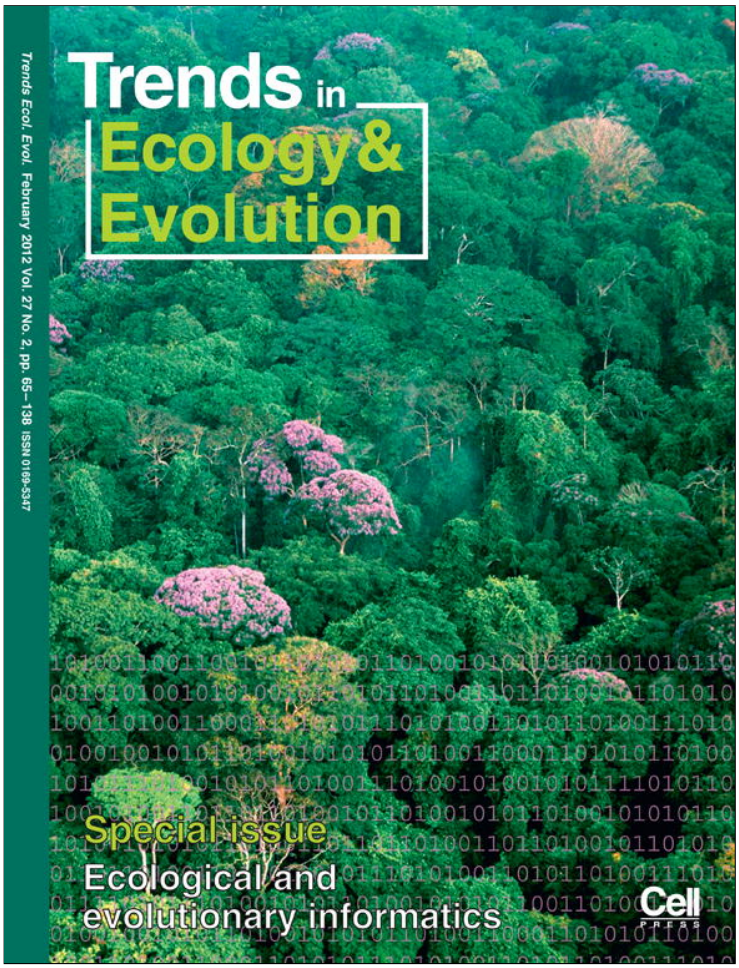

Review

Cell PRESS

*Special Issue: Ecological and evolutionary informatics*

# Ecoinformatics: supporting ecology as a data-intensive science

**William K. Michener[1] and Matthew B. Jones[2]**

[1] University Libraries, University of New Mexico, Albuquerque, NM 87131, USA
[2] National Center for Ecological Analysis and Synthesis, University of California Santa Barbara, Santa Barbara, CA 93101, USA

**Ecology is evolving rapidly and increasingly changing into a more open, accountable, interdisciplinary, collaborative and data-intensive science. Discovering, integrating and analyzing massive amounts of heterogeneous data are central to ecology as researchers address complex questions at scales from the gene to the biosphere. Ecoinformatics offers tools and approaches for managing ecological data and transforming the data into information and knowledge. Here, we review the state-of-the-art and recent advances in ecoinformatics that can benefit ecologists and environmental scientists as they tackle increasingly challenging questions that require voluminous amounts of data across disciplines and scales of space and time. We also highlight the challenges and opportunities that remain.**

## Ecology as an evolving discipline

Ecology is increasingly becoming a data-intensive science (see Glossary) [1,2], relying on massive amounts of data collected by both remote-sensing platforms [3] and sensor networks that are embedded in the environment [4–7]. New observatory networks, such as the US National Ecological Observatory Network (NEON) [8] and Global Lake Ecological Observatory Network (GLEON) [9], provide research platforms that enable scientists to examine phenomena across diverse ecosystem types through access to thousands of sensors collecting diverse environmental observations. It has been postulated that data-intensive science represents the fourth scientific paradigm following the empirical (i.e. description of natural phenomena), theoretical (e.g. modeling and generalization) and computational (e.g. simulation) scientific approaches, and comprises an approach for unifying theory, experimentation and simulation [2].

Ecologists increasingly address questions at broader scales that have both scientific and societal relevance. For example, the 40 top priorities for science that can inform conservation and management policy in the USA rely principally on a sound foundation of ecological research [10]. As ecology expands its scope, it is becoming more collaborative and network and team based [11–13]. For example, research at individual long-term ecological research (LTER) sites in the USA is conducted collaboratively by teams consisting of an average of 18 cooperating investigators and 20 graduate students; inter-site and network-wide studies add further to the scope and scale of the LTER research enterprise [14].

Ecology is also affected by changes that are occurring throughout science as a whole. In particular, scientists, professional societies and research sponsors are recognizing the value of data as a product of the scientific enterprise and placing increased emphasis on data stewardship, data sharing, openness and supporting study repeatability [15–17].

The changes that are occurring in ecology create challenges with respect to acquiring, managing and analyzing the large volumes of data that are collected by scientists worldwide. One challenge that is particularly daunting lies in dealing with the scope of ecology and the enormous variability in scales that is encountered, spanning microbial community dynamics, communities of organisms inhabiting a single plant or square meter, and ecological

## Glossary

**Cloud computing**: provision of computing cycles, storage resources and software as a service that is accessible from the Internet via a standardized approach that treats these shared resources as a commodity utility.
**Data-intensive science**: a transformative, new way of doing science that entails the capture, curation and analysis of massive amounts of data from an array of sources, including satellite and aerial remote sensing, instruments, sensors and human observation.
**Data life cycle**: the data life cycle encompasses all facets of data generation to knowledge creation, including planning, collection and organization of data, quality assurance and quality control, metadata creation, preservation, discovery, integration, and analysis and visualization.
**Faceted search**: faceted search or faceted browsing enables users to discover specific data products by filtering a set of available descriptors. Each facet corresponds to the array of possible values of a property that is common to a set of data products, such as author, data center where the data are stored, sensors used to collect the data, and ecosystem or habitat type where the data were collected.
**Metadata**: documentation describing all aspects of the data (e.g. who, why, what, when and where) that would allow one to understand the physical format, content and context of the data, as well as possibly how to acquire, use and cite the data.
**Ontology**: a formal representation or classification of concepts and their relationships within a domain of interest.
**Provenance**: in science, data provenance refers to the ability to track data from creation through all transformations, analyses and interpretations, enabling full understanding of the processes used to create derived scientific products.
**Quality assurance/quality control (QA/QC)**: refers to the mechanisms for preventing errors from entering a data set that are used *a priori* to ensure high data quality before collection and to monitor and maintain data quality during and after the data collection process.
**Semantic annotation**: ascribing links from data to classes in an ontology.
**Scientific workflow system**: a computational platform that is designed to compose and execute a series of data acquisition, data processing and analytical steps as part of a workflow, in a scientific application.

