1. Vision

The goal of EarthCube as developed through two Webinars and to be expanded upon at the forthcoming Charrette is to:

*transform the conduct of research by supporting the development of community-guided cyberinfrastructure to integrate data and information for knowledge management across the Geosciences.*

The participants in the Ocean Observatories Initiative Cyberinfrastructure Implementing Organization have been working to integrate sensors and platforms at sea into a system for providing open, near-real-time (latencies as small as seconds) to the Earth sciences community and, at the same time, provide a means for controlling the system components to maximize the performance of the system. The OOI cyberinfrastructure tools for development include:

1. End-to-end data preservation and access,
2. End-to-end, human-to-machine and machine-to-machine control of how data are collected and analyzed,
3. Direct, closed loop interaction of models with the data acquisition process,
4. Virtual collaborations created on demand to drive data-model coupling and share ocean observatory resources (e.g. instruments, networks, computing, storage and workflows),
5. End-to-end preservation of the ocean observatory process and its outcomes, and
6. Automation of the planning and prosecution of observational programs.

Additionally, the cyberinfrastructure systems must provide the required background messaging, governance and service frameworks (e.g. identity management including trust relationships within the network) that facilitate interaction in a shared environment, similar to the operating system on a computer. Such a system must provide a suite of tools capable of serving both basic and applied science simultaneously. Some of these tools are directly applicable to EarthCube and we shall seek to emphasize these in this white paper.

The potential of an interactive social network for the Earth sciences is in its infancy and the community is beginning to conduct pilot experiments that take advantage of the new data and knowledge architectures. There is a growing sense that linking observatories, computation, modeling, storage and network infrastructure into a coherent system-of-systems. The resulting system, in a social context, should enable real-time visualization and access to model outputs to allow for adaptive sampling science.

2. EarthCube Design Principles

We advocate that EarthCube invest in building a system that will allow the geosciences community to easily contribute, leverage, and operate on digital products and thus create a self-sustaining, growing network of accessible body of knowledge. We see the principles below as encouraging such sustainable knowledge sharing and collaboration:
2.1. Digital products as first-class citizens.

Research results can no longer be shared exclusively via papers and discussion at conferences, but in the most convenient, up-to-date form, often as digital products representing and developed from data sets, digitally represented algorithms (e.g., as appliances), images, etc. The ability to instantly rerun a digitally represented algorithm with a new dataset rather than reconstruct it from a paper has potential to reduce the “time to science” metric and thus increase collaboration and productivity. Such digital products can be published and made available to the community via social sharing mechanisms facilitating discussion, contribution, and validation. The support for this form of publication requires providing the underpinnings of ownership management (i.e., licensing and citation), preservation (i.e., annotation, standardization, digital libraries) as well as infrastructure that can manage storage, curation, validation, and access. We advocate that EarthCube infrastructure should combine with general and community-specific methods implementing open access to data, algorithms, and other scientific artifacts that, where possible, could be shared in an easily “consumable” form. There are many exemplars for accomplishing this open licensing including BSD, Apache and Creative Commons.

2.2. Foster opportunities for data integration and cross-referencing.

Extracting information from data often requires cross-referencing with existing, published data sets. While data as digital products will offer a substantial library of potential targets, integration will require the ability to identify data sets of interest, manage appropriate storage and mirrors, and above all, develop a Common Data Model that allows for high fidelity transfer of data (such that the scientific information is preserved) between observation and modeling and modeling to modeling communities, which invariably use different data structures (i.e., topologies) to analyze their data. The core issue is the ability to interpolate a source data structure onto a target data structure (i.e., subsetting and regridding) and preserve the numerical accuracy of the source data [??]. Coupling ease-of-use with high-fidelity data translations will foster data contribution and sharing, reuse, validation, and dissemination. This could be accomplished by providing an integrated approach that will allow a cyber-client to browse or even automatically recommend data sets of interest, and then perform high-fidelity translations enabling faster and more robust connections. Coupling ease-of-use with high-fidelity data translations will foster data contributions and sharing, reuse, validation and dissemination.

