Workshop Report

Cyberinfrastructure and the NSF Dimensions of Biodiversity Program Solicitation

Planning for Success

October 13-15, 2010
Madison, WI

Editors: Corinna Gries, Allen Rodrigo, Matthew B. Jones, David Vieglais

Additional material may be found at: http://lter.limnology.wisc.edu/cidimensions/

Executive Summary

The NSF 'Dimensions of Biodiversity' program recognizes that emerging technologies in computing and cyberinfrastructure are revolutionizing our ability to investigate the broad scale patterns and processes underlying biodiversity. This report documents the outcomes of an NSF sponsored workshop (DBI-1047800) held at Madison, WI during October 13-15, 2010 that was charged with identifying aspects of cyberinfrastructure necessary for supporting successful research in the Dimensions of Biodiversity program. Workshop participants represented a broad spectrum of disciplines, ranging from cyberinfrastructure, informatics, and computer science to biodiversity, biology, and environmental science, and provided excellent insights into the major cyberinfrastructure needs of biodiversity science.

For CI developments in general the importance was emphasized to strike a balance between satisfying specific needs of DoB projects and leveraging of existing developments in the larger biodiversity informatics community. The former will assure researcher buy-in but promote business as usual approaches, while the latter may lead to more cumbersome requirements for documentation, annotation and best practices but will improve interoperability and sustainability. This balance should provide optimal return of investment by promoting an economic ecosystem for collaboration and dissemination of data, tools, and computations while reusing existing technologies, leveraging developments in the larger community and determining the needs that are specific to DoB. A focused, mission oriented and coordinated effort in collaboration between DoB researchers and CI experts is needed to develop a specific road map for identifying, selecting and/or developing resources and analytical tools necessary for biodiversity research.

Specific issues identified by this workshop are

- A governance / management / operational structure that ensures coordination of CI developments and agreements on standards (in particular a taxon concept standard) and best practices. A coordinated effort with input from DoB projects can improve design and coding quality leading to broader applicability of fewer and better tools. Standards and best practices need to be agreed upon in a community process that ensures participant buy-in.
- Leveraging existing initiatives (DataNet, iPlant, etc).
- Sustainable data and tools repository that provides state of the art curation approaches making data and tools discoverable, easy to access, integrate and use.
- Repeatability of analyses will require publication and extensive documentation of data and analytical procedures, preferably in a standardized approach that allows repeating of analyses by others.
- Access to advanced computational cycles; although the computing power is available it is currently not easily accessible for most scientists and standardized procedures need to be developed which should make the underlying hardware transparent to the user.
- Workforce training is needed on several levels (undergraduate, graduate and post-graduate) for scientists to know what tools are available, use them effectively and communicate requirements to software developers.
- CI sustainability needs to be achieved in two areas. First, in a community supported process best of breed CI developments need to be identified for further maintenance and second, negotiations with universities, libraries and professional societies need to insure financial support.

The most challenging of these will, in fact, be to determine the governance / management / operational structure that can deliver CI resources in a way that best meets the needs of a scientific community for which past efforts at coordination have been frequently unsuccessful. As this issue influences several others mentioned, it is also the most important one. Several examples of more successful organizations are for example the US-LTER program, UNIDATA, NCBI, ESIP, EOL.

Included with the report body are a number of side-bars highlighting exemplars of existing infrastructure and projects highly relevant to the DoB cyberinfrastructure, a set of milestones for years 1, 3, 5, and 7 that identify strategic, procedural and operational outcomes which are likely to ensure the success of the CI effort for DoB, and appendices of relevant tools and technologies, a list of workshop participants and a glossary of terms. Additional material (including the proposal, workshop agenda, answers prepared by workshop participants, presentations and the complete text of the here abbreviated side bars, may be accessed at http://lter.limnology.wisc.edu/cidimensions/.

Introduction

The NSF 'Dimensions of Biodiversity' (DoB hereafter) program recognizes that emerging technologies in computing and cyberinfrastructure are revolutionizing our ability to investigate the broad scale patterns and processes underlying biodiversity. Such highly integrative research will rely on the support of novel and effective cyberinfrastructure, tools and organizations to enable scientists to readily mobilize relevant data and analyses across typically 'siloed', incompatible technology frameworks. A successful cyberinfrastructure will provide standards, tools, and frameworks that facilitate cross-dimensional data interoperability, integration, management, analysis, and visualization, while promoting workforce development, collaboration, and other processes that enhance the efficiency and transparency of research and findings for all stakeholders concerned about biodiversity.

To solicit input from a range of scientists on cyberinfrastructure requirements that would support DoB over the program's lifespan, we held a workshop that brought together researchers and developers working in the area of cyberinfrastructure for the biological sciences, as well as researchers who were recently funded as part of the DoB solicitation. The aims of the workshop, as described in the proposal, were to:

1. identify short-term requirements for cyberinfrastructure, as these relate to existing and forthcoming Dimensions research, i.e. inform future solicitations through community engagement to assure appropriate data curation and engender scientific buy-in, and
2. make recommendations on the appropriate procedures that need to be adopted to ensure that, in the long-term, the cyberinfrastructure part of the Dimensions initiative will be adaptable to inevitable changes in the research and IT landscapes.

The workshop proposal also listed 7 areas where recommendations would be made. These were:

1. Cyberinfrastructure requirements and best practices for curation of biodiversity research data assuring accessibility, usability, and interoperability across the Dimensions projects and beyond
2. Strategies for effective community input solicitation for defining CI needs leading to community buy-in
3. Strategies for workforce development and effective interdisciplinary collaboration
4. Areas needing further development of tools and standards
5. Strategies for building CI that is adaptive to changes in research needs and IT
6. Strategies for evaluating the success of CI developments and adoption
7. Strategies for CI developments to become sustainable beyond the lifetime of the Dimensions program

Here we report on the outcomes of the workshop. We discuss the challenges that the participants identified, and the potential solutions to these. While a three-day workshop cannot be expected to provide a well-defined framework for sustainable cyberinfrastructure, we were, nonetheless, able to outline a pathway to achieving the desired outcome -- a sustainable cyberinfrastructure that has community support. We believe that if this pathway, signposted by realistic and achievable milestones, is followed, we will begin to see firm results in 2 - 3 years.

Workshop Structure

This workshop has gathered community input that defines a vision, requirements, procedures and approaches for a cyberinfrastructure to support integrative biodiversity science. Leaders in genetics, genomics and metagenomics, taxonomy and systematics, ecology, and biodiversity informatics research and cyberinfrastructure development identified the key types of cyberinfrastructural support (standards, tools, frameworks) that will be needed by researchers working on Dimensions of Biodiversity projects. Participants were asked to prepare for the workshop by answering the following three question in a short write-up (answers are linked to each participant's name http://lter.limnology.wisc.edu/cidimensions/participants):

- Please list the cyberinfrastructure research and infrastructure projects and products in your program, and describe how they might service and support the Dimensions of Biodiversity program
- From your perspective, identify the 2 most critical informatics research issues on hardware, software, or standards that still need to be resolved to enable an effective Dimensions of Biodiversity program
- From your perspective, identify the 2 most critical pieces of cyberinfrastructure that can be built with existing technology to enable an effective Dimensions of Biodiversity program

The workshop itself started with an introduction of the DoB program and goals for the workshop by NSF program officers and a report of results from earlier related workshops. The feedback to the above questions was summarized and the charge for the workshop set in a third presentation. The participants were asked to discuss three evolving questions in small groups over the next three days:

- What would we like to see in 10 years, in cyberinfrastructure (e.g., hardware, software, data) for biodiversity research?
- Given the vision that you articulated yesterday, and taking into account the discussion at the end of the day, identify a set of milestones and barriers at 3, 5, and 7 years that will lead to or inhibit fulfillment of your vision.
- Design an abstract architecture that enumerates the functional components and provides a basis for interoperability across key functional areas needed for a DoB cyberinfrastructure focused on the 3 year milestones.