*Corresponding author:* Michener, W.K. (william.michener@gmail.com)

0169-5347/$ – see front matter  doi:10.1016/j.tree.2011.11.016 Trends in Ecology and Evolution, February 2012, Vol. 27, No. 2 

processes occurring at the scale of the continent and biosphere. The diversity in scales studied and the ways in which studies are carried out results in large numbers of small, idiosyncratic data sets that accumulate from the thousands of scientists that collect relevant biological, ecological and environmental data [18]. Such heterogeneity can be attributed, in part, to methodological specialization to address specific scientific hypotheses, but also to a lack of standard protocols for acquiring, organizing and describing data and language barriers and cultural differences across disciplines, institutions and countries.

In the remainder of this paper, we define ecoinformatics, describe existing tools and approaches, and highlight recent advances. We then identify remaining challenges and opportunities and recommend approaches for better incorporating ecoinformatics into the research enterprise.

## What is ecoinformatics?

Ecoinformatics is a framework that enables scientists to generate new knowledge through innovative tools and approaches for discovering, managing, integrating, analyzing, visualizing and preserving relevant biological, environmental, and socioeconomic data and information. Many ecoinformatics solutions have been developed over the past decade, increasing scientists' efficiency and supporting faster and easier data discovery, integration and analysis; however, many challenges remain, especially in relation to installing ecoinformatics practices into mainstream research and education.

## The data life cycle

Knowledge is derived through the acquisition of data and the transformation of those data into information that can be incorporated into the corpus of scientific facts, principles and theories. Figure 1 illustrates the different stages that data might progress through during the processes that lead to new information and knowledge. Two stages are reflected in this depiction of the data life cycle. First, projects that include collection of new data typically proceed through steps 1–5 (i.e. plan, collect, assure, describe and preserve) and then can proceed directly to step 8 (i.e. analysis). Second, synthesis efforts or meta-analysis can initially start at step 6 (i.e. discover relevant data) and proceed to step 7 (i.e. integration of data from various sources) and, finally, to step 8 (i.e. analysis). The stages are not necessarily exclusive and the steps need not be sequential. For instance, a synthesis effort would probably include step 2 (i.e. assure) after step 6 (i.e. discover) and before step 7 (i.e. integrate). Ecoinformatics tools and techniques associated with each step of the data life cycle are described below.

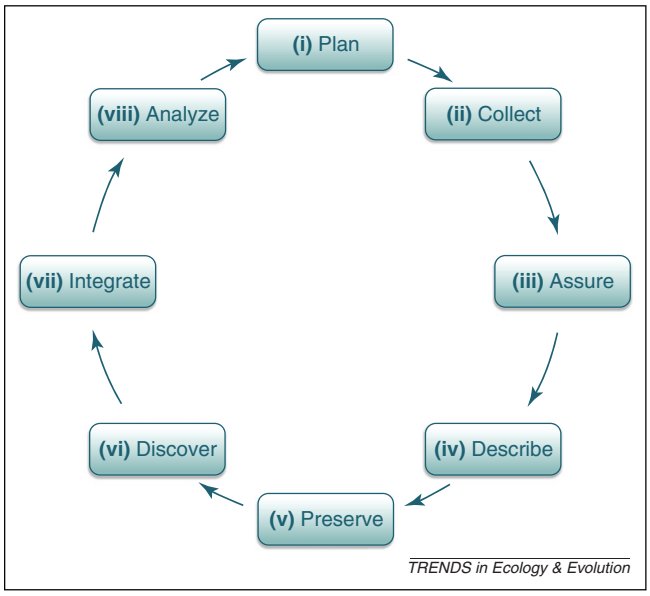


**Figure 1**. The data life cycle includes the following steps: **(i)** plan; **(ii)** collect; **(iii)** assure (i.e. quality assurance and quality control); **(iv)** describe (i.e. ascribe metadata); **(v)** preserve (i.e. deposit data in a secure data repository; **(vi)** discover (i.e. identify data that might be needed to answer a question); **(vii)** integrate (e.g. merge data from multiple data sources); and **(viii)** analyze (e.g. statistical analysis, visualization). Modified after Figure 1 in [22] with the permission of C. Strasser.

### *Step 1: plan*

Data management planning is often underappreciated and underutilized in project design, yet it can save time, enhance research efficiency and, importantly, satisfy requirements of research sponsors that increasingly require explicit data management plans as part of research proposals. Although practical guidelines for data management have been outlined [19,20], there is a need for more comprehensive planning to support open science. In response to these community needs, a data management planning tool (i.e. DMP Tool; http:/dmp.cdlib.org/) was designed to aid researchers in creating, reviewing and revising data management plans, recognizing that data management plans should be living documents that change in response to project needs and availability of new technologies. The DMP Tool is based on a similar tool developed by the Digital Curation Centre in the UK (http://www.dcc.ac.uk/) and includes five components that should ideally be addressed in a comprehensive data management plan (Table 1) [21,22].

### *Step 2: collect*

Ecological data are collected and organized in many different ways, including manual recording of observations in the laboratory and field via hand-written data sheets, tape recorders and hand-held computers; automated data collection via laboratory and field instrumentation; satellites and aerial platforms; and, increasingly, sensor networks that are embedded in the environment. Decreases in the size, cost and power requirements of sensors have revolutionized their use to monitor biota and environmental processes [4,5] and provide access to those data in real or near-real time [9]. New environmental observing systems, such as NEON [8] and the Ocean Observatories Initiative (OOI) [23], will provide access to data collected by aerial, ground-based and underwater sensor networks encompassing tens of thousands of sensors that, when combined, will generate terabytes to petabytes of data annually.

Many different approaches and tools are presently used for data organization and management, ranging from spreadsheets and statistical software to relational database management systems to geographic information systems. Each approach has advantages and disadvantages.

