

# **SIGNIFICANCE of ORGANISM OBSERVATIONS**

## **DATA DISCOVERY and ACCESS in BIODIVERSITY RESEARCH**

Steve Kelling  
Director of Information Sciences  
Cornell Lab of Ornithology



Copyright 2008, Global Biodiversity Information Facility (GBIF)

Significance of organism observations: Data discovery and access in biodiversity research  
ISBN: 87-92020-08-9

Permission to copy and/or distribute all or part of the information contained herein is granted, provided that such copies carry due attribution to the Global Biodiversity Information Facility (GBIF).

Recommended citation format:

Kelling, S. 2008. *Significance of organism observations: Data discovery and access in biodiversity research*. Report for the Global Biodiversity Information Facility, Copenhagen.

## **Contents**

Introduction	1
The Resource -- Observational Data	3
What are observational data?	3
How are observational data on organisms generated?	3
Protocols	3
Types of surveys: Directed, broad-scale and specimen-based	5
The Discovery and Organisation of Observations	6
Discovery: Metadata	6
Organisation: Schemas	8
Existing schemas for biodiversity specimen and observation data	8
Recommended additions to existing schemas for survey data	9
Future directions: Data ontologies	11
Conclusions	13
Literature Cited	14

## Introduction

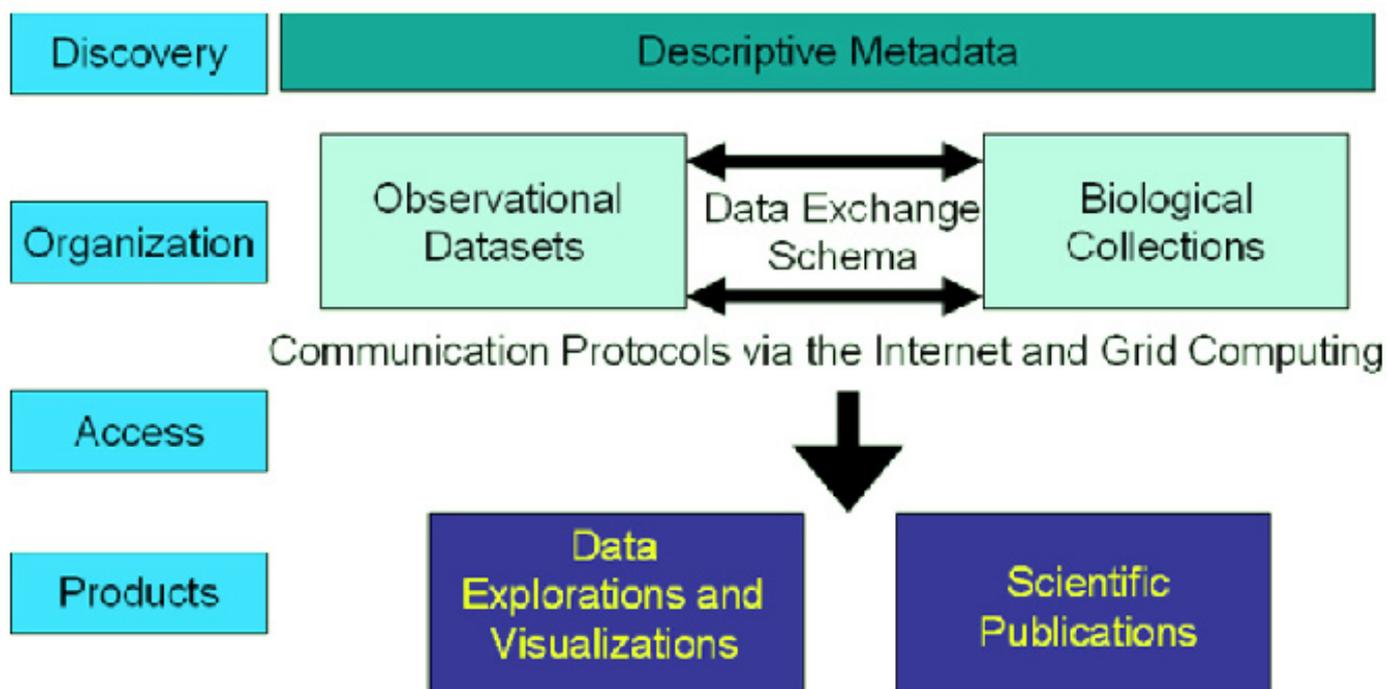
Observations of nature are the foundation of ecological studies, which use observations to search for patterns in nature (MacArthur 1972), and biodiversity conservation (Wiens 1992). They provide the unifying concept for most scientific publications, the basis for biodiversity management plans, and the resource to create sound conservation policy. The need for observational data sharing and interoperability is recognized by the Convention of Biological Diversity as one of the preconditions for improved global biodiversity conservation (UNEP 2005), because the investigation of complex ecological and environmental issues at broad geographic or temporal scales requires the integration of data from multiple research efforts (Andelman et al. 2004, Ellison 2006). Such synthetic analyses must rely on the effective discovery and processing of data from many independent projects (e.g., uncoordinated studies that often focus on restricted thematic issues and spatiotemporal scales). However, there are many challenges to be faced in gaining access to and analysing these data for biodiversity studies because they are stored in highly heterogeneous data structures, are often poorly managed, and a comprehensive organisational structure is lacking (Laihonen and Kalliola 2004).

