1) *Describe your laboratory's overall data entry process. When are data entered into a database? In the field? In the laboratory? At what point are individual tissue samples given a unique identifier? What information is kept with each individual tissue? What database system does your laboratory use? Filemaker, Excel, Access, Oracle?*

**Harilaos Lessios, Smithsonian Tropical Research Institute:**

Samples collected in the field are usually assigned a temporary field identifier. When they are received in the lab, they are assigned a unique 8-character (minimum) organism-ID (described below in #4) and each tissue or sub-sample receives a unique biorepository-ID. At this point a collection label is generated which includes the Organism ID#, Biorepository ID#, Species name, tissue code, collection locality, and a 2-D barcode generated from the Biorepository ID.

Information is added to the organism record in our database (usually via an Excel spread sheet generated by the field collector, which is error checked and formatted by the database manager) which includes taxonomic identification, the collectors name, detailed collection locality information including lat/lon coordinates if available, any physical descriptions, voucher pictures, and any additional notes (e.g. if the specimens were used in specific experiments). This comprises the "Organism Table". Original field collector's identifiers are included in the sample record for any physical sub-samples but can be included in the Organism table as general information. The database also generates links to daughter databases reporting the number and ID of any samples or subsamples, genetic data (sequence and genotypes), taxonomy, attached files and other metrics.

Each whole specimen, tissue sample, sub-sample, or DNA extraction is assigned its own Biorepository ID and a daughter sample record is generated. That sample record includes the associated Organism ID, Biorepository ID, Taxonomic ID, Tissue type and three letter tissue code, summarized Collection locality, preservation medium, and storage location. It also records the Biorepository ID of any parent samples (e.g. for a DNA sample, the tissue sample it was extracted from) or child samples. For DNA extractions we generally enter the type of extraction in the "sample preservative" field. These records are contained in the "Sample Table"

Our database is currently using a Filemaker 12 server.

**Rob Toonen, Hawaiian Institute of Marine Biology**

*When are data entered into a database?*

We normally write everything down in a notebook that is then entered in the field or immediately upon return into an Excel Spreadsheet that each student/post-doc is responsible for maintaining.

*At what point are individual tissue samples given a unique identifier?*

Usually at the time of sampling in the field, but occasionally when samples arrive in the lab. In general we use the formal ddmmyy001 for the samples to result in a unique identifier for each one.

*What information is kept with each individual tissue?*

It varies widely among individuals and each individual is responsible for their own record keeping. We have no standardized information required to be kept with individual tissue samples.

*What database system does your laboratory use?*

We had a Filemaker Pro database but no one kept it up, and we now have individuals each keeping their own data, and most frequently in Excel.

**Cynthia Riginos, University of Queensland**

All the data are entered into an excel database while in the field. We use the same number for collection and in the lab and each individual is given a unique number. In the lab we might append prefixes denoting species and location but keep the same 4 digit unique number.

**Luiz Rocha, California Academy of Sciences**

Data are entered in the database after tissue and specimens are deposited at the California Academy of Sciences (CAS) collections (at the lab, on average less than a month after we return from the field). In the field, every tissue sample and specimen is given a unique identifier (field number) that will be tied to the final CAS catalog number.

Information kept with each individual tissue include species identification, collector, determiner, GPS coordinates of location, date of collection, and additional notes. The original database is in Access, but the web interface is SilverCollection and we have the entire database stored as a Darwin core. We only use Access to enter data.

**Sophie von der Heyden – Stellenbosch University, South Africa**

We use Excel to maintain our sample data base. All samples are given a unique ID (usually determined by the collector, i.e. myself or students with my help) when the samples reach the lab. Each tissue samples is placed in ethanol, with an internal and external label and stored at -20C. We also keep separate data bases per year, i.e start a new one every January

**Gert Worheide, Ludwig-Maximilians-Universitaet München**

*When are data entered into a database?*

We normally enter the data in the field into an Excel Spreadsheet (attached) that forms the import template for our database, such that after return from the field the data can be uploaded in one go.

*At what point are individual tissue samples given a unique identifier?*

At the time of sampling in the field or when external samples arrive in the lab.

*What information is kept with each individual tissue?*

See attached Excel spreadsheet.

*What database system does your laboratory use?*

Custom-made mySQL relational database with webbased front-end and php queries.

**Rachel Gotanco, University of the Philippines Marine Science Institute**

Sample information data are entered in the laboratory

Samples are usually given a unique identifier in the field.