**Table 1. Summary of components that should be described in a comprehensive data management plan[a]**

| Component | Description and examples |
|---|---|
| Information about data and data format | Types of data that will be produced (e.g. experimental, observational, raw or derived, physical collections, models, images, etc.) |
| | When, where and how the data will be acquired (e.g. methods and instruments used) |
| | How the data will be processed (e.g. software, algorithms and workflows) |
| | File formats (e.g. csv, tab-delimited or naming conventions) |
| | QA/QC procedures used |
| | Other sources of data (e.g. origins, relationship to one's data and data integration plans) |
| | Approaches for managing data in the near-term (e.g. version control, backing up, security and protection, and responsible party) |
| Metadata content and format | Metadata that are needed |
| | How metadata will be created or captured (e.g. lab notebooks, auto-generated by instruments, or manually created) |
| | Format or standard that will be used for the metadata (e.g. EML or ISO 19115) |
| Policies for access, sharing and re-use | Requirements for sharing (e.g. by research sponsor or host institution) |
| | Details of data sharing (e.g. when and how one can gain access to the data) |
| | Ethical and privacy issues associated with data sharing (e.g. human subject confidentiality or endangered species locations) |
| | Intellectual property and copyright issues |
| | Intended future uses for data |
| | Recommendations for how the data can be cited (e.g. citation and DOI) |
| Long-term storage and data management | Identification of data that will be preserved |
| | Repository or data center where the data will be preserved |
| | Data transformations and formats needed (e.g. data center requirements and community standards) |
| | Identification of responsible parties |
| Budget | Anticipated costs (e.g. data preparation and documentation, hardware and software costs, personnel costs and archive costs) |
| | How costs will be paid (e.g. institutional support or budget line items) |

[a]See http://cdlib.dmp.org.

For instance, it is easy to introduce errors into spreadsheets, as one can mix diverse types of data within a single column (e.g. dates, numeric values and text) and data summaries are frequently conflated with raw data. By contrast, in using relational databases, one can employ constraints on the types of data that can be entered (e.g. data typing), which can be used to assure data integrity [24]. Within the US LTER Program, relational databases play a key role in data entry and metadata preparation as well as facilitating data integration and analysis [25]. Statistical software tools support many of the functions available through spreadsheet programs and provide the added benefit of supporting robust calculations, data analysis, quality assurance, visualization and data sub-setting. We anticipate that many existing boundaries among data management tools will increasingly become blurred as spreadsheets (or spreadsheet add-ons) enforce data typing and adopt other procedures commonly found in relational databases, and as relational databases provide greater support for geospatial data.

### *Step 3: assure*

Quality assurance and quality control (QA/QC) refers to the mechanisms for preventing errors from entering a data set that are used *a priori* to ensure high data quality before collection and to monitor and maintain data quality during and after data collection. Prior to data collection, QA can consist of defining standards for formats, codes, measurement units and metadata (see Step 4), as well as assigning responsibility for data quality to a specific individual or team. Quality control activities range from using two individuals to independently enter data and then compare results and rectify differences (i.e. 'double entry'), to using database approaches that allow one to minimize the number of times that data must be entered repeatedly, to enforce data typing and to incorporate easily illegal value filters and range checks. Many software packages, such as R and SAS, provide algorithms and procedures that allow one to visualize easily data and identify extreme values and potential outliers.

One active area of research lies in integrating QA/QC with data and metadata management systems and scientific workflow systems so that well-documented data can be automatically assessed for metadata completeness and data quality [25–29]. For instance, the LTER Network Information System (NIS) is designed so that data products are evaluated, processed and classified according to five categories, ranging from raw or minimally modified site data (Level-0) to data that are gap filled and semantically adjusted to meet the needs of specific synthetic data products (Level-4) [25].

### *Step 4: describe*

Metadata provide sufficient documentation so that one is able to understand the content, format and context of a data product. Metadata typically describe: (i) who created, collected and managed the data; (ii) the data content and format; (iii) when the data were collected; (iv) where the data were collected and stored; (v) how the data were generated, processed, assured and analyzed; and (vi) why the data were generated (i.e. the study context) [30–32].

Critically, metadata enable a scientist to understand and use the data; this is particularly important for study reproducibility and for synthetic efforts, such as meta-analysis, where the goal might be to integrate and compare data across many studies, looking for general trends or emergent properties as the scale of study expands. Various metadata standards and tools have been developed to provide consistency in the content and format of metadata, and to facilitate the creation and management of metadata (see the supplementary material online) [33,34].

### *Step 5: preserve*

Data preservation encompasses the deposition of data and metadata in a data center or data repository where the data can be verified, replicated and actively curated over time (e.g. including migration of data to new storage media as old media are replaced or become outdated) [35,36]. Data centers support different levels of data verification, replication and curation, because of the costs involved and the difficulty in automating many of these procedures. Data centers can be associated with governmental and nongovernmental organizations, universities, libraries, environmental observatory networks, and commercial and non-profit enterprises. Frequently, a data center supports a specific community of practice that can be associated with a particular research sponsor, home institution, or thematic area. User help-desk support, peer-review of data products and assignment of Digital Object Identifiers (DOIs) to data products so that they can be uniquely identified and cited, represent some of the services that data centers can offer to their stakeholders. The Oak Ridge National Laboratory Distributed Active Archive Center for Biogeochemical Dynamics, for example, assigns DOIs to its data products and tracks the usage of data products as a service to the data providers and research sponsor [37].

### *Step 6: discover*

Data discovery remains one of the greatest challenges facing environmental scientists as they attempt to scale up research to broader spatial and temporal scales. On the one hand, many valuable and relevant data products are not readily available as they are stored on laptops and computers in the offices of individual scientists, projects and institutions. In essence, these data reside in thousands of data silos disconnected from the web, requiring one to learn of their existence through word-of-mouth. On the other hand, a simple search for a particular type of environmental data, such as 'wave height', 'soil carbon', or 'caribou', might result in millions of 'hits', of which only a small fraction are pertinent. The first problem can be addressed as scientists and organizations recognize that data are valuable products of the scientific enterprise and, accordingly, describe, preserve and make those data available for broader use [17,38]. The second challenge is being addressed through projects such as DataONE. These projects support sophisticated, user-friendly search tools that enable scientists to search by time and space and also drill down further using faceted search techniques that allow one to filter the results by parameter, sensor employed, author and other properties of the data, as well as data-subsetting tools for extracting only those data that scientists desire [39]. In addition, the use of controlled vocabularies and community thesauri are useful for assigning key words to data products and can facilitate discovery of desired data [40]. Observational data models and ontologies will be central to achieving even more precise discovery of specific data [41–43]. For example, by annotating observational data to ontology concepts, searches can be automatically expanded to search for related terms (e.g. a search for biomass would also discover variables associated with dry weight, wet weight and other pertinent descriptors).

