In 2014, Monique put together a series of slides that defined what the goal of the ITS project was to be in 2015. The basic table is shown here:

Scientific Analysis	Long-term Goals	Near-term science objectives
Data Synthesis	Identify key parameters and condense the time series information	(1) Separately synthesize surface, midwater and benthic time series into a manageable number of representative data sets
Synchronous variability	Identify cause/effect relationships between the time series	(2) Compare the time series from (1) with each other and with climate indices to identify similar variability at the seasonal and interannual scale
Prediction	Predict key parameters in the future, a few years ahead and under climate change	(3) Identify lagged-relationships based on (2) and identify species for which those are robust enough to enable prediction
State of the ecosystem	Define and share proxies related to ecosystem health	(4) Investigate calculations of biodiversity indices
Outside datasets	Enlarge the above studies to the California Current	-none for 2015-
Carbon budget	Quantify carbon fluxes from surface to bottom and harvest	(5) Start combining midwater animal counts with biomass and respiration numbers

She states that one objective is to generate comparable “ecosystem indicies” for the 3 time series that capture key ecological patterns. Some examples are:

1. count/biomass of key species
2. biodiversity indices
3. indices of common variability (PCA)
4. Productivity (surface)
5. Carbon flux
6. vertical distribution (midwater)
7. etc.

She identifies several steps in the processing of data.

1. Data harmonization (monthly) (I would tend to call this pre-processing or cleaning of data)
    1. Temporal coverage
    2. Treatment of missing data
    3. Vertical Averages (Midwater)
2. Data Preprocessing (generating standardized time series)
    1. log transform
    2. remove mean/seasonal cycle/ trend
    3. smoothing
    4. species relative weights
3. Data Synthesis/Analysis (generating a manageable number of ecosystem indices)
    1. Principal component analysis
    2. Biodiversity indices
    3. Clustering

Some of the decisions and assumptions that she is making are:

1. Binning data monthly
2. Vertically integrate animal counts (average) for midwater but only from 200-1000 meters
3. When generating anomalies, she does a log transform, then removes the seasonal cycle, then removes the trend.

Some questions she presents are:

1. Can we interpolate over missing data?
2. there are issues with removing seasonal and trend, are we comfortable with those?
3. Should we do smoothing? FFT? Window Smoothing? 3-points? other?
4. How do we weigh the time series? Are certain organisms more important than others? Do we weight by trophic level? Classification level?
5. For analysis/synthesis, what methods should we use: PCA? Bidiversity indices? Key species? Clustering?

My main task is the synthesis task where I take the parameters that are defined by the group (mostly Monique) and synthesize them into a single representative data set with a central data store. Then I can work on the visualization and analysis work and integrating it with climate indices and with external data sets. At some point, it would be nice to allow people to annotate things in the data set to help identify relationships. Also open data to machine learning. Also integrate with calculations of biodiversity indices. Eventually the data will be available to help model carbon flux and budget.

As I look over her slides from multiple presentations, I can extract some of the parameters that she is interested in using:

1. Midwater
    1. Choosing a specific level in the taxa
    2. Getting counts of organisms at the different taxa level
    3. For each chosen level in taxa, she uses counts per meter^3 of water combined by month
    4. Then, she takes integrated counts from 3, log-transforms them and removes seasonal cycle
        1. She poses questions about what are valid transformations of data. This might be nice to be able to chose from the user interface.
    5. She calculates 'anom' which looks to be deviations from some average calculated by month. This is combined over all depths (as are the previous analyses).
    6. Then calculates depth distribution for all specified organisms. This is the percent of the total animal counts grouped by depth. For example, maybe 80% of the Nanomia 200m.
2. Surface
    1. Temperature at M1 (sea surface)
    2. Salinity at M1 (surface I think)
    3. Calculates anom with both SST and Salinity, and it looks like they are grouped by month again.
    4. Primary production which is in mg m^-3
    5. Chlorophyll mg m^-3
    6. Diatoms C m^-3 (I think that is counts per)
    7. Dinoflagellates C m^-3
    8. Picoplankton C m^-3
    9. She takes phytoplankton groups (acid, chill, choano, etc.) and calculates Carbon per liter, grouped by month/year
    10. Calculates phytoplankton biomass which is grouped by diatoms, dinos, etc. and is in units of mg C/m^3 and is grouped by month/year
3. Benthic
    1. Density of 10 key species (individuals m-2)
4. Physical data sets
    1. She wants to add ROV temp, O2 (should we do this for benthic?)

