying the database to answer questions like e.g. "What is the geographic-distribution of similar genes BLAST found?" or "Does the order of BLAST hits correlate with temperature gradient?" This saves the need for implementing and maintaining scripts for each question. The whole architecture scales linear: twice as many BLAST jobs need only twice as many compute nodes to achieve the same results in the same amount of time. The scaling is transparent i.e. no changes to the system is needed except of adding compute nodes. Scaling the database would require sharding. However, this is unlikely to be needed because it only imposes minimal load and resource consumptions on the database level so that a single database server can handle thousands of BLAST job requests per second without significant decrease in performance.

PostBIS: efficient storage of sequence data

PostgreSQL Bioinformatics Information System (PostBIS) was developed by MPIMM. Main development was by Michael Schneider in his master Thesis "Efficient Representation of Biological Sequences in a Relational Database Management System" in collaboration with Prof. Dr. Stefan Kurtz University of Hamburg.

[3] See https://wiki.postgresql.org/wiki/Skytools for technical detials.

PostBIS is an open-source PostgreSQL extension project to facilitate sequence-based Bioinformatics in PostgreSQL. It offers:

- highly efficient specialized sequence data types incl.
  - nucleotide sequences
  - protein sequences
  - alignments
- additional domain-specific functions
- indexing of sequences

The storage requirements for sequence data using PostBIS are ~2 bits per nucleotide base which is around 25% of PostgreSQL native text data type and ~5 bits per amino acid compared to ~9bits PostgreSQL native text data type (Figure 6).

The loading of complete sequence databases using PostBIS is also in all tested cases significantly faster (Figure 7).

All in all, query Access is around 1000x faster compared to using PostgreSQL native text data type.

PostBIS is used in production since more than a year and no issues were encountered.

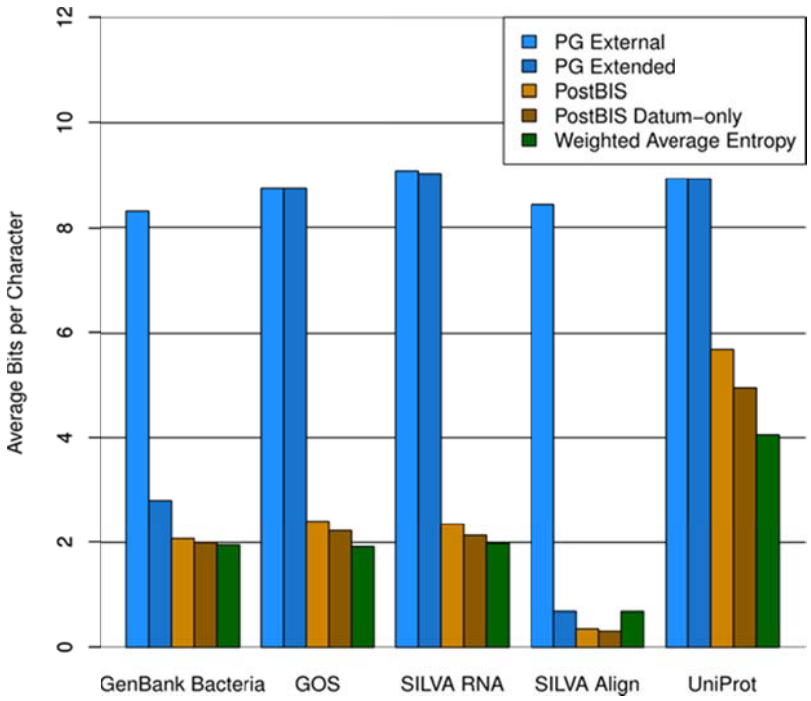

Figure 6: The PostBIS space-efficiency (brownish) compared to PostgreSQLs built-in compression (blue). The graphic shows results for complete genomes (GenBank Bacteria), short reads (GOS), RNA sequences (SILVA RNA), aligned RNA sequences (SILVA Align) and amino acid sequences (UniProt).

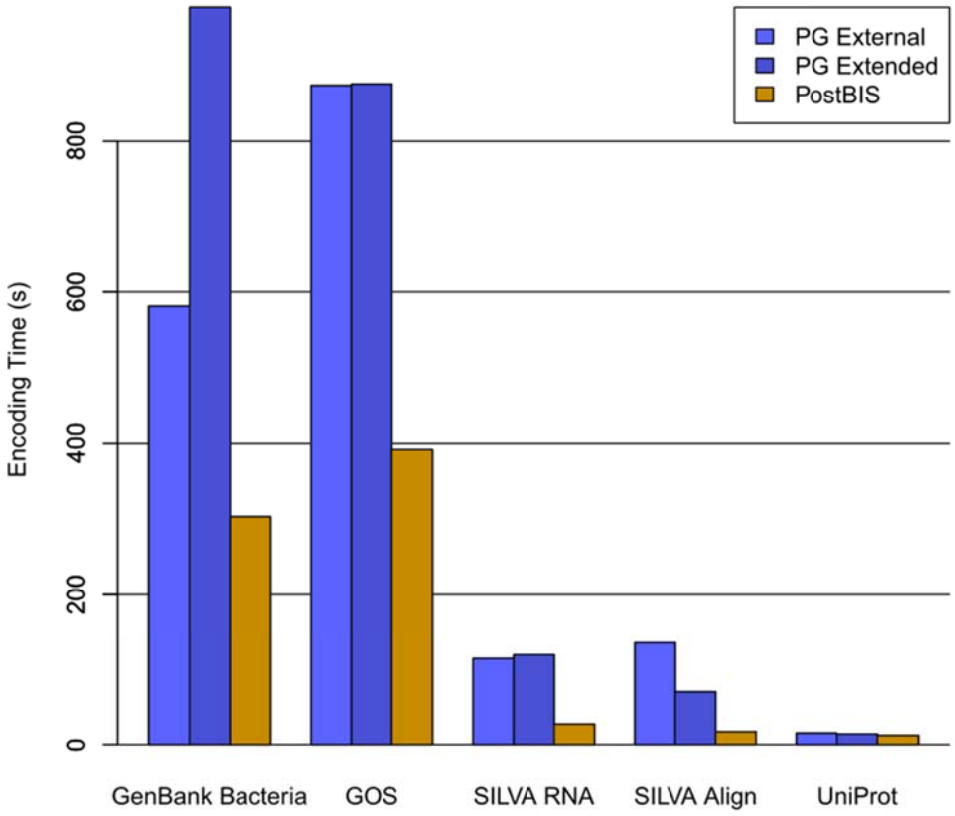

Figure 7: Encoding time (s) of complete sequence databases without and with PostBIS. Tested data are: GenBank Bacteria, Global Ocean Survey metgenomes (GOS), SILVA RNA unaligned and aligned and UniProt protein database.

Further Development

Future development includes more efficient storage of environmental data. Currently, we are collaborating with LifeWatch Belgium to perform several performance tests on different storage layouts and usage scenarios. We are also improving the storage of sequence data for optimizing queries which ask for sequences from certain environmental conditions. On-going work is on in-cooperating and integrating interoperability structures developed by WP3 and WP4. This work will be reported in D5.13.

Availability

- Data model implementation documentation: http://resources.megx.net/megdb-doc/index.html
- PostBIS: https://colab.mpi-bremen.de/wiki/display/pbis/PostBIS
- A dump of the database structure is available on request rkottman@mpi-bremen.de