### *Step 7: integrate*

Collaborative, large-scale synthesis studies in ecology require the integration of data from many disparate studies and disciplines (e.g. population studies, hydrology and meteorology). Integrating source data from such studies is labor intensive and time consuming, because it requires understanding methodological differences, transforming data into a common representation, and manually converting and recoding data to compatible semantics before analysis can begin. Data integration for crosscutting studies is generally a manual process and can consume the majority of time involved in conducting collaborative research [1,11]. Although data integration is challenging, a set of approaches is emerging that explicitly encodes the semantics of observational data and then reasons across these semantics to semi-automate the process of data integration [44,45]. In these approaches, semantic models are built from the bottom up by explicitly capturing the semantics of measurements, which are generally well understood but rarely explicitly captured. Both the Extensible Observations Ontology (OBOE) and the Observations and Measurements specification provide compatible models of data semantics that capture these measurement semantics and that can be used to streamline data integration (Box 1) [41,46]. By contrast, ecologists have traditionally used tools such as Excel to manipulate and convert data manually for integration; however, this process is error-prone and is not reproducible because of the lack of provenance regarding these operations. Scripted analysis environments, such as R and Matlab, improve this by providing a record of data manipulations, but are still largely a record of procedural manipulations of data. Several approaches are emerging that capture a provenance trace that describes the precise derivation of data objects [29,47,48], thereby linking transformation processes to the source and derived data that they produce, and enabling open and reproducible scientific studies.

### *Step 8: analyze*

Ecological systems exhibit high variability and are interconnected in complex ways, thereby stimulating the need for various forms of statistical and geospatial analyses and modeling to distinguish significant ecological processes from background variability. These analytical processes are fundamental to most published results in ecology. Ecologists use a wide variety of programming and general-purpose statistical tools along with a variety of specialty tools and custom built simulation and analytical models to reach conclusions about significant ecological processes (see the supplementary material online) [49–56]. Ironically, these analytical processes are also rarely documented

**Box 1. Ontology-mediated data integration**

Ecological synthesis is labor intensive in large part owing to the difficulty of integrating heterogeneous data for use in analysis. Crosscutting studies typically involve combinations of data spanning multiple disciplines (e.g. ecology, hydrology and atmospheric science) and tremendous variability in the structure and semantics of data within disciplines [73]. Although sensor- and satellite-derived data sets are often fairly uniform, many relevant studies collect data using customized methods and experimental designs meant to address particular hypotheses. Meta-analysis techniques allow for combining even highly specialized experimental studies. For example, meta-analytical techniques have been used effectively to understand cross-system results regarding consumer control of productivity, but data compilation involved manual extraction and coding of data from 191 experiments in 83 individual journal papers [74]. Thus, for synthesis to be more efficient, capabilities for semi-automated data integration need to be massively improved.

One approach to this problem is being developed by the Scientific Observations Network (SONet), a collaboration attempting to define a core model of scientific observations that can be used for advanced data discovery and integration operations. In this model, scientists formally annotate data sets with semantically precise terms drawn from an ontology to specify the semantics of the data. These annotations can then be used to reason about the compatibilities and incompatibilities across source data sets, and then transform those compatible data sets to a common target structure and format to be used in downstream data analysis. The transformations can include unit conversions to a common set of scientific units, alignment and concatenation of semantically compatible variables, downscaling and upscaling to a targeted scale, and calculation of derived values that might be present in one data set but not another.

The example in Figure I illustrates the semi-automated process of integrating two source data sets (b,c) into a common target data set (d) by using an ontology (a) to clarify the semantics of measurements. The ontology follows the OBOE model [41] in modeling Entities (things on which observations are made), Characteristics (the properties of those things that are measured) and Standards (the allowable values for observations). Each of the source data sets makes observations of the size of samples of populations of two salmon species, *Oncorhynchus tshawytscha* (king or chinook salmon) and *Oncorhynchus kisutch* (coho or silver salmon) at various life stages. By annotating both data sets against a common ontology, it is possible to determine that the highlighted cell in source data set 1 is a Number (Standard) representing a PopulationCount (Characteristic) of a KingPopulation (Entity), all indicated by green solid arrows. Source data set 2 is similar, except that the Characteristic is explicitly a SmoltPopulationCount. Because it is known that all SmoltPopulationCount measurements are also PopulationCount measurements, the requisite knowledge is available to transform the data by using this along with annotations of the other columns in the data sets to produce the desired integrated data set (d).

Ontology-driven data integration is among the most promising approaches for streamlining the laborious process of assembling and transforming data in preparation for cross-cutting synthesis studies.

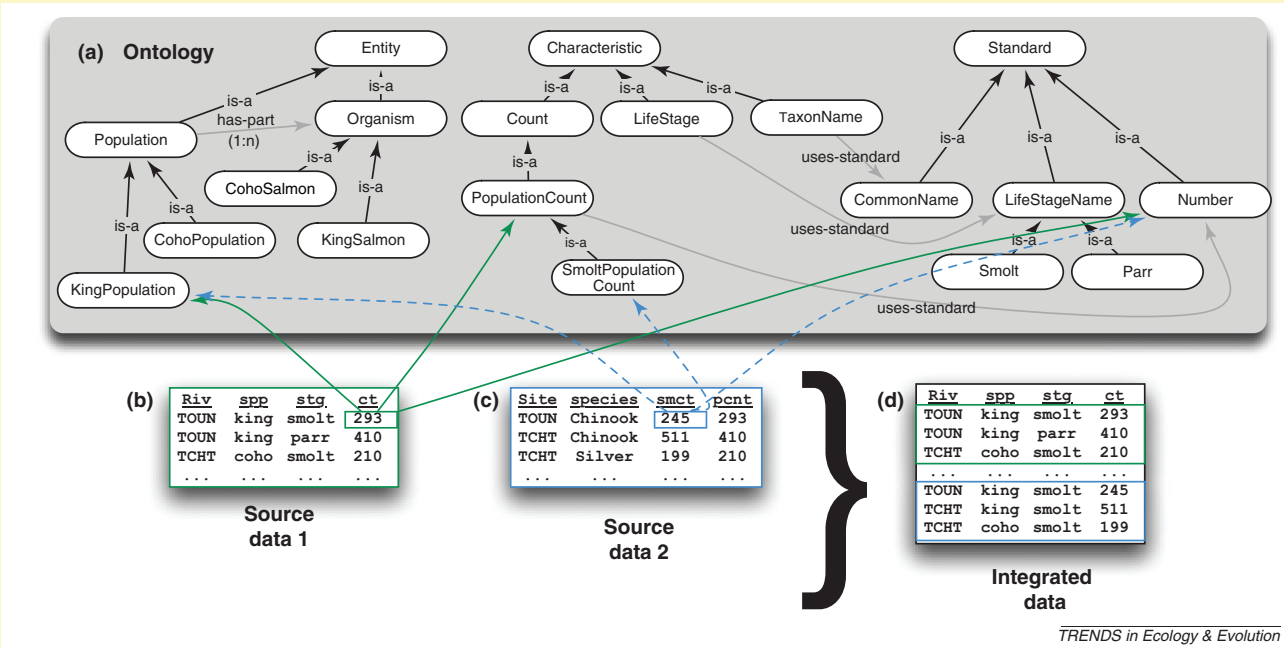


**Figure I.** Ontology-driven integration occurs when semantically explicit scientific terms from an ontology **(a)** are linked to source data sets **(b,c)** and then used in a reasoning process to transform the source data sets to an integrated product **(d)**. In the Extensible Observations Ontology (OBOE) model, linkages between source data explicitly define the Entity being observed, the Characteristic of that Entity that is being measured, and the Standard used to interpret the measured values.

