1. **Data contextualization for decision analysis and fitness-for-use flagging**

Globally, the magnitude of archived data is doubling every couple of years. Federal agencies must process mounting quantities of data to address highly complex problems that examine potential shifts in critical species habitats, key spatial, temporal and taxonomic resolutions required for long-term planning needs, and forecasting information useful to protect marine ecosystems and associated biodiversity. Processing increasing quantities (stores) of multi-use multi-source data will require improved methods to optimize the retrieval of data and to identify data with greatest relevance for a defined study.

Work plan tasks will enhance data contextualization and fitness-for-use flagging used in the Data Assessment Module (DAM) currently under development for OBIS-USA. Supplementary information provided below describes the broader conceptual framework and current DAM prototype structure for OBIS-USA. Work will focus on developing fitness-of-use flags for evaluating relations between biodiversity data and environmental condition data. A secondary effort will focus on scale disparities between data types and scales required for management purposes. These work components are expected to contribute to three focus areas (Prototype OBIS-USA Data Assessment Module (DAM), “classification flags”, “Support scale contextualization”, and “Environmental contextualization”).

**Task 2 - Conceptual framework:**

Publically accessible databases, spreadsheets, and images linked to rich metadata are often used to show scientific reproducibility and to allow generic re-use of data for open-ended applications. This traditional ‘blank-canvas’ approach valuable for base-level documentation places the raw data at the forefront, as the supreme entity, but often overlooks critical information needed for data assessments including metrics of data quality, suitability for the intended study, customized interests of the user, and spatial or temporal resolutions needed to make management decisions or to conduct scientific studies. Agencies typically have specific objectives for limited environments such as examining changes in biodiversity of coral reef species sensitive to climate warming. Without proper context and guidance, data stores are underutilized, and of little practical importance beyond basic archiving and access.

Drawing the most relevant and powerful portions of a data store for an intended study is facilitated through data contextualization and fitness of use measures. Contextualization of data is enabled by searches that are tailored to demographic data, historical data, and relational data to improve relevancy of the data query. Data contextualization may involve location targeted content, optimized content and functionality, and adaptive designs. Fitness-of-use is a measure of data suitability dependent on the intended application. Data may be fit-for-use for a particular analysis at a particular scale of resolution but inappropriate for other applications. It is therefore imperative to provide an array of measures of fitness-of-use that may be used independently or in concert to evaluate suitability of the data for the intended use. Visualization techniques coupled with basic statistics and indicators of data content provide effective tools for data assessments. Visualization and basic statistics have too often become end products of scientific analysis, rather than exploration tools that scientists or managers can use throughout the research life cycle, including data selection.

Organisms are influenced by multiple environmental factors working at different spatial and temporal scales. Relations between biodiversity variables and underlying environmental factors are increasingly assessed in an explicit way. Fitness-of-use measures are needed to evaluate properties at different scales both in space and time and to determine the representativeness of the transformations. Moreover, identifying commonality between spatial or temporal scales for resource management or scientific application and scales supported by the data, or measures derived therewith, is also a critical component in data evaluation. Managers and scientists desire near effortless access to the most appropriate types and forms of data at representative scales of interest. Developing more intelligent and customized methods for data selection, evaluation, and transformation will work toward this goal.

**Task 2 – Prototype OBIS-USA Data Assessment Module (DAM):**

Note: highlighted topics pertain to the current work plan

**Fitness-of-use flags**

1. ***Data centric flags*** *(independent)*
2. *Data errors*: individual data errors identified and associated to root causes wherever possible to show data transparency, used in part to evaluate data quality.
3. *Data cleansing:* indication of data correction, procedure(s) used for the correction, and date the correction was performed.
4. *Data confidence:* uncertainty in observer’s understanding of taxonomic identity, characterization, or other attribute pertaining to the data type.
5. *Data notation:* a generic or coarsely categorized indication of important information related to the observation that could influence data analyses.
6. ***QA/QC flags*** *(mainly independent, associated by data set)*
7. *Grouped indicators*: measures of data precision, accuracy, transparency, completeness, consistency, uncertainty, and believability may be used to evaluate data quality. Data quality is subjective and evaluated by the user for a defined application. Group measures are linked back to individual data for maintaining provenance during atomization for cases where pieces of multiple data sets are combined for a specific application.
8. *Performance measures:* performance checks on location correctness, verification by qualified individuals (taxonomists) or agencies, period of verification, value correctness, citation indexes to measure community acceptance, etc.
9. *Outlier detection:* geographic, statistical, and environmental outliers are identified using deterministic or conditional probability methods. Outliers do not necessarily reflect data errors but can leverage statistical procedures sensitive to outlier presence.
10. *User feedback/rating:* authors and subsequent data users publically rate data quality, discuss applicability, and describe details of the data to inform and therefore benefit the community. In this manner, data packages would be continually reviewed and potentially updated similar to providing product ratings for a commercial enterprise, which enhances user confidence.
11. ***Classification flags*** *(associated to attribute, region, or classification)*
12. *Environmental conditions:* classification indexes for geography, climate, habitat, ecosystem disturbance (e.g. storm frequency), and physiological stress (e.g. coral bleaching). Environmental indexes could be related to NEPA activities.
13. *Management designations:* classification indexes for government and international regions, and environmentally protected and/or affected areas.
14. *Biodiversity:* classification indexes for supplementary information on organism characteristics (e.g. predator versus prey, organism longevity, species size, habitat, inherent traits). Identification of keystone species sensitive to marine ecosystem health, societal impacts, or climate change may also provide useful constraints for data evaluation. OBIS-USA currently queries organisms by taxonomy.
15. *Human activities:* proximity indexes to ocean transportation routes, oil exploration and development, alternative energy farms (wind, wave), terrestrial population centers.
16. *Ecosystem characteristics:* classification indexes for ocean ecosystem name, functional type, and associated key properties. OBIS-USA currently uses a couple of major ecosystem classification schemes.
17. *Ocean attributes:* classification indexes for bathymetry, sea level, sea-surface temperature, ocean circulation, nutrient transport, and biochemical conditions (e.g. salinity).