Each breakout session was followed by a report back to the entire group and a consensus building plenary discussions.

Vision

We envision the coordinated development of a national cyberinfrastructure to enable the efficient access and use of shared, cross-disciplinary data, tools and services that support biodiversity research that can address unforeseen challenges and threats. Such development needs to provide a framework that promotes an international awareness of what biodiversity is, and why it is important.

In 10 years, data storage and processing should be completely independent of research facilities. Existing knowledge will be accessible in a way that can help with all aspects of the data life cycle. A new experiment can be designed and the data models defined in frameworks that promote re-use of existing data models, ontologies, and practices and automatically generate the software necessary to support the collection and management of those data. All content collected re-uses core primary keys of relevance (temporal, spatial, taxonomic, principals, …) and can be interrelated as necessary for synthesis. The data mesh provides crucial programmatic interfaces (APIs) that support search, discovery, annotation, subscription, extraction, and transformation of data through services that exist as nodes within the mesh infrastructure. Metadata is automatically associated with all data at all processing steps and is seamlessly incorporated with synthetic data products. Published information is identified through unique identifiers, as are the workflows used to process information, and those identifiers are guaranteed to be reliable well into the future. The predominant activity of researchers moves from data management and manipulation to critical research, assisted by gap analysis against existing information and emergent properties apparent through integration of diverse data sets through the common data mesh and processing services. The entire mesh, associated nodes, and related infrastructure are sustained through a diverse set of contributions including pure financial, in-kind donation, and commercial service for a fee.

General Issues

Workshop discussions emphasized that developments should focus on CI issues dealing with DoB specific research and less on the general CI issues that cross programs and domains. However, DoB projects should leverage existing initiatives such as DataNet, iPlant, and other efforts to piggyback on their data management, data interoperability and visualization services (full data lifecycle). It will be critical at the beginning of a cyberinfrastructure development project to focus on the fourteen currently funded DoB projects to capture their management and infrastructure requirements. As a result of this focused effort a realistic definition of what DoB specific CI requirements are should arise and the expectations from the other general CI projects should become clear. A clear roadmap for the unique needs of DoB and how data and technology from individuals and institutions will come together under this effort will be required. In addition a set of specific DoB resources and analytical tools necessary for biodiversity research should result, however it will be important to identify the best of breed from the many parallel initiatives. This approach will provide a strong return on investment because individual and institutional efforts will automatically build into a knowledge base that can spawn further DoB research. In order to do so it is important that DoB promotes an economic ecosystem for collaboration and dissemination of data, tools, and computations while reusing existing technologies and determining the needs that are specific to DoB. In order to achieve these improvements stepwise over 10 years a model for sustained support of CI initiatives from design to implementation and continuing through maintenance will need to be developed.

Workshop participants generally recognized that a coordinated, mission oriented approach to cyberinfrastructure development is necessary to support research in the DoB program. Several scenarios for achieving coordination were discussed and can be described as a biodiversity center, a biodiversity service, or a shared set of capabilities which reflect varying degrees of being distributed, virtual and top-down. However, regardless of distribution, virtualization, and organizational structure it is expected that the degree of coordination needed for a successful CI will require dedicated staff (parallels to the LTER community with its dedicated set of data managers were drawn). Rapid adaption to changing demands might be most effi-

NCBI was often brought up during this workshop as a putative model for the DoB CI in delivering robust tools and services of high quality, in a sustainable manner, to the scientific community conducting biodiversity research. As NCBI, DoB CI would provide long-term storage of data, tools for importing and exporting data, as well as online analytical and visualization tools. Unlike NCBI, it would also provide automated data capturing and data processing pipelines with an interface for manual validation of processed data. The integration mission of DoB CI of centralized and distributed data and services also demarks it from NCBI.

http://www.ncbi.nlm.nih.gov/

A contrasting online data system to NCBI that can serve as a useful model for DoB is the Barcode of Life Data Systems hosted by the University of Guelph. A direct outgrowth of the multitude of ongoing species barcode projects, BOLD has emerged as the primary repository of DNA barcode data, with over 1,000,000 sequences now included (including a subset hosted at NCBI). An important aspect of BOLD that is relevant to the DoB program as well, is the ability of users to retrieve data through different means depending on their technical ability and needs.

ciently achieved by a mixed core and distributed development model, where a pivotal team of computer scientists and programmers, that would focus their efforts on developing new tools for the DoB CI, would collaboratively generate online tools, but would also be available for consultation for software developers outside the core team, and to absorb externally developed software: 1) due to popular demand and when developers would prefer a takeover by DoB CI, and 2) when developers run out of funding and cannot maintain or continue providing access to their software and databases.

Most, but not all, workshop participants saw the need for this coordination in a center like structure, which may be achieved by collaboration of existing NSF supported centers representing the different dimensions of biodiversity or a newly formed center integrating the dimensions of biodiversity. The CI aspects of this 'organization or center' were compared to UNIDATA (http://www.unidata.ucar.edu/), NCBI (http://www.ncbi.nlm.nih.gov/) and others (see sidebar). Around this 'center or organization' a community of practice needs to be developed comparable to the ESIP (Earth Science Information Partners http://www.esipfed.org/) which will be responsible, for example, for developing or adopting existing community accepted standards, data models and best practices, the adoption of a community wide identifier system, identification of relevant tools and approaches, and how much and what to archive. In other words, a community accepted 'governance' structure for cyberinfrastructure needs to be developed.

In addition, such a center would serve a major communication and workforce development function. The benefits of developing a strong cyberinfrastructure framework need to be apparent to researchers in order for them to "buy in" to the larger effort. Organizing and annotating data such that it is understandable by others and can be integrated is a time and resource consuming process. If the effort barrier is too high, especially at the beginning, researchers will be hesitant to support the effort. To help overcome this challenge, the long-term benefits of data sharing need to be clearly communicated to the wider community. It will be a challenge not only to try to convince people that this is important, but rather having concrete suggestions as to how to bring the communities together and provide a useful resource with the recognition that there will be tradeoffs among all of the participants. And a balance will have to be struck between emphasizing future use of data vs. use of data in the current DoB project. Developing and promoting a system to track data

Relative to NCBI, the range of analytical tools available through BOLD reflects its narrower focus on species identification and biodiversity modeling. BOLD is centralized at Guelph and supported by 10-15 staff. The DoB project will generate a broader range of geographic, ecological, genetic and phylogenetic information, but the structure and content of BOLD can serve as a useful starting point for construction of a system that is broader in scope. www.barcodinglife.org

CI for the DoB should serve scientists, citizen scientists, educators and the general public. Encyclopedia of Life (http://www.eol.org) is an example of one way this can work. EOL integrates a broad array of descriptive information, maps, multimedia, and links about the world's living species and higher taxa from many scientific databases and public sources such as Flickr and visitors to EOL pages. An API exposes the integrated, curated information to third-party tools. Scientists may be most interested in diagnostic descriptions, digital literature, and links to molecular and biomedical resources. Citizen scientists might use field guides and smart-phone organism lookup services driven by our API. EOL's aggregation architecture relies on a standards-based XML schema. At the schema's heart are the TDWG Darwin Core standard and the Species Profile Model ontology (SPM).