with sufficient detail to enable meaningful reproduction; journal articles typically contain a brief overview that names statistical and modeling approaches but does not capture the details of what was done or how analyses were implemented. Research and development in ecoinformatics focuses on improving this situation through new approaches to documenting the entire set of processes used to reach scientific conclusions [57–64]. Scientific workflow systems, such as Kepler, Taverna, VisTrails and Pegasus [57,59–61,63,64], provide an executable and complete description of analytical procedures that allows scientists to link together processes drawn from multiple different analytical systems. The Kepler scientific workflow system [57] is one such system that has been specialized for use in ecological research (Box 2, Figure I). For example, it includes specific components for accessing data described in the ecological metadata language (EML), has special components for incorporating access to ecological sensor

**Box 2. Adaptive management using Kepler scientific workflows**

Benefits of reproducible science are pronounced in areas where ecological and environmental research are applied to issues of societal importance. Past examples, such as the East Anglia climate research controversy, demonstrate the negative impact that can occur when scientific processes are insufficiently open [75]. In South Africa, water policy has a critical impact on ecological systems in Kruger National Park. Local policy dictates that municipalities upstream of the national park provide sufficient water flow in terms of both total throughput and peak flow events, but traditionally the data for monitoring have been managed in a way that they are unavailable for automated analysis. Park management at Kruger has worked with scientists to establish well-defined Thresholds of Potential Concern (TPCs) that, when exceeded, trigger management evaluation of the state of the system that leads to management action and possibly to changes to monitoring protocols in an iterative adaptive management cycle. Until recently, the data for these TPCs were manually collated and analyses were performed manually in a labor-intensive process. By using a scientific workflow system that captures the complete process of calculating the water flow TPC, provides a mechanism for accessing data from a park-wide repository and periodically executes the workflow on current data, park managers can readily calculate TPCs efficiently and view visualizations of historical performance.

In Figure I, the scientific workflow encapsulates a complex set of analyses and models but communicates the analytical flow in an intuitive way. The workflow can be scheduled to run on an appropriate daily or weekly schedule to produce a visualization that highlights in red the time periods during which the minimum flow threshold has been exceeded. Whenever these conditions occur, park managers can be automatically notified, allowing them to evaluate the reasons for the issue and act accordingly, either by modifying the TPC, the monitoring system, or by contacting municipal authorities about flow issues. In addition, the workflow system provides a complete provenance trace linking the exact versions of data used for the analysis, the analytical procedures executed and the output results, thereby allowing an open and transparent view of the science that is used to guide resource management. Scientific reproducibility as enabled by workflow systems is fundamental to the successful incorporation of science in applied policy and management.

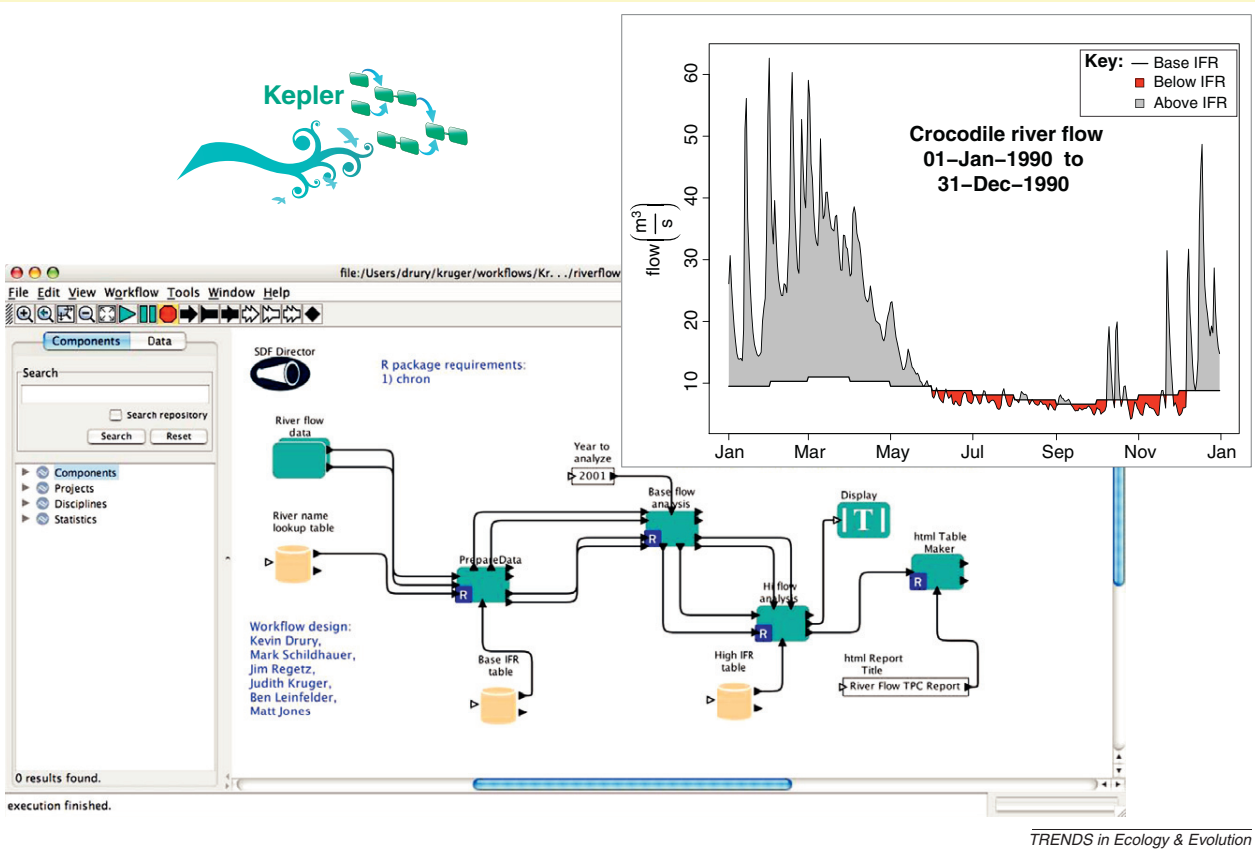


**Figure I.** The Kepler scientific workflow system can be used to capture a complete analytical procedure, even when multiple statistical analysis and modeling systems are in use. This scientific workflow encapsulates all of the processes used to model and visualize water flow characteristics over time, and effectively communicates periods when system thresholds have been exceeded so that adaptive management processes can be employed.