Attempts to organise observational data and make them available are underway through a variety of national, international, and domain-specific initiatives. For example, the Biodiversity Information Standards Working Group (TDWG) is developing standards for information management, integration, and access, and the Global Biodiversity Information Facility (GBIF) is implementing these standards to federate biological collections into a single global data management structure for improved access.

The capability to organise the rich and varied data resources of the ecological and environmental communities has emerged only recently, primarily because of major advances in four related technologies, which have transformed how we access and manipulate information (Friedman 2005):

1. Computers have become ubiquitous because of rapid advances in computing power coupled with decreasing unit production costs (Moore's Law).
2. The Internet and web browsers have engendered global standards for passing information among computers.
3. Pervasive installation of fiber optic cabling has globalised computer networks.
4. Developments in information description languages, data management processes, and software application integration have created seamless workflows for access, manipulation, and processing of almost limitless data resources.

For these technological advances to transform biodiversity information resource discovery, access and scientific analysis, several interrelated processes must be implemented (Figure 1). First, the *resources* must be identified, catalogued, and made *discoverable* (that is, potential users must be able to know that these resources exist). This is accomplished through metadata that explicitly describes the data resources or information products (visualisations and reports) that comprise the resource. Third, the data resources must be *organised* into an interoperable format. Fourth, the infrastructure must be implemented to *access* the resources via communications protocols between networked computers and the Internet applications that provide information access. Finally, *information products* must be developed that utilise these organised data resources. Such products integrate data resources (e.g., observational data with land cover, human demographic or climatic variables) to produce visualisations, via maps, graphs, and tables, scientific and technical publications or the like.



**Figure 1: The discovery and organisation of biodiversity information for accessibility and use.**

The objectives of this paper are to

- formally identify the subset of observational data focused on species occurrences, and
- describe the special opportunities offered to ecological understanding and biodiversity conservation by the special characteristics of these observational datasets on organisms.

While many biodiversity information initiatives have focused on data from specimen collections (a special subset of observation data), this paper focuses on broadening the scope of these endeavors to include all observational data. Then, the paper provides an overview of current efforts to organise and provide access to these data, and makes recommendations for better community-wide resource integration.

## The Resource - Observational Data

### What are observations of species occurrences?

All observations share similar thematic components, and these provide the foundation for the organization and synthesis of biodiversity data. An observation of an organism is some measurement of a particular attribute of an organism made within a particular context. The context is the set of conditions under which the measurements were made, such as the time, date, and location of the observation. The attributes collected could include the size, age, sex, behavior, or the number of organisms. The measurement describes the process in which the data were collected. Observations of organisms can be gathered through myriad mechanisms that are as different as molecular biology techniques that differentiate genotypes of virus strains and a citizen scientist recording birds that visit a backyard feeder, but all resulting observational data are fundamentally similar.

The vast majority of organism occurrence observations are made by human observers (Fig. 2). This is because of the difficulty in training autonomous sensors (i.e. a measuring instrument which converts a physical quantity into an electrical signal) to identify organisms. While inroads have been made in the use of autonomous sensors in ecology (Hamilton et al. 2007), they primarily serve as the source of information on the variables that influence species occurrence, and little information on actual species occurrence (Hochachka 2007).

