Document iMB2/2012/5

---

# iMarine Biodiversity Cluster

## 1. IMARINE BIODIVERSISTY CLUSTER VISION STATEMENT

**The iMarine biodiversity cluster summary vision statement is proposed as:**

- **To provide a complete and diversified data and analytical tool kit to build species distribution and species richness maps that support the Ecosystem Approach to Fisheries Management and Conservation of Marine Living Resources. The vision includes all the work done from preparing good quality data to exploiting maps, visualizations and analyses, for biodiversity management purposes.**
- Outlooks: A Species Occurrence data manager across institutions, transformation of datasets between organizations sharing reference data standards based on Darwin Core, Spatial data analysis, and geolocation based information services, and openModeller.

**High level vision** (medium term e.g. 5 years)

The Long Term Goal of this cluster is to offer a single entry point to aquatic species occurrence data, and iMarine VREs that provide filtering, visualization and geo-statistical analysis. The combination in one e-Infrastructure of biodiversity and environmental data with predictive modeling algorithms enable cross-domain understanding of the spatial patterns in species occurrence, and their reaction to large scale changes to the natural environment caused by human interventions (fisheries, protection), and climate change. The openness of the e-Infrastructure enables scientist to develop their own algorithms, and use the computing power and combined expertise in collaborative science. This enhances the flow of data, by making us of a limited number of well-understood protocols that includes metadata, thus building confidence in data quality.

**Workable vision** (achievable within iMarine project – 2 years)

The quality of species occurrence data products has been enhanced through a capacity to access regional and global sources of taxonomy and biodiversity, and of remote sensing and in-situ data, and to exploit those sources for data discovery, data reconciliation, spatial analysis, and distribution modeling. OpenModeller has enriched iMarine with a range of species distribution and biodiversity prediction tools, and together with “R” offers a potential for executing predictive and analytical algorithms. A few scientific user communities have successfully validated the outcome of the predictive and analytical models.

Action from Advisory Board members:

Advisory council members are asked to provide an opinion on such vision statement. These opinions will be delivered during the second iMarine Board Advisory Council meeting, towards the formulation of a final version.

## 2. USE CASES

### Use case groups

The full activity in the cluster can be described as from data to maps. The business cases are organized along that flow and fitting the development of dedicated VRE rather than scenarios encompassing all VREs. This fits more with the reality where some groups are more involved in handling and preparing data whereas other s are only using the outputs for further analyses or management activity. However, the full workflow must be integrated so some groups can perform their research work from data handling to map use (e.g., OBIS, AquaMaps, … communities).

- Taxonomic/biological names-related Data Management Tools
- Darwin Core related Data Management Tools
- Species Distribution Data Management Tools

### Taxonomic/biological names-related Data Management Tools

Background

Most species information, including occurrence data, is hooked to a scientific species name that is coined to designate a taxon at specific level. Data from various sources can be integrated for a species only because they are attached to the same name. Other alternative would be to attach information either directly to specimens, or to gene sequences (often done for microbes); these alternatives will be explored later. However both taxonomy (the way to split the living organism diversity in well-defined and identifiable species in a hierarchical classification) and nomenclature (the proper way to assign unique names to the different taxa) is a work constantly in progress leading to the difficult situation where a species may be designated by several names (synonymies), or that a name may designate several species (homonymies), along the time or simultaneously according to different authors. Several major efforts are on-going to collect and collate all taxonomic names.

The Catalogue of Life (CoL) aims at gathering all species names and their synonymy and homonymy relationships so that interoperability between various information datasets can be automated. As of April 2012, CoL gathers the names for almost 1.4 million species over the 1,9 million estimated to be known to science from more than 110 separated Global Species Databases. It is most important to have an easy access within the iMarine ecosystem to facilitate huge dataset compilation from various sources (e.g., not using the same name for the same species) or to link them to other sources.

The World Register of Marine Species (WoRMS) is engaged in a similar venture, but specifically for the marine domain. WoRMS and CoL are collaborating, and WoRMS is the main provider of GSDs to CoL for marine groups; it is WoRMS that is used as the primary (but not exclusive) name reference by OBIS. It is important to realise that CoL and WoRMS are collaborating, but are separate activities; the collection of names they hold obviously overlaps, but significant amounts of names are held by only one of the two systems.