networks, and allows scientists to incorporate commonly used analytical tools, such as R and Matlab [28]. Kepler and other workflow systems also capture provenance information about scientific analyses; each workflow represents a precise record of the processes used in an analysis, and the systems record the provenance of derived products of the analysis, allowing others to understand fully the relationship between data, processing and results, significantly improving the replicability of scientific findings. Ecological models can also be computationally demanding, but ecologists are rarely able to take advantage of advanced computing resources because their models often are constrained by built-in user interfaces and lack the modularity to be incorporated into other execution frameworks. By designing analytical components that can be executed in scientific workflow systems, scientists significantly improve documentation of their processes, tame complexity of the models and enable the models to be

run on powerful distributed computing systems. For example, Kepler includes facilities for easily executing models on pre-existing computing grids, in cloud-computing environments and in ad hoc networks of workflow systems [65,66], while capturing a full provenance trace of the process; and VisTrails is built to generate effectively scientific visualizations while also capturing the provenance of the analysis [61].

## Supporting the full data life cycle

New ground, aerial and satellite-based environmental observing systems coupled with the rapid growth in the use of in situ environmental sensor networks for field research and monitoring, as well as an ever-growing number of citizen-science programs, will soon push ecology and the environmental sciences into a new era where petabytes of data are being collected annually. Powerful informatics platforms will be required to support scientists as they move into this age of data-intensive science. Several such platforms are being designed and built at various scales, including the LTER NIS, the DataONE Federation, LifeWatch, NEON, GLEON and OOI.

The US LTER Network is presently building a network information system that will support synthetic science by: (i) using standardized metadata management and access approaches; (ii) providing middleware programs and workflow solutions that facilitate the creation and maintenance of integrated LTER data sets; and (iii) supporting standardized applications that facilitate discovery, access and use of LTER data [25,67].

DataONE represents a new type of research platform that is specifically designed to support the full data life cycle and to enable new, data-intensive science. It is a federated network providing infrastructure and services for environmental science, enabling new science and knowledge creation through anytime, anywhere access to data about life on Earth and the environment that sustains it [39]. DataONE comprises three principal components. First, DataONE Member Node organizations provide data, computing resources and services, such as data replication. These organizations include data repositories, libraries, universities, research networks, governmental and nongovernmental agencies, computing centers and commercial enterprises. Second, DataONE Coordinating Nodes support network-wide services that enhance interoperability of the Member Nodes and support indexing and replication services. Coordinating Nodes make it easy for scientists to discover data wherever archived, make it easy for data repositories to replicate their data and make Member Node data and services more broadly available to the international community. Third, the DataONE Investigator Toolkit provides tools that are familiar to scientists and that can support them in all aspects of the data life cycle.

LifeWatch is a platform similar to DataONE that is being designed in the EU. NEON [8] and OOI [23] are terrestrial and oceanic observational programs, respectively, that also include integrated informatics infrastructure. We envision that a Federation of such platforms will be needed to support data-intensive, cross-domain research at the biosphere scale. Moreover, the data underlying such research must be openly available and the approaches used in deriving scientific findings must be transparent to ensure that science and society maximally benefit (Box 3).

**Box 3. Open science for society**

Global problems require open access to global data from many disciplines. Such data arise from scientific disciplines that often have very different cultures with respect to data sharing, development and adoption of standards, and practice of good data stewardship. Incentives from research sponsors, societies and institutions (e.g. requiring data management plans) combined with the availability of new informatics tools and platforms, such as DataONE, will be necessary to facilitate data intensive science. Three avenues of research and development offer particular promise: (i) automated provenance-tracking mechanisms that allow scientists to understand and replicate scientific findings fully [76]; (ii) advanced visual analytics that enable scientists to interpret complex, large data volumes more rapidly [68]; and (iii) usability analysis and software engineering support that enable scientists to use advanced ecoinformatics tools more easily.

Tracking the provenance of scientific results is particularly important as advances in environmental science are applied to issues important to society. Open data provide the feedstock on which good science is based, replicable analysis and modeling practices lead to robust findings, and open-access publication disseminates these critical results to the broadest audiences, ensuring the greatest impact of open science for society.

## Remaining challenges

Despite the emergence of ecoinformatics solutions that enable science, several technical and sociocultural challenges and research opportunities remain. First, from the technical side, it is difficult to transport terabyte- and petabyte-sized data sets. Possible solutions include adding computing capabilities to data repositories so that data sets can be processed prior to transport and colocating high-performance computing with large data resources. Second, new visualization approaches and technologies are needed to reduce the time and costs associated with generating visualizations of increasingly large and complex data relationships [68]. Third, little attention has been paid to preserving the algorithms and workflows that scientists use in assuring, analyzing and visualizing data (i.e. activities that support reproducible research) [69–71]. myExperiment was developed as a site where scientific workflows may be stored and shared [72]. However, many standard data center services, such as replication, verification, and migration and conversion to new technologies are more challenging for algorithms and workflows. Finally, despite the promise of semantic technologies, data integration for large-scale studies is still largely manual and time consuming.

Sociocultural challenges can exceed the difficulty of the technical challenges. First, ecoinformatics must be inculcated into mainstream ecological research. Two problems, in particular, need to be addressed: (i) increasing application awareness (i.e. making ecologists aware of the informatics tools and approaches that are available; and (ii) increasing application literacy (i.e. showing ecologists how to use tools properly). Ecoinformatics is, in essence, the 'new statistics' and should be included in undergraduate and graduate curricula, as well as training workshops at professional society meetings. Second, funding agencies can play an important role as they are key stakeholders in the scientific enterprise. Requiring data management

plans is a necessary first step in the short-term, but must be matched by a long-term commitment to sustain data repositories and coordinating organizations [15] and to further encourage data sharing and data stewardship throughout the entire project lifecycle.

## Concluding remarks

In a manner analogous to the transformation undertaken in the physics domain, new environmental observational systems are moving ecology into the realm of big science, whereby scientists and institutions share observation platforms, accumulate and analyze massive amounts of data, and collaborate across institutions to address environmental grand challenge questions. NEON, GLEON, OOI and other observational platforms play a key role in this scientific transformation, much like telescopes, supercolliders, gravitational observatories, and other shared facilities have for physicists. Nevertheless, ecological understanding will, for the foreseeable future, continue to depend upon data collected across a wide range of scales by both individuals and large teams from all countries. New integrative informatics platforms, adoption of standard informatics protocols and good data stewardship practices, as well as sociocultural changes such, as promoting informatics literacy, data sharing, and scientific transparency and reproducibility are central to understanding the nature and pace of ecological and environmental change. The alternatives are for ecological data to remain largely hidden from view in a myriad of disconnected data silos and for ecology to be destined to generate a huge assortment of conclusions from local studies with little way to judge how general or idiosyncratic those scientific findings might be.