Observations are being gathered at an enormous rate, and their numbers are growing exponentially. For example, the volume of data stored in biological collections worldwide is estimated to approach a billion specimens (Beaman et al. 2004), and significantly more are available as species occurrence records gathered through survey and monitoring techniques (Sarvala et al. 2005).

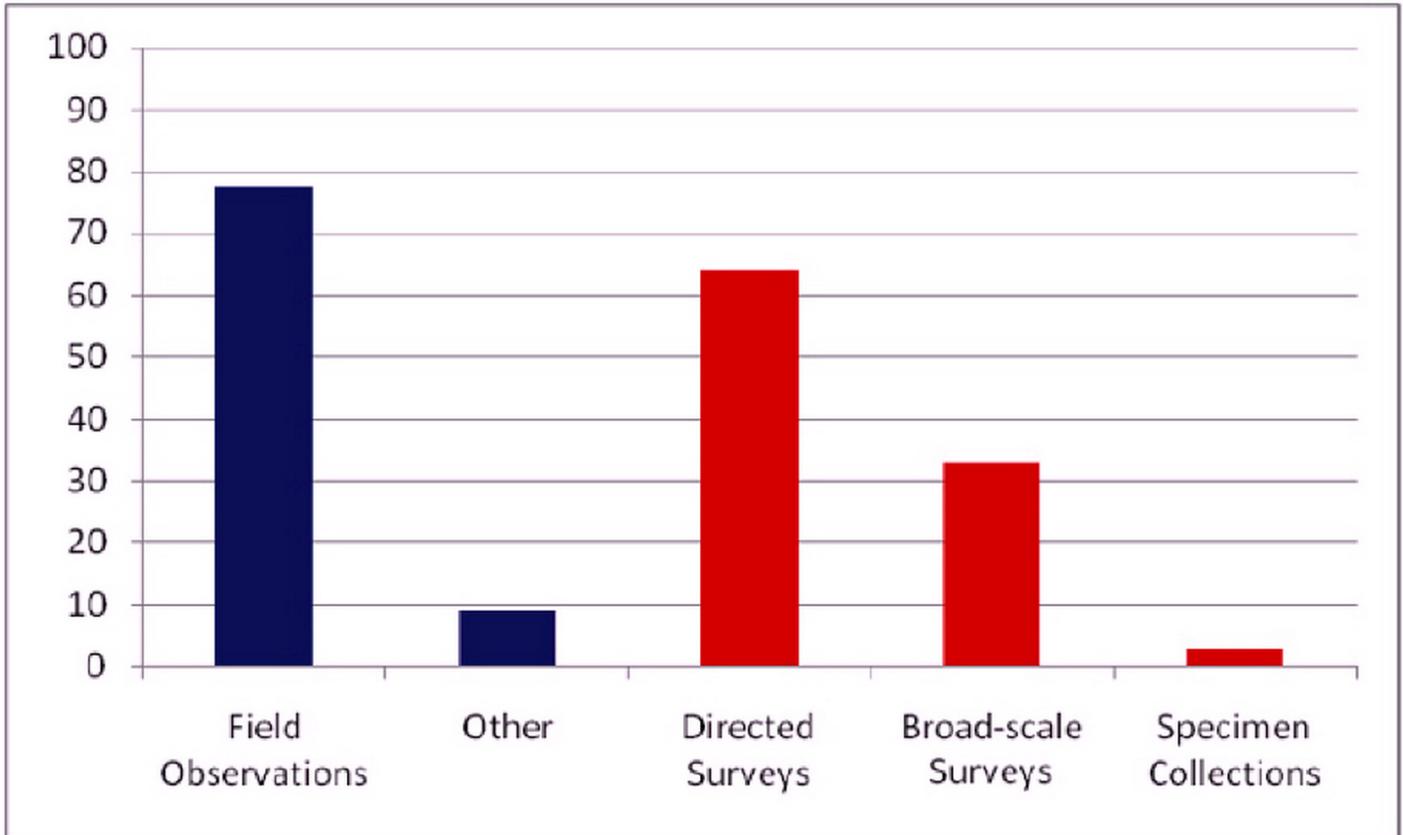
Occurrence observations are gathered and organized within 3 broad classes: directed surveys, broad-scale surveys, and biological collections (Fig. 2). While each class of observation shows some overlap in data gathered, each is distinct. For example, all observations of organisms contain information on the species and location it was observed. But, natural history collections document this occurrence through physical specimens, while surveys and monitoring approaches only maintain a data record that an organism or assemblage of organisms was observed. Thus, for observational data to be properly used it is essential to understand the opportunities and limitations that each of these categories of data. In the next section the processes that are used to gather observations and descriptions of the three categories are discussed.