The Atlas of Living Australia (ALA) (http://www.ala.org.au) is funded by the Australian government to build cyberinfrastructure in order to manage Australian biodiversity data as a tool for research, policy and education. The core component of the Atlas is a comprehensive discovery portal which acts as a broker between different information sources, presents an integrated map of biodiversity data and supplies tools and content to use anywhere to address the needs of the ALA's user communities.

impact statistics (similar to a citation index for publications) should be part of this effort, and should be coordinated with similar data citation efforts in DataONE and other projects. Such an effort will help create a sustained and supportive community. In addition, cross disciplinary communication and mutual understanding need to be promoted as well as capacity building. Workforce development needs to lower the bar for all users to employ community tools.

It will be critical to provide formal mechanisms for engagement between DoB funded projects and the supporting CI community so that the science focus of the cyberinfrastructure is maintained and technology does not drive the solution without consideration of the needs of biodiversity science. This should be initiated early in the science process, possibly even during the proposal period, but certainly no later than in the first few months after funding begins, and should ideally persist through the lifetime of the project. This consultation will produce several benefits. First, it will create an opportunity for scientists to learn what relevant CI resources are available and how to use them effectively, whether involving data discovery and acquisition, data management, analysis, and/or visualization. Second, it will also provide CI developers with real targets that can be used to refine and adapt the development process.

A critical barometer for success of the DoB CI initiative will be the demonstration of real research benefits for one or more DoB projects. Early deliverables will help to affirm the utility of CI development. At the same time, this does not necessarily mean that CI will cater to business-as-usual approaches to data management and analysis. The mission of the CI program should include a broader outreach component that communicates the benefits of good data management and innovative practices over existing approaches. The CI program will face tension over giving the scientist what they want, even if they don't yet know that they want it.

Specific Issues

Standardized Representation of Taxon Concepts

Names for taxonomic concepts (cf. TCS standard at TDWG) are central to Biodiversity informatics (Kennedy et al. 2006). Since these concepts provide "Keys" into many types of biodiversity-relevant information resources, the lack of a sanctioned and widespread approach to representing taxon concepts is an issue that must be solved, and tools and information frameworks based upon it must become generally available. The problem also extends into other naming systems for biotic "classes / categories", including genomic, metagenomic, annotation approaches, etc. as these constitute the named "components" of biodiversity.

Suggestions: For the DoB identify, agree upon and adopt appropriate standards for naming taxonomic concepts and other important classes/categories.

To address the need for standardized representation of taxon concepts the Taxonomic Name Resolution Service project (TNRS) was initiated as a iPlant Incubator Program. This effort is a collaboration between the iPlant Collaborative, the Botanical Inventory and Ecology Network (BIEN), and the Missouri Botanical Garden (MoBoT). This will allow interested data providers and taxonomists to submit names for consideration to then render and explicitly map multiple taxonomic perspectives. Automated and assisted taxonomic standardization will then use Taxon Matching to recover validly published names using exact and fuzzy algorithms to check submitted names and authors against the Tropicos names database returning the canonical spelling of each name or provide suggestions based on the available information.

https://pods.iplantcollaborative.org/wiki/pages/viewpage.action?pageId=3873275

Requirements: Unique references to taxa, clades, phylogenetic nodes etc. must be possible. Transformative search engines are envisioned that, for instance, link via taxonomy to genetics, images, abiotic/biotic information (ecological context) and appropriate tools for analysis, or link to a dynamic tree-of-life to add new forms or recalculate relationships

Accelerating and Supporting Original Data Acquisition

For some dimensions of biodiversity, data capture can be a slow and a mostly manual process involving extensive fieldwork and travel. Robotic, local and remote sensor assistance could increase data acquisition dramatically (e.g. barcoding, video traps, remote missions).

Suggestions: When considering acquisition hardware, a carefully designed network offering support for streaming of data with quality of service (QoS) and integrated with commodity hardware is required. In addition, the hardware infrastructure needs to scale well with growth of data, different structures of data (e.g., raw, relational databases, virtual machine images, spreadsheets, tools), amount of computation (e.g., memory intensive, computationally intensive, communication intensive), high-volume live streams, and users. Scalability can be achieved with good integration in the upper software layer (e.g., data caching , data replication, data quality, job adaptation, co-location of data and processing, and security services).

Requirements: Hardware and software to handle large real time data streams. Data quality control services that are integrated with the data management infrastructure. Network bandwidth and quality of service, integrated with data storage and replication infrastructure. And data caching services, integrated with data sharing infrastructure.

The Integrated Taxonomic Information Service (ITIS) [www.tits.gov] is the standard taxonomic name service for the US Federal Government. It is an interagency effort and has been maintained by a community of experts as a consensus based, synonomized, taxonomic system online since 1996. Its comprehensive business rules and data standards allow it to reliably serve as a full provider directly to other databases. It serves as a significant portion of the Catalogue of Life which populates the taxonomy of EOL and many other database systems. The standard Taxonomic Serial Numbers (TSNs) used, maintained, and resolved by ITIS provide robust name resolution through common web services and profile pages.

Encyclopedia of Life's names infrastructure accommodates multiple hierarchies, including both classifications and phylogenies, and uses synonymies and other algorithms to reconcile name strings to a canonical taxon name. One can use taxon identifiers from one provider to determine the corresponding identifiers from other providers based on matching algorithms, and mine unstructured text for scientific name strings and resolve them. These services are available by API (http://www.eol.org/api, http://www.eol.org/content/page/namelink). EOL and other partners are developing and take advantage of the Global Names Architecture (http://globalnames.org).

Hardware Accessibility for Scientists

Although large computing power is available, it is not considered accessible by most scientists. Typical requirements of high performance computing facilities (e.g., prohibiting or limiting use of non-optimized code, requiring time consuming cycle allocation procedures, requiring use of idiosyncratic batch processing languages) produce extreme barriers to access to the computing resources needed for biodiversity research. The DoB community requires integration of and access to commodity computing hardware (processor, disk, memory and networks) as well as a variety of acquisition hardware (e.g., sensors, geospatial, sequencing).

Suggestions: Improve accessibility to the point where the underlying hardware becomes transparent to the user. Many currently deployed tools will need to be advanced to be able to take advantage of increased computing power.

Requirements: The computational infrastructure providers (e.g., TeraGrid, Open Science Grid, Amazon EC2, FutureGrid, Microsoft Azure, local clusters, campus-wide HPC clusters, and other cloud computing providers) need to provide standard interfaces for a base set of services that will allow users to transparently (a) schedule, launch, monitor, migrate, adapt, recover, manage and audit computing tasks, and (b) store, cache, replicate, migrate, update and audit data. This infrastructure should not require any specialized knowledge by scientist users, and access to these resources should be transparently handled by tools typically used by the biodiversity science community (e.g., R, Matlab).

Long-term Sustainable Data Repository

Storage and access to data beyond any grant cycle currently cannot be guaranteed by most projects. Public data repositories are a basic operational need for a successful DoB program. However, biodiversity data are incredibly disparate, spanning molecular data, characterizations of biotic and abiotic contexts, geospatial, etc. and a rich multitude of associated metadata. While consistency among disparate data types and formats poses obvious challenges, reconciliation of concepts within sub-disciplines must first be settled. For example, unique references to taxa, clades, and phylogenetic nodes, and even consensus over what constitutes an observation etc. remain unresolved. Publicly available data organized without establishment of such basic frameworks inhibits data interoperability and usefulness while making search and the data discovery process difficult.