### Acknowledgments

This work was supported by National Science Foundation awards #0619060, #0743429, #0722079, #0753138, #0814449, #0830944, #0918635, and the National Center for Ecological Analysis and Synthesis [funded by NSF (Grant #EF-0553768), the University of California, Santa Barbara, and the State of California].

## Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.tree.2011.11.016.

### References

1 Kelling, S. *et al.* (2009) Data-intensive science: a new paradigm for biodiversity studies. *Bioscience* 59, 613–620
2 Hey, T.S. *et al.* (2009) *The Fourth Paradigm*, Microsoft Corporation
3 Jensen, J.R. (2006) *Remote Sensing of the Environment: An Earth Resource Perspective*, Prentice Hall
4 Collins, S.L. *et al.* (2006) New opportunities in ecological sensing using wireless sensor networks. *Front. Ecol. Environ.* 4, 402–407
5 Porter, J.H. *et al.* (2009) New eyes on the world: advanced sensors for ecology. *Bioscience* 59, 385–397
6 Rundel, P.W. *et al.* (2009) Environmental sensor networks in ecological research. *New Phytol.* 182, 589–607
7 Benson, B.J. *et al.* (2010) Perspectives on next-generation technology for environmental sensor networks. *Front. Ecol. Environ.* 8, 193–200
8 Keller, M. *et al.* (2008) A continental strategy for the National Ecological Observatory Network. *Front. Ecol. Environ.* 6, 282–284
9 Kratz, T.K. *et al.* (2006) Toward a global lake ecological observatory network. *Publ. Karelian Inst.* 145, 51–63
10 Fleishman, E. *et al.* (2011) Top 40 priorities for science to inform US conservation and management policy. *Bioscience* 61, 290–300
11 Hackett, E.J. *et al.* (2008) Ecology transformed: the National Center for Ecological Analysis and Synthesis and the changing patterns of ecological research. In *Scientific Collaboration on the Internet* (Olson, G.M. *et al.*, eds), pp. 277–296, MIT Press
12 Peters, D.P.C. *et al.* (2008) Living in an increasingly connected world: a framework for continental-scale environmental science. *Front. Ecol. Environ.* 5, 229–237
13 Michener, W.K. and Waide, R.B. (2008) The evolution of collaboration in ecology: lessons from the United States Long Term Ecological Research Program. In *Scientific Collaboration on the Internet* (Olson, G.M. *et al.*, eds), pp. 297–310, MIT Press
14 Gosz, J.R. *et al.* (2010) Twenty-eight years of the US-LTER program: experience, results, and research questions. In *Long-Term Ecological Research* (Müller, F. *et al.*, eds), pp. 59–74, Springer
15 Reichmann, O.J. *et al.* (2011) Challenges and opportunities of open data in ecology. *Science* 331, 703–705
16 Whitlock, M.C. (2010) Data archiving in ecology and evolution: best practices. *Trends Ecol. Evol.* 26, 61–65
17 Whitlock, M.C. *et al.* (2010) Data archiving. *Am. Nat.* 175, 145–146
18 Heidorn, P.B. (2008) Shedding light on the dark data in the long tail of science. *Libr. Trends* 57, 280–299
19 Borer, E. *et al.* (2009) Some simple guidelines for effective data management. *Bull. Ecol. Soc. Am.* 90, 205–214
20 Cook, R.B. *et al.* (2000) Best practices for preparing ecological data sets to share and archive. *Bull. Ecol. Soc. Am.* 82, 138–141
21 Donnelly, M. *et al.* (2010) DMP online: the Digital Curation Centre's web-based tool for creating, maintaining and exporting data management plans. *Int. J. Digit. Curation* 5, 187–193
22 Strasser, C. *et al.* (2011) DataONE promoting data stewardship through best practices. In *Proceedings of the Environmental Information Management Conference 2011 (EIM 2011)* (Jones, M.B. and Gries, C., eds), pp. 126–131, University of California
23 Cowles, T. *et al.* (2010) The Ocean Observatories Initiative: sustained ocean observing across a range of spatial scales. *Mar. Technol. Soc. J.* 44, 54–64
24 Vanderbilt, K. and Michener, W.K. (2007) Information management standards and strategies for net primary production data. In *Principles and Standards for Measuring Primary Production* (Fahey, T.J. and Knapp, A.K., eds), pp. 12–26, Oxford University Press
25 Michener, W. *et al.* (2011) Long term ecological research and information management. *Ecol. Inform.* 6, 13–24
26 Barseghian, D. *et al.* (2010) Workflows and extensions to the Kepler scientific workflow system to support environmental sensor data access and analysis. *Ecol. Inform.* 5, 3–8
27 Barseghian, D. *et al.* (2011) Sensor lifecycle management using scientific workflows. In *Proceedings of the Environmental Information Management Conference 2011 (EIM 2011)* (Jones, M.B. and Gries, C., eds), pp. 33–38, University of California
28 Gries, C. and Porter, J.H. (2011) Moving from custom scripts with extensive instructions to a workflow system: use of the Kepler workflow engine in environmental information management. In *Proceedings of the Environmental Information Management Conference 2011 (EIM 2011)* (Jones, M.B. and Gries, C., eds), pp. 70–75, University of California
29 Lerner, B. *et al.* (2011) Provenance and quality control in sensor networks. In *Proceedings of the Environmental Information Management Conference 2011 (EIM 2011)* (Jones, M.B. and Gries, C., eds), pp. 98–103, University of California
30 Michener, W. (2006) Meta-information concepts for ecological data management. *Ecol. Inform.* 1, 3–7
31 Fegraus, E.H. *et al.* (2005) Maximizing the value of ecological data with structured metadata: an introduction to ecological metadata language (EML) and principles for metadata creation. *Bull. Ecol. Soc. Am.* 86, 158–168
32 Jones, M.B. *et al.* (2001) Managing scientific metadata. *IEEE Internet Comput.* 5, 59–68
33 Rugge, D.J. (2005) Creating FGDC and NBII Metadata using Metavist 2005, Gen. Tech. Rep. NC-255, US Department of Agriculture
34 Higgins, D. *et al.* (2002) Managing heterogeneous ecological data using Morpho, In *Proceedings of the Fourteenth International Conference on Scientific and Statistical Database Management*, pp. 69–76, IEEE Computer Society