**Data Contextualization**

1. ***Environmental contextualization*** *(environment specific, taxa relational)*
2. *Ocean biochemistry:* physiological relations to biochemical properties or responses to changes in biochemical properties (e.g. changes in ocean nutrients or salinity) are related by convincing findings in the scientific literature, observed relations in the data source, or post audit analyses to an open-source geospatial coverage of credible origin. Data queries are refined by biochemical condition and/or relation to organism attributes (e.g. diversity indices, changes in population densities), allowing improved contextualization of data. Wherever possible discrete data locations are overlaid on maps of biochemical properties to further enhance data selection and visualization.
3. *Ocean thermo-physics:* physiological relations to ocean patterns or responses to changes in ocean patterns (e.g. ocean circulation, sea-levels, sea surface temperature, sea ice coverage, and natural disturbances) are related by convincing findings in the scientific literature, observed relations in the data source, or post audit analyses to an open-source geospatial coverage of credible origin. Data queries are refined by ocean patterns and/or relation to organism attributes (e.g. diversity indices, changes in population densities), allowing improved contextualization of data. Wherever possible discrete data locations are overlaid on maps of ocean attributes to further enhance data selection and visualization.
4. *Climate:* physiological relations to climate properties or responses to changes in climate properties (e.g. climate warming, climate variability) are related by convincing findings in the scientific literature, observed relations in the data source, or post audit analyses to an open-source geospatial coverage of credible origin. Data queries are refined by climate properties and/or relation to organism attributes (e.g. diversity indices, changes in population densities), allowing improved contextualization of data. Wherever possible discrete data locations are overlaid on maps of climate properties to further enhance data selection and visualization.
5. *Human factors:* physiological relations to human factors or responses to changes in human factors (e.g. changes in coastal populations, use of natural resources, ocean transportation routes, alternative wind or wave energy) are related by convincing findings in the scientific literature, observed relations in the data source, or post audit analyses to an open-source geospatial coverage of credible origin. Data queries are refined by human factors and/or relation to organism attributes (e.g. diversity indices, changes in population densities), allowing improved contextualization of data. Wherever possible discrete data locations are overlaid on maps of human factors to further enhance data selection and visualization.

*C.* ***Statistical contextualization*** *(application specific)*

1. *Sample populations:* sample representativeness for each data query is examined using population statistics prioritized to extensive properties. Biodiversity attributes are examined using parametric and nonparametric tests to determine statistical differences between samples by major classifications (e.g. management districts, marine ecosystems) to guide data selection.
2. *Spatiotemporal patterns*: spatial analyses of statistical measures (e.g. shift in species abundance, shift in habitats, biodiversity indices) are used to evaluate the potential of data derived from multiple sources resolved over major regions of interest (e.g. management districts, marine ecosystems) to guide data selection.
3. *Trends:* parametric and nonparametric trend analyses are used to visualize temporal changes in attribute values (e.g. temporal change in species abundance) and direction (up, down, no change) at defined levels of confidence. Trends are displayed visually perhaps as arrows along with environmental or human attributes to guide data selection.
4. *Associations*: parametric and nonparametric correlation analyses are used to reveal basic quantitative relations between biodiversity attributes and environmental or anthropogenic attributes to guide data selection.

*D.* ***Support scale contextualization*** *(query specific)*

1. *Data support scales: geostatistical and spatial probability measures are used to identify spatial scales (resolutions) reasonably supported by the distribution of available data.*
2. *Management support scales: managed areas are often governed by agency priorities. Obvious “first-cut” delineations of management areas could use governed regions or major ecosystem classifications currently implemented in OBIS-USA, possibly further partitioned using choropleths. Tradeoffs between support scale and prediction uncertainty should be evaluated.*
3. *Adjustment of scale: Several circumstances arise where support scales must be adjusted or interrelated. These cases may arise from comparing multiple forms of data represented at different scales or transforming results to a different scale of representation more meaningful to scientists and decision makers. Geostatistical modeling or other types of spatial modeling, such as self-affine correction, linearized co-regionalization, and nested hierarchical approaches offer imperfect but useful pathways to accomplish these goals.*