Suggestions: Establishment of federated data repositories that can guarantee data availability independently from the original data producing projects is urgently needed. Successful data repositories would establish standard data formats, metadata, ontologies, data provenance and workflow documentation and assure interoperability and usefulness of data well beyond the scope of primary research projects. Several data repository efforts are currently under way and DoB should take advantage of and integrate with these projects (DataNet, iPlant). However, different data types may fall under the purview of different repositories and the challenge for the DoB program is that data remain connected, interoperable, and discoverable across these efforts.

Requirements: Important requirements for a federated data repository include but are not limited to persistent unique identifiers for data with integrated provenance tracking and protection from corruption through replica-

The major goal of NSFs DataNet program is to catalyze development of a system addressing the vision outlined in Chapter 3 (Data, Data Analysis, and Visualization) of NSF's Cyberinfrastructure Vision for 21st Century Discovery in which "science and engineering digital data are routinely deposited in well-documented form, are regularly and easily consulted and analyzed by specialists and non-specialists alike, are openly accessible while suitably protected, and are reliably preserved." Towards this end, two awards have thus far been made to establish the DataONE and DataConservancy projects. The two projects differ in their approach and target science domains, though expect to provide interoperable services aiming to improve storage, preservation, access, discovery, and integration of science data both within and between their representative science communities.

DataONE is a federated data network built to improve access to Earth science data, and to support science by: (1) engaging the relevant science, data, and policy communities; (2) facilitating easy, secure, and persistent storage of data; and (3) disseminating integrated and user-friendly tools for data discovery, analysis, visualization, and decision-making. The Data Conservancy will research, design, implement, deploy and sustain data curation infrastructure for cross-disciplinary discovery with an emphasis on observational data.

Each of these DataNet projects offers significant benefits to the research community through the infrastructure and services they provide for data management, and given the general Dimensions In Biodiversity goals of inter-disciplinary research, the data access, and integration capabilities being developed by the DataNet projects seems particularly relevant.

tion (mirroring). Data, metadata, and semantic annotations management software needs to be compatible with the repository. Metadata need to include enough information for the user to judge applicability and usability of the data. Consistent ways to express data quality are needed and the system should automatically annotate and flag issues and allow for review, rating, and annotation from other users.

Security (authentication, authorization, access control, and audit) is an integral part of a CI that spawns all layers of hardware and software. As such, establishment of security mechanisms and policies for DoB CI will facilitate development of and interoperability between all components of the CI.

Coordination of Tools and Services Development

Most NSF-sponsored tool development happens within a grant cycle with limited time, funds and the requirement to produce something creative, innovative, and new. The current funding model generally does not include generalization of the tool, integration with other development, and long term maintenance of the code. This results in uncoordinated, one-off, and sometimes redundant developments. Tools developed by NSF projects rarely transition out of the research prototype stage, largely due to a lack of stable funding. Although a model focused on CI innovation produces interesting directions for computing, it is an ineffective model for building a stable and sustained infrastructure to support science.

Suggestions: A coordinated effort is needed to avoid further developments of highly specialized and partially redundant software as well as a mechanism to generalize and integrate tools and maintain the code base for DoB. In the short term, DoB CI development activities should be largely geared towards meeting the specific needs and requirements of DoB projects, albeit in a way that leverages standard technologies when possible (e.g., service APIs, data services). However, components of this infrastructure should then be generalized over time and with experience. This generalization process should be based on input from the broader community, particularly in cases where we need to reconcile idiosyncratic, incompatible technologies developed by independent groups (e.g., ontologies, exchange formats). If successful, CI development activities will produce a flexible, extensible system that can address unforeseen biodiversity challenges and threats in the coming decade, adapt to new scientific purposes, and ultimately provide value beyond the immediate DoB community.

Requirements: An integrated service infrastructure that serves as a platform for the development of tools. Such infrastructure is based on standards, supports libraries and service APIs for distributed data access, authentication,

The desire to study populations and species across scales demands robust capabilities that connect data elements across modalities (see e.g., Yang et al., 2010, which highlights many of the common cross-cutting needs for spatial computing) . The ability to drill down from top level concepts to individual field samples (and their genetic data) requires the ability to infer relationship and identify connections in a reliable manner on a global scale, which is greatly dependent on the underlying reference data sources. Putting disparate sources of information together into a "Map of Life" (Kidd, *2010)* is a daunting task for scientists with a background in one of the relevant disciplines, but several tools are now available to do this, each with its own focus. **Geophylobuilder** is a plug-in for the established ArcGIS package that has been used for many different types of biodiversity data. **GenGIS** is a standalone application that offers 3D integration of readily available map data, ecological information, and hierarchical information about geographic locations, and includes Python and R for statistical hypothesis testing. Several applications additionally take advantage of highly visible tools like Google Earth for widespread reporting and sharing of data. Future developments in this field will need to access and display very large data sets, and deal with an exponential increase in the number and types of hypotheses concerning the relationship between the abiotic and biotic environments.

etc. It enables people in the community to research and develop data integration, analysis, and visualization software.

Duplication of Data and Redundancies of Tools, and Services

Discovery of appropriate data and tools relevant to a particular project is difficult and leads to collections of duplicate data, and redundant software tools, database, standards, and ontologies simply because existing solutions were not found before work commenced on the project. It is expected this problem will be exacerbated as researchers bridge into less familiar domains while engaging in DoB activities.

Suggestions: Data and software should be documented with appropriate metadata, and those metadata should be indexed by a variety of catalogs (academic and commercial) to assist with discovery. Metadata documents should be identified with persistent identifiers that can be used in publications (e.g. referenced in the methods sections for studies) so that other researchers are made aware of exactly which tools and data were utilized. Annotations associated with the metadata documents (i.e. linked via the persistent identifiers) can provide a basis for increased awareness and ease of discovery of relevant tools and data.

Requirements: Reliable metadata descriptions of data, tools and services. Indexing services for parsing and enabling search for relevant materials. Persistent identifiers for metadata documents that can be utilized in publications. An infrastructure that supports documenting relationships between metadata documents, utilizing the persistent identifiers as linking points.

Standardized Observational Data Model

Biodiversity research often requires analyzing genomic, phylogenetic, and taxonomic information along with a variety of additional biotic and abiotic data (e.g. ecological, geospatial, etc.). Much environmental and other earth science information that can inform biodiversity studies is stored in data tables, where the cell values represent measurements of features and properties of interest (e.g. average rainfall and soil type at a specific location), associated together as the rows or "records" within a data set. These data can encompass observations of a broad range of specialized concepts, and can come from many different original sources. There are currently no easy ways for scientists to readily discover and access these data online, much less interpret and integrate these for analyses. Moreover, the volume and diversity of relevant data becoming available are rapidly growing.

While the specific details vary between observational data formalizations, these all provide a consistent schema for mapping the values of disparate scientific measurements onto terms derived from domain (disciplinary-) specific vocabularies and ontologies. This renders them amenable to various forms of semantic reasoning services, for example for assisting with data discovery, interpretation, and even integration. The NSF-funded Scientific Observations Network (SONet) is working jointly with the DataONE and Data Conservancy projects, along with other institutions and standards bodies, including OGC, TDWG, and GBIF, to develop a unified, "core" model for observational data, that should help resolve some of the most pressing informatics challenges confronting biodiversity researchers.