Apart from WoRMS and CoL, which will be the main sources of taxonomic names, there are several other systems that should be taken into account. The Integrated Taxonomic Information System (ITIS) is an initiative of the federal government of the USA, and is the oldest player in this game. It is one of the two partners in the Catalogue of Life, the other being Species 2000. LIke between WoRMS and CoL, there are significant number of names that are only available in ITIS.

The National Center for Biotechnology Information (NCBI) is one of the organisations hosting GenBank. In order to manage the genetic sequence information, they also keep a register of taxonomic names, which is compiled independent from CoL, WoRMS or ITIS. As such it is an important 'second opinion' on biological names. Especially for microbes, the NCBI is often the best source.

Last but not least, the Interim Register for Marine and Non-marine Genera was created within the OBIS community, with the objective to enable to discriminate between marine and non-marine, and between fossil (extinct) and extant taxa.

Anticipated impact

- Significantly decrease the human resources spent to cross-check names over multiple datasets;
- Increase the quality of the primary data sources by sending them feed-back on issues (up to provide the service for themselves to check their data);
- Integrate the developed tools in the efforts developed by the Biodiversity Informatics community to establish a Global Name Architecture in order to implement in the VRE the technological solutions proposed by this project. The main goal of this topic is to identify, in a formal way, some critical characteristics for describing species coming from different data sources, in order to understand when two entries refer to the same one. Such identification task is not trivial because of the deep differences in the nomenclature protocols which are followed in different areas of biology. Nomenclature can vary moving from Zoology to Botany and Bacteriology. This section has the daring scope of investigating the margins for building a merging algorithm which solves the above issue, as an automatic solution has never been found up to now.

Proposed measures and status

- Develop a VRE for species and species names discovery within datasets available in the infrastructure (together or not with discovery of occurrence datasets). Start with CoL, WoRMS, ITIS, and OBIS. OBIS is a different type of dataset with respect to names, whereas the 3 first are supposed to be references for names, and whereas OBIS use these names. However, OBIS receives names that are not registered and not validated, so all names in OBIS are not in the three others altogether.
    - o Status: the Biodiversity Research VRE was implemented using CoL, WoRMS, ITIS, and OBIS and demonstrated during the technical session.

- Develop a VRE to reconciliate/synchronize names between datasets;
  - Potentially could be extended to a check name VRE for primary data providers to check their names themselves (since the primary providers are the ones who know their own data best), before they provide to aggregators. See also quality control below.
  - Potentially could be extended to a check name VRE for data users to check the names they use.
    - o Status: To be created. Work plan to be established.

New proposed measures

- The CNR team has collaborated with the University of Cardiff in the framework of i4Life to implement their cross-checklist algorithm. The Board should give its opinion to make this VRE available to the iMarine community. CNR to evaluate the difficulty of it. The algorithm is different from the taxamatch one but complementary because it addresses another case when names must be reconciled.
- Check GBIF activities in that domain.
- Check the progress of GNA, and maybe, establish contacts.

**Darwin Core related Data Management Tools**

Background

The production of Darwin Core data is the main activity expected to be supported by the D4Science infrastructure. The CoP already provides software that can be evaluated, although these have been used with varying results in the past. The Community seeks in iMarine the development and release of an environment for the management of marine species observational data that offers:

- Selection of data from a source (browse repository, filter).
- More flexibility in occurrence datasets that can be used today (OBIS, ...),
- Interactive occurrence selection in a GIS-viewer (e.g. sliders to-from dates),

- Occurrences data cleaning: DarwinCore is already the standard for exchange, but there is no standard to exchange corrections (e.g., if we find an error, how do we send the feed-back to the provider in such a way that he can use what we send easily: obviously data are structured by DwC but what about the indications of what was changed and why).

  This includes also:

  - the private data, they will be integrated through the DwC, but what about conveying assessment on quality, on possible traps and known issues.
  - data editing incl. through interactive maps.
  - environmental data associated with biological observations, and to be used in performing the extrapolations between observed (point) data to the distribution range (polygons)
  - Statistics on probabilities, e.g., a summary on a given area, bootstrap or any other robustness measures.
  - change of scale (e.g. size of grid) up to changing the grid system (for an equal area grid system: the current cells have 4 times more surface near equator than near the poles) and provide transformation standards.
  - eanaging shape files and related statistics, graphs, etc: can be a transect, along a coast line, a geographical area. This includes sharing maps, and analyses.