2.3. Support for Observatory Science.

Oceanographers have historically collected data in the ocean and the seafloor from ships during cruises of limited duration. This expeditionary research approach has resulted in major advances that span understanding a range of ocean-related phenomena and has also demonstrated a need for sustained sampling spanning temporal and spatial scales that are not effectively carried out within the expeditionary approach. Filling these informational gaps required the oceanographic community to develop an observatory approach, geared towards long time scales (~25 years as is the case for the OOI) (Schofield et al., in press). Fortunately in this mission the ocean community was able to take advantage of technical advances, in particular in sensor technology and wireless networking leading a transformation from experimental science towards observatory science where detailed monitoring of an environment can be achieved based on sensor networks rather than direct observation. The OOI project is leading this transformation where data describing various phenomena are obtained via a wide-ranging network of heterogeneous sensors. The availability of this up-to-the-minute information requires new methods of data processing as the existence of various phenomena such as hurricanes, algal blooms, or a tsunami become known and must be analyzed on-demand; it therefore requires the ability to rapidly and reliably provision compute and storage resources capable of processing the incoming information and storing and disseminating the new model outputs. The design considerations for building an observing network can be generalized to other domains where reactive systems support peak needs that occur when multiple scientists request processing to support ongoing experiment needs, after experiments when multiple parties are trying to process data, or before various collaborative events as scientists are bringing their research to a conclusion so that results can be shared and contributed to the general body of knowledge. As rapid progress in obtaining new experimental data makes results obsolete quickly, a direct relationship often exists between the ability to rapidly provision resources for analysis via an adaptive system and the rate of progress in a research area.

2.4. Support for open science.

Creating infrastructure where more scientists can experiment with digital products extends the amount of talent that can be brought to bear on important scientific problems by providing open access to those products. Such openness brings to
scientific problems fresh perspectives, unbiased by traditional perceptions, and allows different backgrounds, expertise, and skills to play a role in finding new solutions. In addition, increasing the sheer numbers of researchers, and reaching out to interested students and early career scientists creates a growth potential that is likely to significantly accelerate the discovery cycle. However, open science brings with it an increase the size and capabilities of the user community and changes in the types of users that will demand new approaches for enabling the types of data analysis, data movement, and data storage on computing resources that are widely available and adaptable to the resource needs and culture of the broader scientific community. To cater to this new audience it is essential that new methods are developed that make data and computing easy to access to the average scientist. More importantly, it is necessary to develop methods that will not only handle data and computation at far larger scales than currently supported, but that will also gracefully manage growth. The supporting infrastructure should allow contributions to the general body of knowledge in a relatively loose, uncoordinated fashion, resulting in a body of knowledge that can be collectively modified and supported over time.

3. Approach

To realize these principles we propose an approach informed by the ideas and experiences developed in the context of the funded OOI project as a model that can be built upon, adapted and extended, and also ensure collaboration with the Earth science communities via a commonality of interfaces and technologies. We envision a system where data products and other digital contributions can be shared via a publish/subscribe mechanism, which integrates high-fidelity data translation to facilitate data sharing as well as data modeling and other operations yielding a variety of data products. The underpinning of this system are built on top of a flexible provisioning base, integrating resources from a variety of sources and relying on infrastructure clouds to provide on-demand elastic provisioning accommodating both activity growth (e.g. driven by data returned by sensors) and community growth (e.g. allowing independent researchers to interact with the system via resources independently in the cloud).

Building on the discussions and experiences that lead to the design of the OOI we advocate the adoption of the following ideas.

3.1. Management of Data Products

Today, problems such as climate or the run up of a tsunami on a complex coastline are too complicated to express as analytical solutions and are too large to be replicated in a laboratory. Simulation, however, can be used to study the complex questions using large scale computing and necessarily incomplete data to approximate reality (Howe, 2006). While observations in the Earth sciences is growing at a very high rate, it remains impossible in any realistic future to observe at scales sufficient to understand specific currents, waves, or earthquakes. As observing systems grow and gain sophistication, the expenses cannot be dedicated to the study any specific hypothesis. Howe (2006) notes that the creation of a comprehensive data repository will allow future scientific discoveries on a larger scale; he terms this *Downloading the World* (DW).