Suggestion: The adoption of a standard data model with explicit semantics should provide a foundational framework for cross-disciplinary discovery, access, and interpretation of data for integrative biodiversity research.

Requirements: A formalized concept of an observation as a common data structure that can be used to integrate and link disparate data within and across their respective disciplines. An observation should be defined as the realization of a measurement of some characteristic or property of some entity (or a process), collected according to a protocol, with the (measured) value expressed in some specified standard or unit. The capability to associate an observation with a context, such as where and when should also be provided.

The Value of Ontologies

Ontologies will represent a key component of a DoB CI. Ontologies capture the semantics of all elements involved in knowledge production, including data, tools, models, analyses, and even hardware. Ontologies can be leveraged to unify how scientists interact with the disparate and cross-disciplinary kinds of data, tools, and predictive models that will be the hallmark of Dimensions-funded research. Ontologies are also key to making data and tools interoperable among each other. The languages in which ontologies are expressed, such as RDF and OWL, are more expressive for modeling domain data than the relational database or XML Schema models, and they can be extended without breaking legacy software. Moreover, ontology-based annotation and domain data models can form the basis for automating quality control and provenance tracking for data generated within the Dimensions program, which is important as the reproducibility of science depends on metadata with sufficient quality and information.

iPlant is pursuing a program in semantics to deliver infrastructural capability for the description, discovery, engagement, and response handling of semantically-aware web services. The drive is to enable computers to better enable scientists to automate and semi-automate data and service integration based on the semantics (the "meaning") of web-based, distributed data and services. The approach is to deliver to the community a protocol, platform, and capability for the enablement of community-driven semantic integration.

http://www.iplantcollaborative.org/communities/developers/semanticweb

Suggestions: To demonstrate the value of formal ontology-based semantics, and thus the investment into developing the ontology infrastructure, the CI will need to include tools early on that exploit ontology annotations of data in ways that present immediate benefits to users. In particular, such tools could reduce the effort for integrating, manipulating, and analyzing data in complex workflows, which will be common for Dimensions research. In the same way, the semantic annotation of services and data sources should come with a CI component that is capable of semantic mediation between heterogeneous data sources and their different exchange formats, such as the various exchange format flavors currently in use in phylogenetics. Another direct benefit would be an infrastructure for the composition and execution of workflows by the automated semantically-guided discovery of the tools needed for a particular analysis goal.

Various efforts exist already (e.g., Bioportal / OOR, Kepler, MyExperiment, SADI, Semtools, SONet, SSWAP, Taverna) that taken together cover nearly the entire landscape of the CI functions enumerated here at various levels of interoperation and maturity. This includes semantically explicit domain models for scientific observations (e.g., OBOE, O&M, EQ). The main technical challenge for assembling an effective ontology infrastructure within the Dimensions CI will be to integrate the existing products and approaches into a shared platform. Compared to the CI, greater challenges will likely need to be overcome to develop the necessary best practice-compliant ontologies and to effectively train a broad user base.

Requirements: In order for ontologies to fulfill this potential, a DoB CI must provide a number of functions within its architecture for the effective development, management, and utilization of ontologies. Specifically, it must include a platform for authoring and continuously maintaining ontologies in a way that promotes reuse of existing ontologies rather than the proliferation of new ones, and that broadly engages domain experts to ensure community vetting and acceptance. To be effective, such an ontology authoring platform will need to support the mapping between ontologies that overlap in scope, and include a brokering component that makes the addition of terms a lightweight and programmable process while not compromising the term reuse and community vetting objectives. As a software component, this platform will also need to be fully integrated into the CI architecture so that data and metadata management tools can fully leverage ontologies for organizing and annotating data, beginning as early as at the time of data acquisition. This platform will also enable the development of CI components for capturing and augmenting metadata in an automated fashion, a requirement for scaling the semantically rich annotation of data across the Dimensions program.

Documentation and Repeatability of Data Analyses

Currently most data analyses are *ad hoc* and not documented in a way that they could be repeated easily even by anybody directly involved in the original analysis. Any publication of data should be accompanied by documentation that makes the employed workflow repeatable for anyone interested.

Suggestions: Within the DoB identify and agree upon specific scientific workflow systems in which components may be arranged into well documented and repeatable workflows. As a community make tools interoperable within that framework even if each component in a workflow may need to utilize a different underlying analysis or modeling framework.

Requirements: Components are interoperable tools that can be chained into desired workflows. Workflow software needs to automatically generate metadata and annotations, keep track of data provenance and processes, and integrate with data federations and high performance computing facilities to consume data that is processed through scientific workflows to produce and archive new scientific data products.

Sustainability of CI Developments and Data

As pointed out repeatedly most tool developments and data storage currently are not sustainable because they are dependent on funding cycles and the notion that only new and innovative developments are fundable.

Suggestion: To sustainably support research in the Dimensions of Biodiversity program new Cyberinfrastructure innovations may not be the goal. More important are decisions of which innovative development in the community should be sustained. New models of subscription based service or pay per use need to be developed and need to be negotiated with universities, libraries, professional societies, or other service providers.

Requirements: A commitment needs to be made to maintain data and tools from the DoB beyond the individual projects producing them.

Some segments of the biodiversity informatics community have recognized the need for record level annotation with some tools available or under development. Managing annotations involves several distinguishable life-cycle elements, most of which do not differ from those of primary data. However, many familiar CI issues apply to annotations: should they be managed centrally or distributed? What controlled vocabularies and ontologies are needed to allow interoperability between annotation systems and stores? How can they be provided identifiers so that annotations can be referenced in other annotations, and can support data mining in ways that assure the same annotation is not considered twice? What is the scale of storage required if annotations become primary objects of interest and must have the same longevity as traditional primary biodiversity data? Addressing many of these presently are in the remit of a number of recently funded NSF projects, but some of these projects work on combinations of production deployments for specialized sub-communities and at the same time on CI research. This means that some of their accomplishments fall into the milestone timeframes a DoB CI program might set forth. Since these and related projects are funded in several different NSF programs, special attention may be needed to insure that their outcomes become known to DoB CI planners. http://etaxonomy.org/mw/TDWG2010_Annotations_Session_1:_Existing_Systems NSF #0956271, NSF #0851313, NSF #0960535

Education, Outreach and Training

Workforce development is lagging behind CI developments in the area of DoB, contributing to many of the specific issues outlined above. Bridging between CI needs of DoB and solid software engineering knowledge is one area that needs drastically improved mutual understanding and communication skills on both sides.

Suggestion: Social network integration is important. Raising awareness and training even at the early stages of academic training is key to any CI development. Any infrastructure needs to absorb, propagate, and support education, outreach and training. Tools are needed:

- to automate production of best-practice recommendations from validated data annotation, management, and analysis workflows.
- to detect, notify, and correct any data and annotation quality issues for knowledge-transfer (by crowd-sourcing).

Requirements: Scientists should be able to obtain the training necessary to know what tools are available, use them efficiently and help define what is lacking. Scientists should be able to more efficiently enter and store their data in comparable standard formats and effectively communicate with software developers enabling them to construct tools that are useful to a broader community of scientists.

Planning for Success: Suggested Milestones for CI development in DoB

Workshop participants were asked to identify the milestones that would need to be met to ensure the successful implementation of the vision of a coordinated effort to develop CI for the DoB program.