### How are observational data on organisms generated?

#### Protocols

The data context that facilitates the combination of observations by defining (either implicitly or explicitly) the inferential population described by the data is the protocol under which the data were collected. Protocols define a formal design or action plan for gathering attributes of an entity (University of Washington Health Services 2000) and in so doing facilitate the combination of observations made by multiple participants in many locations. Protocols are important for understanding how observational data can be combined and analyzed, and can be classified into three general data gathering types (Kutner and Stein 2006).

- *Place-based protocols* characterize a defined locality and typically result in a checklist and measure of abundance of the species observed.



**FIGURE 2.** The data sources for contributed papers in the journal *Conservation Biology*, which is published by the Society for Conservation Biology. Over 300 papers were reviewed and dichotomously categorized. The primary data sources (blue columns) for 78% of the papers were field observations of organisms made by humans and 22% of the papers used molecular biology, human demographic, autonomous sensor, or geospatial data sources. Field observations were further divided (red columns) into directed survey techniques (60% of the field observations), broad-scale surveys (35% of the observations), or gathered or used specimens (5% of the observations) [see text for definitions]. Note journals that publish on specific taxon (i.e. *The Auk*, which is published by the American Ornithologists Union) have appreciably more papers published using specimen data, while those focused on ecology (i.e. *Ecology*, which is published by the Ecological Society of America) are almost entirely based on directed survey data.

- *Taxon-based protocols* record the occurrence of a particular species or ecological element at a location.
- *Monitoring protocols* can follow either place-based or taxon-based approaches but are unique in that the data are gathered on repeated visits over a period of time.

The data collection protocol chosen for a project places constraints on the methods of data organisation and analysis that may be used. For example, if data are gathered using similar protocols, they can be aggregated and analysis conducted on the combined dataset. But, if different protocols are used, analysis is difficult because biases that could not or were not controlled for during data collection and synthesis must be accounted for. Therefore, data sets should be accompanied by information about the factors that affected data collection and aggregation (metadata). If this information is sufficiently detailed, potential sources of bias may be investigated and ameliorated during analysis (Kelling et al. 2008, submitted). Also, many survey protocols require the observer to report all organisms detected (e.g., most bird surveys require reporting of all birds that were identified). Thus, it can be inferred that an unreported organism was in fact not present (or at least not detected). This has significance in data analysis because locations at which effort was made but the organism was not detected can be distinguished from locations that were not sampled.

## Types of surveys: Directed, broad-scale and specimen-based

**Directed surveys** are often used in ecological studies where *a priori* knowledge of a given system or biological mechanism already exists. The experimental design of such surveys attempts to control for known sources of variation, or to sample well defined inferential populations. As such, directed surveys are the form of observational data collection that closest resembles experimental studies. These studies seek to establish the causal relationship between some experimental treatment and its effect, describe certain characteristics of a well defined target population, or a combination of the two such as making causal inferences on a target population (Nichols and Williams 2006). With these data researchers can draw strong inferences via multiple competing hypotheses leading to refined future analyses (Platt 1964). For example, Maschinski and co-workers (2006) developed several competing hypotheses to predict the impact of global warming on the Arizona Cliffrose (*Purshia subintegra*), a federally endangered arid plant species. She found that fine spatial-scale modelling is necessary to accurately differentiate sites having a high risk of extinction with those that could serve as potential refugia. The survey techniques used required deep understanding of the causal impact of changing climate on the cliffrose, and a clear enumeration of the characteristics that should be measured or summarized. This prior knowledge leads to a well defined statistical / analytical framework to make inferences, and quantify uncertainties.