Information kept with each individual sample: location info (species ID, date of collection, name of collector, place, depth, general habitat)

Records currently in Excel

**Cecile Fauvelot, Institut de Recherche le Developpment, Noumea, New Caledonia**

I did not have a proper sample database as you could call it.

Most of my samples from previous works are in Perpignan in Serge's lab (and the current project are in collaboration with Serge, he has the samples), and since I arrived in New Caledonia few moths ago, I did collect only samples for population genetic studies, and boxes are stored, and we just have excel files (with basically Dascyllus aruanus and tridacna maxima samples from different places and time in New Caledonia).

Data are entered in the lab, coming back from the filed (easy for us, as our field is almost in front of the building ... so data can be entered everyday back from the field, in the lab)

Individual tissue samples are given a unique identifier, and tissue samples are stored in 80% EtOH, in individual tubes,and stored in 9x9 samples boxes

Usually, only the scientific name plus ind number (and sometimes location) is provided on or within the vials.

**Eric Crandall for Paul Barber, University of California, Los Angeles**

We have two separate sample databases, one for my laboratory’s collections and one kept in collaboration with Dr. Kent Carpenter to keep track of samples collected on the CTPIRE grant. Both databases were created with approximately the same fields and are maintained with roughly the same protocols, the main difference being that the CTPIRE samples reside in their country of origin, while those in my laboratory’s database are in Los Angeles.

*When are data entered into a database? At what point are individual tissue samples given a unique identifier?*

In the field, we collect population samples of 20-50 individual tissues per species per sampling site. These population samples are assigned a four-digit serial number upon their return to the laboratory and relevant metadata are recorded in an excel spreadsheet. For CTPIRE samples, this serial number is prefaced by two letters for the country of origin e.g. IN0123. Upon DNA extraction, individual tissues receive unique identifiers as decimal points of the four-digit population sample identifier e.g. XXXX.01, XXXX.02.

*What information is kept with each individual tissue?*

Individual tissues keep their individual serial number XXXX.XX through which all of their metadata can be tracked.

*What database system does your laboratory use?*

My laboratory’s database is kept in Filemaker, while the CTPIRE database is kept in excel.

2) *What types of sample metadata are collected? i.e. what are the names of the fields/columns in your database?*

**Harilaos Lessios, Smithsonian Tropical Research Institute**

*a) from the field*

Collector's Name

Collection Date

Collection Region

Collection Locality

Specific Collection Locality

Detailed Collection Locality info

Latitude

Longitude

Depth/Altitude

Country

Field Collector's ID

Tissue collected

Preservation medium

Sex

Physical Characteristics (description)

General Information

*b) in the laboratory*

Organism ID

Biorepository ID

Date Received

Sample Owner

Storage location information

*c) photo/voucher information*

Where available, digital files are identified by the field collector, and attached to the organism record as miscellaneous attached files. There is also a field in the sample table for voucher information if the sample originated in a different collection (for example, borrowed or subsampled from a museum specimen).

**Rob Toonen, Hawaiian Institute of Marine Biology**

It varies widely among individuals and projects.

**Cynthia Riginos, University of Queensland**

*Field:*

Indiv-ID          Species            Country          Region Site      local     habitat depth (m)        date (dd/mm/yyyy)    time     long-Deg                      Long    Lat-Deg                       Lat       Source for location (GPS, other)         sample-type (tissue, whole organism)          Preservative    Whole animal preserved separately (Y/N)      permit permit authority         errors   notes

*Lab:*

Person DNA extraction date   Extraction method       tissue left?       Extraction clean up?    stock' extraction strorage        CO1 sequence?           Primers used (seq)other          PCR product clean up            Macrogen order no.     GenBank Accession no.  Other: sequence?         Other: Primers used    Other: PCR product clean up Other: Macrogen order no.

*Other:*  
GenBank Accession no.          Msats  Other markers?           Lab notes

We keep photo vouchers for some species. They are stored on our server and labled by the id label.

**Luiz Rocha, California Academy of Sciences**

All of the above, we have a very complex database fully available here: <http://collections.calacademy.org/ich/>

**Sophie von der Heyden – Stellenbosch University, South Africa**

These are the rows we usually have ID number, species name, common name, Date collected, Where collected (inc GPS coordinates), photo taken (Y/N), name of person who collected, other notes (I.e. samples sent to collaborator, photos taken, non-targeted species, i.e. bycatch, vouchers deposted with collections etc

**Gert Worheide, Ludwig-Maximilians-Universitaet München**

Sequence data associated with each entry is stored in Geneious (at least that is the theory/intention).