What follows is a set of milestones that have been classified as Strategic (i.e., milestones that ensure that broad strategic plans are in place to ensure that CI development and deployment are successful), Procedural (i.e., milestones that define the types of processes necessary for the CI effort to be implemented), and Operational (i.e., milestones that relate to the nuts-and-bolts of CI for DoB).

These milestones are set against a timeline of 10 years. In the first year, efforts should be directed at identifying and funding critical aspects of the coordinated effort. There was much discussion at the meeting about appropriate structures for coordination. For instance, some participants were unconvinced that a coordinating center would be a workable solution because of the scientific community's antipathy to standards and coordination. Others thought that there was a need for some oversight of effort, but felt that this would be better cast as a consortium or partnership. Consequently, a key challenge will be to ensure that coordination happens to the extent that there is community buy-in, while also acknowledging the resource flexibility that biodiversity scientists demand.

Year 1

Strategic Milestones

- Articulate what the coordination effort addresses: technical, scientific, social issues
- Identify structure of the coordination effort: Is it centralized, federated, distributed, or "ecosystem"-like? Should it be a **Center, Consortium, Partnershp, Community**?
- Is there a governance/executive structure? Is there a physical or virtual presence? Is this suitable for a software institute?
- Define the roadmap (incl. sustainability and training plans) for next 2 yrs, 4 yrs, 9 yrs.

- Begin process for partnerships (*e.g.*, Google, NCBI, NASA, Atlas of Living Australia, etc.) and synergies (*e.g.*, NSF OCI, cross-project collaborations), community engagement, and EOT (Education, Outreach, and Training).

Procedural Milestones

- Delineation of specific needs and requirements as articulated by the current DoB funded projects. One or more workshops and more intensive, sustained mechanisms to bring together scientists and CI people to identify key needs and gaps.
- Establishment of funding programs to build the CI that arises from coordinated activities (tiered to support some big frameworks, other smaller components).
- Assembly/support of one or more teams to develop first projects.

Operational Milestones

- Survey of biodiversity informatics landscape. Survey existing software tools, integrate with existing infrastructure programs (e.g., DataNet).
- Establishment of an online portal of links to existing tools.
- Identification of example data, tools and projects that will immediately benefit from developments and efforts to assure scientist buy-in.

Year 3

Strategic Milestones

- Acceptance of standards/data norms within the scientific community.
- Development of good demo projects showing utilization of the tools/standards and motivate additional adoption.

Procedural Milestones

- Establishment of procedures to ensure that analyses live beyond the snapshot of published papers.
- Measurable progress in better communication between users and CI providers (explicit and demonstrative community engagement).
- Training, education, and outreach components engaged in communities of users and CI providers.

Operational Milestones

- Development of coordinated data catalog /ontologies. There is better abstraction of data, so users can focus on science not technology details.
- Development of frameworks for hardware integration (e.g., linking large repositories /computing facilities).
- Development of a software platform; provides a way to move away from one-off, incompatible, community-specific software solutions.
- Visualization and analysis tools that prove the value of exposing & integration of data.

Year 5

Strategic Milestones

- Increased integration of citizen, informal, and formal science.
- Deep community engagement between users and providers of CI.

- Review of collaborative and EOT engagement.
- New and trained workforce competent in biodiversity CI.
- Prototypic scientific advances in biodiversity, as a consequence of CI developments.

Procedural Milestones

- Mid-term review of processes.

Operational Milestones

- Location- and context-aware data acquisition.
- Consistent use of primary keys and infrastructure for retrieval and relations.
- Consistent ways to express data quality and automatically annotate data.
- Datasets under coordination (priority information resources identified and interoperable; cross-disciplinary data integration).
- Tools under coordination (community development and public APIs [Application Programming Interfaces]).
- Services under coordination (matchmaking and provisioning).
- Hardware under coordination (national centers, cooperative agreements).

Year 7

Strategic Milestones

- Well developed plan and strategy for Yr 10+ sustainability and adaptability.
- Demonstrable scientific advances in biodiversity facilitated by CI.
- Next-gen informatics technology is included in existing systems or has been accounted for.

Procedural Milestones

- All processes are mature and well developed.

Operational Milestones

- Fully realized semantic model of DoB data, tools that can use and navigate/access this.
- Crowdsource annotations as people work with existing data.
- Development of tools to automate some of the metadata capture for DoB data.
- Mandatory data deposition is part of all science disciplines in DoB.
- Dedicated workflow infrastructure that is capable of doing data-intensive work and automates documentation of the data and procedures applied through the full data lifecycle.
- Automatic generation of new databases, recycled and in the ecosystem.
- Real time handling of streaming data and effective ways to filter, cluster and analyze it.
- Robotic and assisted data capture.

References

Kennedy, J., Gales, R., Kukla, R., Hyam, R., Wieczorek, J., Hagedorn, G., Döring, M., Vieglais, D. (2006). Developing a Core Ontology for Taxonomic Data. In: Belbin, L., Rissoné, A., Weitzman, A. (Eds.) Proceedings of TDWG (2006), St Louis, MI, USA

Kidd D. "Geospatial phylogenies and the Map of Life". *Systematic Biology* 59 (2010) 741-752.

NSF's Cyberinfrastructure Vision for 21st Century Discovery
http://www.nsf.gov/pubs/2007/nsf0728/index.jsp

Yang C, et al. "Geospatial Cyberinfrastructure: Past, present and future". *Computers, Environment and Urban Systems* 34 (2010) 264–277.

Appendices

Tools and Technologies

The following is a compilation of technologies as they were represented by the workshop participants and described by them in more detail at http://lter.limnology.wisc.edu/cidimensions/participants . This is by no means an exhaustive list of relevant technologies available today.

Name	Description	Web site
AKN	The Avian Knowledge Network (AKN) is an international organization of government and non-government institutions focused on understanding the patterns and dynamics of bird populations across the Western Hemisphere. Currently almost 50 organizations have contributed more than 85 million bird observations.	http://www.avianknowledge.net
Australian National Data Service (ANDS)	A large multi-faced project with social, educational, and technical projects about the management of scientific data of which an ambituous one is "transform the disparate collections of research data around Australia into a cohesive collection of research resources."	http://ands.org.au/
CAMERA	CAMERA - Community Cyberinfrastructure for Advanced Microbial Ecology Research & Analysis. The aim of this project is to serve the needs of the microbial ecology research community, and other scientists using metagenomics data, by creating a rich, distinctive data repository and a bioinformatics tools resource that will address many of the unique challenges of metagenomic analysis.	http://camera.calit2.net/
Center for Tropical Forest Science	A global network of large tree plots in tropical forests, all utilizing a standardized data collection methodology. Now includes temperate forest plots in North America and Europe.	http://www.ctfs.si.edu/
DataOne	A distributed global network of Member Nodes (i.e., data repositories) that provide open and persistent access to well-described and easily discovered Earth observational data. In addition, a smaller number of Coordinating Nodes (i.e., metadata repositories and service centers) support network-wide services such as data replication and access to an array of enabling tools.	https://www.dataone.org/
Distributed Active Archive Centers (DAACs)	The centers process, archive, document, and distribute data from NASA's past and current research satellites and field programs and additional supporting data. Each center serves one or more specific Earth science disciplines and provides data products, data information, services, and tools unique to its particular science	http://nasadaacs.eos.nasa.gov/