The aforementioned Common Data Model that allows for high fidelity transfer of data (that is, preserving the scientific information) from observation to modeling to modeling communities creating ensembles of results, which invariably use different data structures (topologies) to analyze their data. The core issue is the ability to project (or interpolate) a source data structure onto a target data structure (subsetting and regridding) and preserve and extend the accuracy of the source data; in other words, use the function space of the source to calculate the interpolation). The middleware needed to support this task is referred to as the gridfield model. The use of the DW model for managing huge, organized filesystems depends upon a language for extracting portions of files across disciplinary stovepipes to expose the integrated data as gridfields.

We believe (e.g. Howe, 2006) that the data products achieved by combining the gridfields with computational models will supplant the research paper as the currency of scientific communication – the data product can deliver a scientific message succinctly and dramatically and that is more understandable by non-specialists and the general public.
OOI can contribute to the Earth science community the following four production service capabilities to operate on any Earth observing data:

1. Federated Spatial-Temporal Indexing for all structured Earth observing data sets,
2. Evaluation of mathematical expressions over any Earth observing data set,
3. High fidelity projection of any data set structure onto any other data structure (by regridding and subsetting), and
4. Standard calculable addressing scheme for data elements within any structured data set across all feature types from simple time series to C-grids to arbitrary heterogeneous mesh topologies.

3.2. Management and Integration of Compute Resources

To manage computation and storage resources we advocate the development of a reactive system that can adapt to the needs of incoming sensor and other data products in terms of resource acquisition, participant notification, and failure. Such reactive architecture is based on information coming from hardware and software sensors that can express both application-
actions can then be realized by provisioning actions applied to local systems, national infrastructure, or community and commercial infrastructure clouds. We have successfully used this pattern to provide overall resource management working the OOI project for data distribution networks and will build on and extend facilities developed by the OOI project (OOI CI and generalize them to provide additional functionality necessary to adapt them to the needs of other scientific projects.

Practical application of this pattern to scientific application on a large scale will present additional challenges including storage and network management in the cloud to keep up with computational demands, interoperability in various cloud environments, evaluation of choice among different cloud providers, provisioning of instances to support tightly coupled applications and other topologies, policy and fault-tolerance support, and assurance of security, privacy, and trust via a variety of mechanisms ranging from private networks and certificate management to encryption.

4. The OOI Cyberinfrastructure

The OOI is establishing a high-speed network linking several sites in the US to the fiber-optical seafloor network (the OOI Regional Scale Nodes [RSN] and the Coastal-Global Scale Networks [CGSN]). The specific spatial network is shown in Figure 1. The RSN connection to Portland features two separate 10Gbps for redundancy and the triangular configuration of the network provides connection redundancy to all sites in the event the network is inadvertently broken at a point. The fiber optical connections (to be selected by an open bidding process; e.g. National Lambda Rail (NLR) or Intenet2) feature repeaters approximately every 200km allowing interested users and their campuses to establish a direct connection to the OOI network.
4.1. Design Considerations for Ocean Observatories

Ocean observing networks are designed to address a specific need, which is used to define the required sampling resolution in space and time. Defining the appropriate scales can be a difficult problem as many large-scale (1000’s of kilometers), long period (annual-to-interannual) processes are determined by small-scale, short-period variations in atmospheric forcing, and small-scale, relatively short-lived, oceanic processes (Large et al., 1991; Milliff et al., 1996, 1999, Milliff and Morzel 2001). As highlighted by Munk (2002), 95% of the oceanic kinetic energy is associated with mesoscale currents having time and space scales less than about 100 days and 100 km. Other forcing factors can operate over inertial or diurnal time scales (Milliff et al., 2001); therefore, a comprehensive understanding of the oceans will require nested sampling capable of resolving the feedbacks between processes operating over different scales. This generally requires a multiplatform strategy as each system samples a specific time and space domain. Once the sampling requirements have been defined, it is possible to choose 1) the appropriate platforms, 2) the required measurements, 3) the data latency needs for a particular observatory, and 4) the funds available for construction. The costs generally increase with the flexibility of the system. Increased power on a platform allows for greater flexibility in carrying a wider range of sensors.