The directed survey approach has been widely accepted in the scientific community—in part because of the formal mathematical justification these models impart. But, this level of data gathering and analytical rigor is expensive, managed by researchers working independently, and conducted on small spatial and temporal scales. This creates a network of heterogeneous data repositories with little opportunity for integration or reuse and eventual data degradation and loss (Michener 2006). Data collected through directed surveys are most susceptible to loss, and require a great deal of data curation attention to prevent loss or degradation.

**Broad-scale surveys** generate probabilistic estimates of species occurrence that can be used to elucidate patterns over broad geographic and temporal scales (Ralph et al. 1995). While these techniques do not provide direct evidence for the causes of species occurrence, they do identify priorities for more targeted directed surveys, and to prioritize species for conservation action (Van Horne et al. 2007). For example, researchers in North America studying the spread of the bacterial pathogen *Mycoplasma gallisepticum*, which causes severe conjunctivitis in wild House Finches (*Carpodacus mexicanus*), have used a continental scale network of contributors to monitor the disease. First detected in 1994 (Luttrell M.P. 1996), the disease killed an estimated 60% of eastern North American House Finches within three years of disease emergence (Hochachka 2000). Over a 14-year period, the study has engaged thousands of volunteer-observers who feed and watch birds in their yards to report birds with conjunctivitis (Dhondt 2001). This observer network has documented the expansion of the epidemic throughout most of the House Finches' range (Fischer 1997, Ley 1997, Dhondt 1998), and the results provide the basis for theoretical models and intensive and experimental studies designed to understand host-disease dynamics.

Broad-scale surveys engage a diversity of participants that range from trained observers to interested citizens, and currently gather tens of millions of observations annually. Often these projects gather data opportunistically in an effort to collect as much information as possible. Broad-scale survey projects are inexpensive to operate, as they rely on volunteer participation, covering huge regions over long periods of time. Presently, these studies provide the bulk of non-specimen observational data available via the Internet (GBIF 2006). However, because the data collection protocols used for many broad scale surveys are less stringent than directed surveys, they are subject to more potential

biases. For example, broad-scale surveys tend to have geographically patchy coverage (e.g. along roadsides, near human population centers), and primarily survey charismatic fauna (i.e. there are many more projects that collect data on birds than there are that collect data on isopods). Finally, using these data to make statistical inferences about the target population requires careful analysis to control for known biases.

**Specimen-based surveys** result in zoological, botanical, and paleontological collections that are housed in museums, living collections in botanical or zoological gardens, or microbial strain and tissue collections (Berendsohn 2007). They provide the foundation for taxonomic and historic occurrence of species, which are documented through physical specimens (Chapman 2005). While most use of specimen collections has been for taxon-oriented research, they are now being used for a diverse set of questions unrelated to their establishment (Winker 2004). For example, Becker and Beissinger (2006) used tissue samples from specimens collected over a 100 year period to show that the endangered Marbled Murrelet (*Brachyramphus marmoratus*), which occurs along the Pacific seacoast of the northwestern United States and Canada, has shifted to feeding on prey sources with lower total energy content. Murrelets fed primarily on the high-energy food source (needed for egg production) Pacific Sardine (*Sardinops sadax*) prior to the collapse of the fishery during the 1940s. With this collapse murrelets were forced to shift their primary food source to less energy rich food sources such as krill. This shift has been implicated in the decline in Marbled Murrelet populations.

The data from biological collections, taken together, comprehensively cover all known organisms, and the taxon identifications are made by trained personnel. However, these are presence-only data, and often reflect significant and/or undocumented bias in sampling. There are often large geographic or temporal gaps in these data, which result from shifts in collecting effort or emphasis and that create an uneven record of species distribution (Ponder et al. 2001). Furthermore, while the majority of specimens in collections are associated with a location, this is often represented by a geospatially imprecise textual description (Beaman et al. 2004). Although the value of biological collections is immense, particularly with regard to documentation of historic occurrences of species, it seems that their future role will diminish because of apparent declines in collecting effort (Winker 1996, Cooper and Steinheimer 2003).

## The Discovery and Organisation of Observations

### Discovery: Metadata

To access biodiversity data sets, users must first know what is available, and how the data can be accessed. Just as library catalogs detail information about a book and how to discover that book in the library, in the online information environment metadata describe data resources and their accessibility.