DRYAD	A general-purpose digital repository for published scientific data. Dryad partners with an expanding consortium of biological journals, implements incentives for data depositors, researches ways to automatically augment metadata to improve findability, and uses hand-shaking protocols with specialized repositories to maintain cross-links to related data records hosted by those.	http://datadryad.org
Earth Observing System Data and Information System	The Earth Science Data and Information System (ESDIS) Project processes, archives and distributes NASA Earth science satellite data (e.g., land, ocean and atmosphere data products) and field campaign data, provides tools for this work, and ensures that scientists and the public have access to this data. It also includes additional supporting data. ESDIS manages and provides earth science data through the Earth Observing System Data and Information System (EOSDIS). EOSDIS manages data for more than two dozen NASA satellites and instruments. EOSDIS is designed as a distributed system, with major facilities at data centers located throughout the United States.	http://esdis.eosdis.nasa.gov/
eBird	Maximize utility of 50 million (and rising) Citizen Science and professional bird occurrence observations.	http://www.ebird.org
Ecological Metadata Language (EML)	A formal metadata specification, intended to encompass the widely heterogeneous data formats encountered in ecology	http://knb.ecoinformatics.org/software/eml/
Electronic Field Guide (EFG)	An open source platform that provides field biologists with simple ways to build web accessible or standalone taxonomic identification tools aimed at different audiences starting from a single table of morphological, phenological or other attributes of a set of taxa.	http://efg.cs.umb.edu
Encyclopedia of Life (EOL)	EOL provides a number of tools, including mechanisms for aggregating summary information, references, multimedia, and maps about taxa across the entire tree of life on subjects spanning all of biology. These mechanisms handle all kinds of original sources, from databases to web services, spreadsheets, PDFs, and HTML pages. An open-source tool (LifeDesk) for capturing descriptive data about species and their taxonomy in a way that provides both an independent web presence for a research community but also exports of standards-compliant data to be shared with EOL and other data consumers. A names infrastructure that accommodates multiple hierarchies, including both classifications and phylogenies, and uses synonymies and other ways to reconcile name strings to taxon concepts. A tool for finding scientific names in any digital source. An API that allows machine access to all of our information. Soon to come will be processes to propagate placeholder data to children in the tree of life or to designate exemplars where little data are available, thus filling	http://www.eol.org/

	gaps until research results are available.	
Filtered Push	Develops tools for the production, exploitation, management and distribution of annotations of scientific data, with emphasis on biodiversity data. The current focus is the use of distributed annotations of distributed data for data quality control.	http://etaxonomy.org/mw/FilteredPush
FutureGrid	An experimental, High-Performance Grid Test-bed, FutureGrid makes it possible for researchers to tackle complex research challenges in computer science related to the use and security of grids and clouds. These include topics ranging from authentication, authorization, scheduling, virtualization, middleware design, interface design and cybersecurity, to the optimization of grid-enabled and cloud-enabled computational schemes for researchers in astronomy, chemistry, biology, engineering, atmospheric science and epidemiology.	http://futuregrid.org
GenGIS	A standalone application that offers 3D integration of readily available map data, ecological information, and hierarchical information about geographic locations, and includes Python and R for statistical hypothesis testing. Several applications additionally take advantage of highly visible tools like Google Earth for widespread reporting and sharing of data.	http://kiwi.cs.dal.ca/GenGIS
Genomics Sequence Consensus Data System	A data system for the storage, annotation, query, and analysis of genomic sequence data. The system is extending the existing open source SeqWare software developed at UCLA and UNC. This system leverages the Hadoop/HBase technologies to allow for scaling to 1000s of genomes while capturing data on both genetic variants and on quality scores at variant and non-variant genomic locations.	http://www.renci.org/
In-VIGO ICBR	In-VIGO ICBR is a web portal that allows Interdisciplinary Center for Biotechnology Research (ICBR) clients to run extremely large bioinformatics (BLAST) jobs on high performance computing resources managed by a queuing system through a user-friendly web interface in a fault-tolerant manner.	http://invigo.acis.ufl.edu/icbr/
iPlant	A major NSF initiative to develop CI for many aspects of botany. Special emphasis on two Grand Challenges. One is to construct a phylogenetic tree of all plant life. Second is to advance understanding of how genes influence the phenotype.	http://www.iplantcollaborative.org/
iRODS@RENCI	Aids in the application of the iRODS data grid technology to different scientific domains. iRODS is a software technology for managing distributed data collections, with a focus on scientific collections	http://www.renci.org/
Kepler	A scientific workflow application that builds upon a mature set of java numerical libraries, to allow orchestrating heterogeneous analytical components into an executable, sharea-	http://kepler-project.org

	ble, extendible, archivable visual workflow format.	
Knowledge Network for Biocomplexity	A distributed data repository that provides the scientific community with a means for describing, sharing, archiving, discovering, accessing, and interpreting datasets. Its ability to accommodate arbitrarily heterogeneous types of data makes it ideal for serving the data management needs of institutions, field stations, and individual researchers collecting highly variable forms of biodiversity data.	http://knb.ecoinformatics.org
MASSIVE	Multimodal Australian Sciences Imaging and Visualisation Environment. An Australian National Computational Infrastructure facility.	http://www.massive.org.au
Metacat	A Metadata Catalog database, that provides distributed, authenticated, replicated services for sharing of EML and other XML documents. Basic construction involves implementation of a DOM model in a web-enabled RDBMS.	http://knb.ecoinformatics.org
MOA Database	A custom in-house MySQL database at Dalhousie University that integrates microbial genomic, metagenomic, and EST data to support analyses via either the Web, user applications such as GenGIS, or directly through scripts	
Morpho	A java-based desktop tool that enables creation of EML documents for managing one's own data, but also a capable querying client for a local Metacat, or onto the global KNB.	http://knb.ecoinformatics.org
NBCR - National Biomedical Computation Resource	One of NBCR's roles has been to anticipate and to ease the transition through technology changes for the biomedical community, thus allowing the community to harness the new capabilities without having to invest time to either redevelop codes or algorithms or implement on new machines.	http://www.nbcr.net/ , http://www.nbcr.net/tools.php
NeXML, PhyloWS	Standards for exchanging (NeXML) and accessing (PhyloWS) phylogenetic data with rich metadata, external cross-references, and explicit semantics.	
Phenoscape	Applies an ontology-based logic formalism, called Entity-Quality (EQ), to the evolutionary phenotype diversity recorded as free text character descriptions in the systematics. The Phenoscape Knowledgebase demonstrates how this approach can render descriptive data amenable to computational knowledge integration and reasoning on a large scale.	http://kb.phenoscape.org
RENCI GeoViz	This platform is looking at new ways to enable collaboration around geospatial data using shared geo-workspaces.	http://www.renci.org/
Renci Science Portal	The RENCI Science Portal is a software application that manages the submission of software codes to different computational resources, including the Open Science Grid, Teragrid, and RENCI-based computational capabilities.	http://www.renci.org/