<u>Anticipated impact</u>

- If the project manages to provide a solid environment for the management of biodiversity data, the potential interest in an even wider community will certainly be raised. However, to make an impact, several conditions must be met: Open source, using well known technologies, etc.

<u>Proposed measures</u>

- Dynamic links with WoRMS, CoL, ITIS, NCBI and other code lists will facilitate integration and quality control of taxonomc names; see also names-based activities below
  - <u>Status</u>: Done and demonstrated during the technical session.
- Tools to check GBIF holdings for marine data that are not available in OBIS will allow to detect and ingest these datasets in OBIS, and make OBIS a one-stop shop for marine biogeography
  - <u>Status</u>: Done at name level and demonstrated during the technical session. To check if it is possible to elaborate reports. Checklists can be created and exported in DwC-A.
- Tool for visualising the data - most important through a map interface
  - <u>Status</u>: Done and demonstrated during the technical session. Point lists can be exported in DwC and csv formats
- Tools for quality control, such as detection of duplicates; detection of outliers in environmental space...
  - <u>Status</u>: Development of a complete VRE: Statistical manager environment demonstrated during the technical session.

New proposed measures

- Use WoRMS as the complete classification as a thesaurus to be queried for higher level taxa that are not in the main levels, e.g., Crustacea.

VRE: Statistical manager environment

*Depiction of outliers*

Using clustering methodologies, there are two options using spatial distance or density, and 3 approaches: DBScan point density, Minimum point number, and distance between two points (Kmeans, Xmeans.

The result show that outliers depend on the algorithm used, and that notwithstanding of the examination of results by biologists, density-based algorithm are more suitable. Density include a part of temporal component which is important, and could be captured as well by linking in a closer way the environmental parameters at the most detailed spatio-temporal scale.

In total the Statistical manager handles 26 supported algorithms: 2 modelers, 3 clusters, 3 evaluators, 6 projectors, 12 transducers (see technical reports for details). In a short sentence, we should be able to get the closest values of the environmental parameters from one set of coordinates and date and time.

New proposed measures:

- Focus in the coming months on developing the tools to produce these detailed environmental parameters datasets. Relevant parameters are to be selected by the cluster partners and TerraDue and IRD to be contacted to prepare the datasets. It is suggested to start with the parameters used in the algorithm supported by OpenModeller.
- Display suggestion for using statistical tools in the relevant VREs, just like “buttons” were presented in the slides.
- Implement a weekly discussion between CNR and FIN to follow up development.

*Depicting of duplicates*

The duplicate issue is still a problem in GBIF. The idea is to produce tools to depict those and mark them as such. Five algorithms were tested: Merge, Inters, Nodupl, Onearth, InSea. Some use lexical distances on scientific names and authors.

New proposed measures:

- Implement quality control over the source datasets before testing then for inter-duplicates. Take examples in the business cases, e.g., tunas and deep seas.
- Which may mean to develop annotation processes, that could be use in general for correcting the records. To be explored by the cluster.

**Species Distribution Data Management Tools**

Only a few maintenance work was performed on the AquaMaps VRE. The complete run of maps should be performed during October, and serve for the basis of a paper on Lessepsian migrants (FIN).

New proposed measures:

- It is requested to the Board if the VRE developed under another European Project also involving CRIA, OpenBio, could be made available for iMarine partners. CNR precises that can be done without further efforts for iMarine.
- Make the assessment of the use of OpenModeller if made available.

## 3. INVOLVING THE COMMUNITIES

The main action to conduct in the coming months is to involve the iMarine partners and later their communities to use the infrastructure, if possible through the production of papers, e.g., using the models available, but also through real routine work, e.g., for name and point data reconciliation.

A new philosophy arose in the infrastructure for the creation of VREs: they be produced on demand for the need even for a few combining applications existing in dedicated VREs. Then the workspace become the main tool for linking VREs. This has to be promoted among the partners.

There is a proposal from the management team of iMarine to contract Edward Vanden Berghe to continue to bring in communities and to conduct some design work on specific areas of the Biodiversity Cluster. Contact were established during September resulting in a proposal document.