**Rachel Gotanco, University of the Philippines Marine Science Institute**

Only field collection data are currently in the database. Genetic data have yet to be encoded in the database.

MorphoID\_ScientificName

CommonName\_English

CommonName\_Vernacular

CollectionDate

CollectionLatitude

CollectionLongitude

CollectionCountry

CollectionProvince

CollectionLocality

CollectionDepth

HabitatType

TissueType

TissuePreservation

VoucherName

VoucherHoldingInstitution

ImageFilename

**Cecile Fauvelot, Institut de Recherche le Developpment, Noumea, New Caledonia**

for our particular projects at the moment are kept the size of the individual, the place of collection, the date of collection

**Eric Crandall for Paul Barber, University of California, Los Angeles**

Sample ID

Species

Container

Date

Locality

Trip

n

Notes

Sorted/extracted

GPS N

GPS E

3) *For about how many tissue samples does your lab intend to share genetic data and metadata with the collaborative database? i.e. how many rows?*

**Harilaos Lessios, Smithsonian Tropical Research Institute:**

We currently have ~17,000 specimens, divided into ~33,000 samples. About 7,500 of those specimens are from the Indo-Pacific region.

Our Organism Table is customized to our lab's needs and either records or links to all the associated metadata, and other labs at STRI may have different fields and metadata. However, the Sample Table is shared by all the labs at STRI that are using our database server and this overall sample database (containing 173,000 records) is intended to be shared with a central database containing the cryogenic samples from the National Museum of Natural History and the National Zoo. The Sample Table does not contain the complete metadata - only the unique identifiers, Taxonomic info, Tissue and storage information, summarized collection locality, and sample ownership information (see attached file). This simplified table is designed to be exported and shared.

Metadata can be freely shared, so that everyone knows what samples are available. Information on what specific genes or microsatellites have been obtained is also available. Genetic data are submitted to GenBank at the time of publication. They can also be shared upon request by a particular investigator, stating the use to which he/she intents to make of the unpublished data.

**Rob Toonen, Hawaiian Institute of Marine Biology**

We have genetic data for about 60 species that each have roughly 10 sites and 20-50 individuals per site, so probably on the order of about 18,000 rows currently. However, that is irrelevant to the future needs because in the past month we generated RAD libraries for 60 species in which we average about 5million sequence reads per individual and expect that will become the norm for many in the very near future. That is the type and volume of data that the database needs to accomodate.

**Cynthia Riginos, University of Queensland**

Metadata - anything published. Probably ~1200 COI and control region sequences.

Tissue samples - to be decided on a case by case basis. I don't mind if others know what we have.

Will share any EST data.

**Luiz Rocha, California Academy of Sciences**

So far we have approximately 1,000 samples cataloged in this dataset, growth rate is about 500 per year. All the information associated with the samples is made freely available through our collection database website (above) between 1 and 3 months after collection.

**Sophie von der Heyden – Stellenbosch University, South Africa**

I imagine it is well over a thousand. We are happy to share anything really.

**Gert Worheide, Ludwig-Maximilians-Universitaet München**

Not sure yet, this needs some time to come up with the numbers. Might be difficult because at present we do not have a good link between database entries (which mainly relates to samples) and associated sequences. We are simply not there yet due to lack of (human) resources to enter all the (old) data.

**Rachel Gotanco, University of the Philippines Marine Science Institute**

~400-600 samples

**Cecile Fauvelot, Institut de Recherche le Developpment, Noumea, New Caledonia**

Our two data sets are big ones are thy were created for paternity analysis. So we have the dascyllus aruanus one with about 6000 individuals from a nearly unique location (extending few kilometers)

The Tridacna maxima is smaller, but same configuration (about 1000 individuals from few reefs close to each other)

Another giant clam project of 1000 individuals from a different location, but same configuration.

**Eric Crandall for Paul Barber, University of California, Los Angeles**

The CTPIRE database contains about 22,000 individual tissue samples (most of which have not yet been sequenced/genotyped). The samples themselves are housed in the country where they were collected. Data use is governed by MOUs signed by UCLA and ODU. The Barber Lab database probably contains 10-15K samples, housed at UCLA.