Monique then sent out a Excel spreadsheet that contained the data sets she is using for her work and I started working from that. When looking at the above in summary, the data that we need is:

1. Surface
    1. For each column in the TAXA table in BOG, we want mgC/m3, grouped by month
2. Midwater
    1. For each species group counts by month
        1. One grouping for each depth
        2. One grouping across all depths
3. Benthic
    1. Counts of each species/m2, grouped by month

Loading
After looking at the documents from previous meetings, I wrote some loading software to pull the information from the sources above into a MongoDB instance. Here are the extraction steps

1. Upper water column
    1. Query the BOG database using (note the list of things to query for after depth is specified in a configuration file):
    'select cruise, date_time, seq, ctrb_id, dec_lat, dec_long, depth, allp, allhet, syn, rfp, prym, aflag, adino, cryp, prasy, phaeo, pen, cen, hdino, hflag, choano, hcryp, hcil, acil from taxa where (ctrb_id in ('M1','Mooring1','H3','67-50') and depth=0) order by date_time'
    2. At the time of this writing, that produced 498 documents that looked something like:

```
{
    "_id" : ObjectId("569e86928fd7841393fa718c"),
    "date" : ISODate("2006-12-01T08:00:00.000+0000"),
    "date-resolution" : "month",
    "depth" : NumberInt(4000),
    "name" : "Abyssocucmis abyssorum",
    "value" : 14.0
}
```

2. Midwater/VARS database:
    1. This is a multi-step process. The first step is to find the dives of interest in the midwater database. The query run on the midwater database is:
    'SELECT id, dive, rov, day_night, year, start_datetime_gmt,
    end_datetime_gmt, depth_m, volume_m3 FROM transect_data INNER JOIN (SELECT rov as 'rov2',
    dive as 'dive2', count(*) as 'numRows' FROM transect_data WHERE day_night = 'D'
    AND year >= 1997 GROUP BY rov, dive HAVING count(*) > 3) table2
    ON transect_data.rov=table2.rov2 AND transect_data.dive = table2.dive2
    WHERE depth_m in (50,100,200,300,400,500,600,700,800,900,1000) ORDER BY depth_m, start_datetime_gmt'

2. The results of this query are written to a file and have records of the form:
   Transect Target Depth (m), Transect Start Date (GMT), Transect End Date (GMT), Year, ROV, Dive, Day_Night, Transect Volume (m^3)
   50.0,1997-08-08 20:01:00.0,1997-08-08 20:11:00.0,1997,Ventana,1292,D,1118.7001
   50.0,1997-09-30 20:46:35.0,1997-09-30 20:56:35.0,1997,Ventana,1321,D,1118.7001
3. This file is then passed to the next step in the process which is to extract the annotations from the VARS database. First a prepared statement is built that looks like (note the list of species queried for is specified in a configuration file):
   '**SELECT ConceptName, Latitude, Longitude, DEPTH, RecordedDate, ObservationID_FK, LinkName, LinkValue, VideoArchiveName FROM Annotations WHERE RovName = ? AND RecordedDate >= ? AND RecordedDate <= ? AND (ConceptName = 'peobius' OR ConceptName = '**
   nanomia' OR ConceptName = 'aegina' OR ConceptName = 'chuniphyes' OR ConceptName = 'solmaris' OR ConceptName = 'euphausia' OR ConceptName = 'caecosagitta' OR ConceptName = 'pseudosagitta') ORDER BY RecordedDate ASC
4. The midwater file is then looped over and the ROV name and start and end times are filled in using the next line in the file and the query is then run.
5. The results are then processed one by one with the following steps:
   1. If the LinkName is 'identity-reference' it is recorded in a file that keeps track of entries that may be duplicates
   2. if the LinkName is 'population-quatity', it is written to a separate file that is keeping track of the counts of specific species
6. The processing then takes both files and creates a third file with all entries, but with duplicates removed
7. The cleaned file is then opened, read in line-by-line and each line creates a document in the MongoDB that looks something like:

```
{
    "_id" : ObjectId("569e868d8fd7841393f86131"),
    "date" : ISODate("1999-12-27T04:11:50.000+0000"),
    "date-resolution" : "second",
    "loc" : {
        "type" : "Point",
        "coordinates" : [
            -122.0375747680664,
            36.68381118774414,
            -50.0
        ]
    },
    "depth" : 50.0,
    "name" : "aegina",
    "volume" : 1966.800048828125,
    "count" : NumberInt(1)
}
```

3. Benthic data
   1. The software opens the spreadsheet (Lynda's from the video lab)
   2. It selects the "summary abundance tab"
   3. Then it loops over each row and column and inserts a document in the MongoDB summarizing the counts for the species. Each document looks something like:

```
{
    "_id" : ObjectId("569e86928fd7841393fa718c"),
    "date" : ISODate("2006-12-01T08:00:00.000+0000"),
    "date-resolution" : "month",
    "depth" : NumberInt(4000),
    "name" : "Abyssocucmis abyssorum",
     "count" : 14.0
}
```

4. I then tell MongoDB to perform a series of aggregations to create processed data sets from this loaded data.