RENCI Sensor Bus	A platform for integration, management, analysis, and visualization of sensor data.	http://www.renci.org/
Rocks clustering toolkit	Rocks is an open-source Linux cluster distribution that enables end users to easily build computational clusters, grid endpoints and visualization tiled-display walls. Hundreds of researchers from around the world have used Rocks to deploy their own cluster	http://www.rocksclusters.org
SeqMonitor	A prototype Web application that supports geographic queries on automatically updated reference genetic data sets.	http://ratite.cs.dal.ca/SeqMonitor
Sky Computing	Sky computing combines several technologies and techniques to execute a popular bioinformatics application called BLAST on distributed resources and achieve scalable management and performance. It combines an IaaS cloud toolkit (Nimbus) to create virtual machines (VMs) on demand, a virtual networking middleware (ViNe) to connect VMs, the MapReduce framework (Hadoop) for parallel fault-tolerant execution of an unmodified application, and a skewed task distribution technique to deal with resource imbalance. Massively parallel applications are best suited for sky computing.	http://www.nimbusproject.org/files/Sky_Computing.pdf
SNAP Workbench	A Java program that manages and coordinates a series of programs. The workbench enhances population parameter estimation by ensuring that the assumptions and program limitations of each method are met and by providing a step-by-step methodology for examining population processes that integrates both summary-statistic methods and coalescent-based population genetic models.	http://snap.cifr.ncsu.edu
TARDIS	Federated Repository for Scientific Imaging. Aimed at raw data in nuclear science.	http://tardis.edu.au/
Taxon Concept Schema, TCS	A formal approach towards dealing with the ambiguity of taxonomic names, as these change through time or mean different things according to different authorities.	http://www.tdwg.org/standards/117/
TreeBASE	A community repository of phylogenetic data and trees hosted by NESCent	http://treebase.org
VisTrails	A scientific workflow authoring, execution and management environment with a focus on visualization.	http://www.vistrails.org/index.php/Main_Page

Participants

Reed Beaman, NSF, rbeaman@nsf.gov

Robert Beiko, Faculty of Computer Science Dalhousie University, Halifax, Nova Scotia. Canada, beiko@cs.dal.ca

Paul Bonnington, Monash University, Australia, Director - Monash e-Research Centre, paul.bonnington@adm.monash.edu.au

Christy Geraci, NSF, cgeraci@nsf.gov

Damian Gessler, Semantic Web Architect, iPlant, dgessler@iplantcollaborative.org

George W. Gilchrist, NSF, ggilchri@nsf.gov

Corinna Gries, Information Manager, LTER, cgries@wisc.edu

Gerald "Stinger" Guala, Director of the Integrated Taxonomic Information System (ITIS), and Senior Administrator in the National Biological Information Infrastructure (NBII), U.S. Geological Survey, gguala@usgs.gov

Steve Huter, Network Startup Resource Center at the University of Oregon, sghuter@nsrc.org

Matthew Jones, Director of Informatics Research and Development at NCEAS, jones@nceas.ucsb.edu

Steve Kelling, Information Science, Cornell Lab of Ornithology, stk2@cornell.edu

Jessie Kennedy, Professor of Computer Science, Edinburgh Napier University, j.kennedy@napier.ac.uk

Suzanne Lao, Data Manager, Tropical Forest Science (CTFS) at the Smithsonian Tropical Research Institute, LAOZ@si.edu

Hilmar Lapp, Assistant Director of Informatics at NESCent, hlapp@nescent.org

Allison Leidner, AAAS Science & Technology Policy Fellow, Research & Analysis Program, Earth Science Division, NASA Headquarters, aleidner@gmail.com

Francois Lutzoni, *Evolution of symbiotic systems,* Associate Professor, Duke University, francois.lutzoni@duke.edu

Andrea Matsunaga, University of Florida, ammatsun@acis.ufl.edu

Chris Mentzel, Program Officer, Science Program, Gordon & Betty Moore Foundation, chris.mentzel@moore.org

Nirav Merchant, IT Director for ARL, iPlant, nirav@email.arizona.edu

William Michener, Director, New Mexico EPSCoR State Program, Associate Director, LTER Network Office, PI, DataONE, wmichener@lternet.edu

Robert A. Morris, Emeritus Professor of Computer Science, UMASS-Boston, ram@cs.umb.edu

Phil Nista, Indiana University, Department of Biology, pnista@indiana.edu

Phillip Papadopoulos, Program Director, UC Computing Systems, University of California, San Diego, philip.papadopoulos@gmail.com

Cynthia Parr, Director, Species Pages Group, Encyclopedia of Life, parrc@si.edu

Dan Phillips, USGS Center for Biological Informatics, dphillips@usgs.gov

Jim Regetz, Scientific Programmer/Analyst, NCEAS/UCSB, regetz@nceas.ucsb.edu

Allen Rodrigo, Director, NESCent, a.rodrigo@nescent.org

Mark Schildhauer, Director of Computing, NCEAS, schild@nceas.ucsb.edu

Charles Schmitt, Renaissance Computing Institute (RENCI), cschmitt@renci.org

Dave Vieglais, Director for Development and Operations - DataOne, vieglais@ku.edu

Glossary of Abbreviations

ALA	Atlas of Living Australia (http://www.ala.org.au/)
Amazon EC2	Amazon Elastic Cloud 2 (http://aws.amazon.com/ec2/)
API	Application Programming Interface
Bioportal / OOR	Open Ontology Repository (http://www.oor.net/home/release)
BOLD	Barcode of Life Data Systems (http://www.boldsystems.org/views/login.php)
CI	CyberInfrastructure
DataConservancy	http://dataconservancy.org
DataNet	National Science Foundation DataNet program (http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=503141&org=OCI)
DataONE	Data Observation Network for Earth (http://dataone.org)
DoB	Dimensions of Biodiversity (http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=503446)
EOL	Encyclopedia of Life (http://www.eol.org/)
EOT	Education, Outreach and Training
EQ	Entity-Quality Ontology
ESIP	Earth Science Information Partners (http://www.esipfed.org/)
FutureGrid	High-performance computing grid test bed (http://futuregrid.org/)
HPC	High Performance Computing
iPlant	iPlant Collaborative (http://iplantcollaborative.org)
ITIS	Integrated Taxonomic Information System (http://www.itis.gov/)
Kepler	Kepler Scientific Workflow System (http://kepler-project.org)
LTER	Long-term Ecological Research Network (http://www.lternet.edu/)
Matlab	http://www.mathworks.com/
Microsoft Azure	http://www.microsoft.com/windowsazure/
MyExperiment	http://myexperiment.org
NASA	National Aeronautics and Space Administration (http://www.nasa.gov)
NCBI	National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/)
NSF	National Science Foundation (http://www.nsf.gov/)
NSF-OCI	National Science Foundation Office of Cyberinfrastructure (http://www.nsf.gov/dir/index.jsp?org=OCI)

O&M	Observations and Measurements Standard (http://www.opengeospatial.org/standards/om)
OBOE	Extensible Observation Ontology (http://dx.doi.org/doi:10.1016/j.ecoinf.2007.05.004)
Open Science Grid	http://www.opensciencegrid.org/
OWL	Web Ontology Language (http://www.w3.org/TR/owl-features/)
QoS	Quality of Service
R	The R Project for Statistical Computing (http://www.r-project.org)
RDF	Resource Description Framework
SADI	Semantic Automated Discovery and Integration (http://sadiframework.org/content/)
Semtools	Semantic Tools for Ecological Data Management (https://semtools.ecoinformatics.org/)
SONet	Scientific Observations Network (http://sonet.ecoinformatics.org)
SSWAP	Simple Semantic Web Architecture and Protocol (http://sswap.info/)
Taverna	Workflow Management System (http://www.taverna.org.uk/)
TCS	Taxonomic Concept Schema (http://www.tdwg.org/standards/117/)
TDWG	Taxonomic Databases Working Group (http://tdwg.org)
TeraGrid	https://www.teragrid.org/
UNIDATA	Services, tools and cyberinfrastructure for earth systems science (http://www.unidata.ucar.edu/)
XML Schema	http://www.w3.org/XML/Schema