4) *What format do your tissue identifiers have? How do you ensure that they are locally unique (i.e. unique among all tissues in your collection), and do you make an attempt to make them globally unique (i.e. unique such that they could be found unambiguously online in the equivalent of a google search 100 years from now) or accession them with a museum?*

**Harilaos Lessios, Smithsonian Tropical Research Institute:**

Our tissue samples use a minimum 8-character code, usually composed of a four letter pre-fix which is optionally related to the species id and collection locality (e.g. *Echinometra lucunter* from Puerto Rico - ELPR), and a four number suffix. The Prefix is assigned by the database manager usually during collection intake, and is confirmed to be unique at that time. The database itself forces each bio-repository ID code to be unique but the format described is used only for convenience - as far as the database is concerned a random number sequence would work just as well as long as it is unique in the database. We have not attempted to make these IDs globally unique, nor do we see how this would be possible.

Each tissue sample, DNA extraction, and other sub-samples have a unique biorepository-ID, and a new ID is generated each time a sub-sample is generated from one of those samples. Each biorepository-ID is associated in our database with a parent organism-ID which identifies the original organism sampled - this organism ID is also unique in the database (although normally part of the same series as the ID's used for subsamples) and its database entry contains all associated metadata (collection locality, pictures, physical descriptions, collectors information, publications, etc.).

**Rob Toonen, Hawaiian Institute of Marine Biology**

It depends on the person, and each keeps their own system. Many use mmddyy001 for a unique identifier, but some use Genus species 3 letter codes or some combination of collection site and numbers also.

**Cynthia Riginos, University of Queensland**

Each label is locally unique, not globally. (you cannot write too many numbers on each label and I think it is REALLY important to use the same numbers in the field as the lab)

**Luiz Rocha, California Academy of Sciences**

All of our samples are accessioned at the CAS collection database (<http://collections.calacademy.org/ich/>) and linked to specimens.

**Sophie von der Heyden – Stellenbosch University, South Africa**

Each sample receives a unique ID (determined with the lab and depending on the species) that usually encompasses abbreviations of scientific name, sampling locality and/or date. Others are more simple, i.e. B252 (but then all fishes with a B prefix are one species).

**Gert Worheide, Ludwig-Maximilians-Universitaet München**

A unique running number with a GW prefix.

**Rachel Gotanco, University of the Philippines Marine Science Institute**

Samples are identified by a combination of a unique field collection code and sequential numbering.

No attempt to make tissue identifiers globally unique.

None are currently accessioned with a museum.

**Cecile Fauvelot, Institut de Recherche le Developpment, Noumea, New Caledonia**

Codes are unique: usually starts with the initials of the species: (Daru for exemple, or Tmax), then a code for the location (letters usually), then the number of the individual (number)

**Eric Crandall for Paul Barber, University of California, Los Angeles**

The population samples have a four digit serial number, and a two-digit number following the decimal place to identify the tissue. For CTPIRE samples, there is a two letter prefix for each country.

5) *What types of genetic data do you collect from each sample? What format are these data kept in, and how are they currently associated with sample metadata?*

**Harilaos Lessios, Smithsonian Tropical Research Institute:**

Sanger sequence:

For each sequence we generate a sequence record which is associated with the specific Biorepository-ID from which it was generated (usually a DNA sample). Some of our legacy data (i.e. data collected before the database was created) are associated with the parent organism-ID only. Each sequence record contains the associated organism and biorepository IDs, Taxonomic-ID, the locus being sequenced, the name of the primer used to generate the sequence, run-information from our sequencing center (date, run-id, etc), a field containing the text of the generated sequence, and any associated GenBank Accession numbers and/or publication names. As of this year (2013), each record should also have the chromatogram files attached (the chromatograms for older data can be retrieved from archived data on DVDs, but are not linked to the DB).

Genotype data:

These are usually Allozymes (entirely legacy data) or Microsatellite data (actively accumulating). A record is generated for the data from each sample/locus and is linked to the Biorepository-ID. The record displays the unique identifiers and taxonomic information, and has fields for recording the data type (e.g.. microsatellite), run and/or gel information, locus, alleles, and additional information. As of 2013 the associated chromatogram file from the run is attached to the record as well.

Next generation sequence:

We don't currently have the capability of storing or summarizing this information. Until that can be remedied, when samples are used for this type of data a notation is made in the Organism Record under "General information" and in the specific sample record under "Notes."