35 Jones, M.B. *et al.* (2006) The new bioinformatics: integrating ecological data from the gene to the biosphere. *Annu. Rev. Ecol. Syst.* 37, 519–544
36 Marcial, L.H. and Hemminger, B.M. (2010) Scientific data repositories on the web: an initial survey. *J. Am. Soc. Inf. Sci. Technol.* 61, 2029–2048
37 Cook, R. (2008) Citations to published data sets. *FluxLetter* 1, 4–5
38 Vision, T. (2010) Open data and the social contract of scientific publishing. *Bioscience* 60, 330–331
39 Michener, W. *et al.* (2011) DataONE: Data Observation Network for Earth – preserving data and enabling innovation in the biological and environmental sciences. *D-Lib Mag.* 17, DOI: 10.1045/january2011-michener
40 Porter, J.H. *et al.* (2011) A controlled vocabulary for LTER data keywords. In *Proceedings of the Environmental Information Management Conference 2011 (EIM 2011), Vol. 103* (Jones, M.B. and Gries, C., eds), pp. 18–169, University of California
41 Madin, J. *et al.* (2007) An ontology for describing and synthesizing ecological observation data. *Int. J. Ecol. Inform.* 2, 279–296
42 Madin, J.S. *et al.* (2008) Advancing ecological research with ontologies. *Trends Ecol. Evol.* 23, 159–168
43 Berkley, C. *et al.* (2009) Improving data discovery in metadata repositories through semantic search, In *Proceedings of iSEEK'09*, pp. 1152–1159, IEEE Computer Society
44 Leinfelder, B. *et al.* (2011) Using semantic metadata for discovery and integration of heterogeneous ecological data. In *Proceedings of the Environmental Information Management Conference (EIM 2011)* (Jones, M.B. and Gries, C., eds), pp. 92–97, University of California
45 Buccella, A. *et al.* (2009) Ontology-driven geographic information integration: a survey of current approaches. *Comput. Geosci.* 35, 710–723
46 Reed, C. *et al.* (2007) OGC® sensor web enablement: overview and high level architecture. *IEEE Autotestcon* 372–380
47 Missier, P. *et al.* (2010) Linking multiple workflow provenance traces for interoperable collaborative science, In *Proceedings of the 5th Workshop on Workflows in Support of Large-Scale Science (WORKS)*, pp. 1–8, IEEE
48 Osterweil, L.J. *et al.* (2010) Clear and precise specification of ecological data management processes and dataset provenance. *IEEE Trans. Automation Sci. Eng.* 7, 189–195
49 Kernighan, B.W. and Ritchie, D.M. (1988) *The C Programming Language*, (2nd edn), Prentice Hall
50 Stroustrup, B. (1997) *The C++ Programming Language*, Addison-Wesley Professional
51 Adams, J.C. *et al.* (2009) *The Fortran 2003 Handbook*, Springer
52 Schwartz, R.L. *et al.* (2011) *Learning Perl*, O'Reilly Media
53 Lutz, M. (2009) *Learning Python*, O'Reilly Media
54 Hanselman, D. and Littlefield, B. (2004) *Mastering MATLAB 7*, Prentice Hall
55 Crawley, M.J. (2007) *The R Book*, Wiley
56 Elliott, A.C. and Woodward, W.A. (2010) *SAS Essentials: A Guide to Mastering SAS for Research*, John Wiley and Sons
57 Ludäscher, B. *et al.* (2006) Scientific workflow management and the Kepler system. *Special Issue: Workflow Grid Syst. Concurrency and Comput.: Pract. Exp.* 18, 1039–1065
58 De Roure, D. *et al.* (2009) The design and realisation of the myExperiment virtual research environment for social sharing of workflows. *Future Gen. Comput. Syst.* 25, 561–567
59 Deelman, E. *et al.* (2005) Pegasus: a framework for mapping complex scientific workflows onto distributed systems. *Sci. Program. J.* 13, 219–237
60 Hull, K. *et al.* (2006) Taverna: a tool for building and running workflows of services. *Nucleic Acids Res.* 34, 729–732
61 Silva, C.T. *et al.* (2007) Provenance for visualizations: reproducibility and beyond. *Comput. Sci. Eng.* 9, 82–90
62 Ellison, A.M. *et al.* (2006) Analytic webs support the synthesis of ecological data sets. *Ecology* 87, 1345–1358
63 Ludäscher, B. *et al.* (2009) Scientific process automation and workflow management. In *Scientific Data Management: Challenges, Technology, and Deployment* (Shoshani, A. and Rotem, D., eds), pp. 467–509, Chapman & Hall/CRC
64 Taylor, I.J. *et al.* (2007) *Workflows for e-Science: Scientific Workflows for Grids*, Springer
65 Wang, J. *et al.* (2009) Accelerating parameter sweep workows by utilizing ad-hoc network computing resources: an ecological example, In *2009 World Conference on Services I*, pp. 267–274, IEEE
66 Wang, J. *et al.* (2009) Kepler + Hadoop: a general architecture facilitating data-intensive applications in scientific workflow systems, In *Proceedings of the 4th Workshop on Workflows in Support of Large-Scale Science (WORKS09) at Supercomputing 2009 (SC2009) Conference*, ACM DOI: 10.1145/1645164.1645176
67 Servilla, M. *et al.* (2006) PASTA: a network-level architecture design for generating synthetic data products in the LTER network. *LTER Databits* Fall
68 Fox, P. and Hendler, J. (2011) Changing the equation on scientific data visualization. *Science* 331, 705–708
69 Peng, R.D. *et al.* (2006) Reproducible epidemiologic research. *Am. J. Epidemiol.* 163, 783–789
70 Hollister, J.W. and Walker, H.A. (2007) Beyond data: reproducible research in ecology and environmental science. *Front. Ecol. Environ.* 5, 11–12
71 Cassey, P. and Blackburn, T.M. (2006) Reproducibility and repeatability in ecology. *Bioscience* 56, 958–959
72 Goble, C.A. *et al.* (2010) myExperiment: a repository and social network for the sharing of bioinformatics workflows. *Nucleic Acids Res.* 38 (Suppl 2), W677–W682
73 Halpern *et al.* (2008) A global map of human impact on marine ecosystems. *Science* 319, 948–952
74 Gruner, D.S. *et al.* (2008) A cross-system synthesis of consumer and nutrient resource control on producer biomass. *Ecol. Lett.* 11, 740–755
75 Reay, D.S. (2010) Lessons from Climategate. *Nature* 467, 2010
76 Bowers, S. *et al.* (2008) Provenance in collection-oriented scientific workflows. *Concurrency Comput.: Pract. Exp.* 20, 519–529