As ocean infrastructure is expensive, most large infrastructure networks must often be able to address a range of basic research and applied science needs to justify the investment. Historically, basic and applied research efforts are often treated as separate enterprises; however, major issues confronting the ocean science communities reveal numerous commonalities that reflect the chronic under-sampling of the oceans. Both applied and basic science require information on the physical hydrography, circulation, biological and chemical properties; however, it is often the real time availability of data that...
defines its utility for applied science where data are used to meet real-time needs such as weather forecasting, search and rescue, and national security. When available, however, real time information has a great deal of utility as scientists use the information to optimize adaptive sampling techniques.

4.2. Information Systems for Ocean Observatories

While the range of technologies available to oceanographers has been increasing over the last several decades, it is the availability of global communications and information technology that allow these technologies to transform ocean (and Earth) sciences. Oceanographers have conducted experiments as either individuals or small groups within a single science focus at any given time; however, the broad scientific and civil demands for multidisciplinary and interdisciplinary research coupled with exponential growth in information technology are transforming oceanography. This history of working in small groups has resulted in the traditional data-centric cyberinfrastructure strategy, where typically a central data management system ingests data and serves them to users on a query basis. This approach is not sufficient to deal with the range of challenges that face ocean sciences.

Given this potential, the community is now dedicated to building the cyberinfrastructure that will be central to any globally integrated ocean observing system. A modern cyberinfrastructure backbone will allow globally distributed scientists to operate as a community by aggregating data from individually deployed instruments for any experimental effort. When realized, this will allow anyone with access to the internet to use the global array of sensors to study any ocean process of interest.

The OOI conducted a field experiment in 2009, which allowed a distributed community of scientists to assess how well the software could aggregate data from ships, autonomous underwater vehicles (AUVs), shore-based radars, and satellites and to make it available to ocean forecast models. Scientists used the model forecasts to guide future (next 24 hours) glider missions which then were used to optimize data collection for model data assimilation, which demonstrated the feasibility of two-way interactivity between the sensor web and predictive models. The sensor web included the re-tasking of a satellite. The early software allowed the distributed community to adaptively modify the in situ observation network throughout the experiment (Schofield et al. 2010). The net result was a science driven machine-to-machine interactive loop (Figure 3). These machine networks will increasingly become standard tools for the ocean science community in the future.

As observatories comprise a series of individual components that are linked to form a coherent sampling system, an often underemphasized, yet critical need is the ability to register to all components to a common time stamp. The time stamp functionality in the sensor network is necessary to compare the output of one sensor to another. This is not trivial when sensors are dispersed geographically and the data from sensors need to be integrated with external datasets. The accuracy required is a function of the process being studied and the length of the time series to be collected. For example, seismic studies require time accuracy on the order of milliseconds, acoustic tomography of one microsecond and studies of phytoplankton growth rates require data on the time scale of hours. Additionally, avoiding drift for temporal time series will increasingly become critical as sustained time series become the norm for oceanography. Fortunately, the ability to register time accurately is increasingly improving. A dramatic example is the evolution of small low power atomic clocks. Atomic clocks offer the frequency stability of one part in ten billion, which is equivalent to gaining or losing one second every three hundred years. These advances enable atomic clocks on a chip to be operated on batteries and can be integrated throughout the individual components of the ocean observatories. GPS, of course, has provided highly accurate time for years as long as the antenna extends into the atmosphere.

The potential of cyberinfrastructure tools, such as described above, is dependent on the real-time availability of data; fortunately global communications have improved dramatically over the last few decades. In the early 1990’s the primary mode of communication from ship to shore was via satellite voice calls. This improved over the next decade as ship-based science was provided with limited email communication. Communications have continued to improve and now provide sufficient bandwidth to allow for video-transmission at 100’s of kilobits/s (http://hiseasnet.ucsd.edu/). These improvements are changing the type of science that ships conduct as real-time data allow scientists to adaptively sample the ocean. Additionally the launch of low Earth orbit satellite communication systems have allowed for global communication and enabled rapidly evolving capabilities for communications to autonomous platforms. The communications have improved from one-way communications with data transmission limited to about 16,000 bits/day to global two-way communications at a rate of 2400 bits/second.
5. References