Other data:

Each organism and sample record has an associated sub-table called metrics that can record customized data about that sample or specimen (e.g.. specimen wet weight. Optical Density or 280/260 ratio of a DNA sample, etc.) These fields are customized to each specimen and not indexed. This is a new feature and has been rarely used to date.

**Rob Toonen, Hawaiian Institute of Marine Biology**

We tried to use the Geneious BioCode plugin, but it did not work with the version of Geneious we have on our server. Students in the lab are highly resistent to a standardized system unless it is intuitive and simple – the more complicated and time consuming it is to database the samples, the less likely it will be completed by most students.

**Cynthia Riginos, University of Queensland**

Sequences mostly. Starting on some msats but no system formalized there.

Raw trace files contain ID number and stored on server along with editted sequence files, nexus files, etc.

**Luiz Rocha, California Academy of Sciences**

Type of genetic data varies from simple COI sequences to RAD-seq and UCE, depending on the needs of individual projects. These data are deposited in Genbank and tied to their samples using the CAS catalog numbers.

**Sophie von der Heyden – Stellenbosch University, South Africa**

Our data base is much more simple than the one you attached, but it has given me some ideas. Will get back to you on this….

**Gert Worheide, Ludwig-Maximilians-Universitaet München**

The intention was to use the Geneious BioCode plugin. However, it is non-intuitive and complex, and there is resistance in the lab to use it, except for 96-well barcoding plates of the Sponge Barcoding Project.

**Rachel Gotanco, University of the Philippines Marine Science Institute**

Sequence data, genotype data

Sequence data in text format.

Genotype data in Excel format.

Genetic data currently not (yet) entered in sample database

**Eric Crandall for Paul Barber, University of California, Los Angeles**

Sequence and microsat data are kept in nexus or arlequin format by the student/postdoc/technician that collected them. The taxon names in these files are the four+two digit identifier described above. Each person working on a particular species has the responsibility to keep these files on their computer, together with associated chromatograms and other notes in a pre-specified format. This responsibility is not always maintained by the numerous students that have worked with the samples.

**Cecile Fauvelot, Institut de Recherche le Developpment, Noumea, New Caledonia**

Mainly microsatellites, but for some individuals, we also plan do to some mtDNA sequencing (CR or CO1)

no association with the sample database, genetic data are kept separately in excel forms for microsatellites and individuals and alignment fasta files for sequences.

6) *What queries would you like to make of a central shared database?*

**Harilaos Lessios, Smithsonian Tropical Research Institute:**

Taxonomic identification, collection locality, collection date, storage medium/sample condition, owner of sample, associated sequence data or accessions and the data repository used if the data are in a public repository such as GenBank.

**Rob Toonen, Hawaiian Institute of Marine Biology**

Most often it would be species, locality & DNA marker right now, but in the future I see habitat, depth range and other environmental parameters in relation to genomic, transcriptomic or metabolomic data becoming far more important.

**Cynthia Riginos, University of Queensland**

Would want to query by location (bounded area) and taxonomic attributes.

**Luiz Rocha, California Academy of Sciences**

At least species and location.

**Sophie von der Heyden – Stellenbosch University, South Africa**

Are vouchers/photos available?

**Gert Worheide, Ludwig-Maximilians-Universitaet München**

Typically species, locus, locality, habitat, depth range.

**Rachel Gotanco, University of the Philippines Marine Science Institute**

Collection areas for species

Available genetic data by data type and locus

**Cecile Fauvelot, Institut de Recherche le Developpment, Noumea, New Caledonia**

Have not thought about it yet

**Eric Crandall**

Query by geographic location, taxon, field collection info (who?, how?), laboratory info (who?, how?), type of genetic data (locus, method etc.)

7) *If possible, could you attach a printout from a small exemplar portion of your sample database to help us gain a better understanding of how it is structured?*

**Rob Toonen, Hawaiian Institute of Marine Biology –** see attached

**Cynthia Riginos, University of Queensland –** see attached

**Luiz Rocha, California Academy of Sciences** database fully available here: <http://collections.calacademy.org/ich/>

**Sophie von der Heyden – Stellenbosch University, South Africa**

**Gert Worheide, Ludwig-Maximilians-Universitaet München –** see attached

**Rachel Gotanco, University of the Philippines Marine Science Institute**

**Cecile Fauvelot, Institut de Recherche le Developpment, Noumea, New Caledonia**

**Paul Barber –** see attached