Metadata provide information on the identification, quality, spatial context, data attributes, and distribution of datasets, using a common terminology and set of definitions that prevent loss of the original meaning and value of the resource. This common terminology is particularly important to biodiversity datasets, because different biodiversity projects collect dissimilar types of data and record them in various ways, occur at a variety of scales, and are dispersed globally. Without descriptive metadata, discovering that a resource exists, what data were collected and how they were measured and recorded, and how to access it would be a monumental undertaking.

Metadata in the biodiversity information domain provide:

- An accurate description of the data themselves;
- A description of spatial attributes, which should include bounding coordinates for the specific project, how spatial data were gathered, limits of coverage, and how these spatial data are stored;
- A complete description of the taxonomic system used by the project, with references to methods employed for organism identification; and
- A description of the data structure, with details of how to access the data and/or how to access tools that can manipulate the data (i.e. visualisations, statistical processes, and modelling).

Several initiatives are underway that are developing discovery resources for biodiversity data and monitoring programmes. These initiatives can be identified as open-ended (encompassing all biodiversity resources), or domain specific (only organising the resources within a specific area of interest), and their foci range from description of data generated by monitoring programmes to description of the projects or programmes themselves. What follows are four examples of initiatives that are attempting to facilitate biodiversity observational data resource discovery:

### *1. Biological Data Profile and the Metadata Clearinghouse*

The Biological Data Working Group, working under the auspices of the United States' Federal Geographic Data Committee (FGDC), developed a user-defined, theme-specific profile for describing biological data with the purpose of increasing compatibility in the development, use, sharing, and dissemination of these data (<http://biology.usgs.gov/fgdc.bio/charter.html>). The goal of the Biological Data Profile is to describe datasets that result from data collection efforts. It defines all information required by a user to determine what variables are stored in the dataset, the data quality, and how to access the data (FGDC 2003). The Biological Data Profile employs a series of interrelated metadata elements, which together provide a very detailed description of the dataset's contents. The value of this approach is in the detail of the resulting dataset description. However, this description is often text heavy and very complex, often making it difficult to decipher and thus to compare projects.

### *2. Ecological Markup Language (EML)*

EML was developed by the ecological community to provide a common structure to allow ecologists to discover ecological data (<http://knb.ecoinformatics.org/software/eml/>). EML metadata is based on FGDC standards, and its descriptors are organised into classes that describe the dataset and its research origin, data structure, status and accessibility. The goal of EML is to provide sufficient information for a researcher to be able to use the data in a scientifically correct manner. Consequently, its metadata categories are very deep, and metadata descriptions can be constructed on the dataset, entity, or attribute level. Web-based tools are available for uploading project metadata descriptors. EML serves as an XML schema for documenting and organising ecological data in a standard format for data sharing. Access to EML metadata is through the Knowledge Network of Biocomplexity (<http://knb.ecoinformatics.org>). The project is global in extent, and EML has been applied to the results of 1,500 projects from all continents that have gathered more than 65 billion observations of all types (i.e. observations of organisms, climatic, weather, or chemical flux) (Matt Jones, personal communication).

### 3. *The National Biodiversity Network of the United Kingdom (NBN)*

The NBN is a web-based tool that facilitates discovery and access to biodiversity data across the United Kingdom (<http://www.nbn.org.uk/>). It is a national standard for data description and data sharing. The NBN Gateway provides web enabled tools to allow users to discover and access these data. Fundamentally, the NBN Gateway provides 1) sufficient metadata to allow users to assess the scope and potential uses of a project's data, and 2) a data warehouse, which provides a standard data format to access, when permitted, the project's data. The goal of NBN metadata is to provide the minimum contextual information to enable a user to potentially use a data source. The metadata describe and document geospatially referenced data sets and provide information on dataset ownership, methods and scale of data collection, and potential limitations of interpretation. The metadata standard is based on the FGDC metadata elements with minimal modifications that make it more functional within the United Kingdom.

### 4. *North American Bird Monitoring Projects Database*

An example of ongoing regional or taxon-specific metadata efforts is the North American Bird Monitoring Projects Database. It contains information on several hundred bird monitoring projects conducted in Mexico, the United States and Canada (<http://www.bsc-eoc.org/nabm>). The database contains programme descriptions and contact information (which can be submitted by volunteers) with the goal being to facilitate the development of new projects. While the North American Bird Monitoring Projects Database and other similar initiatives provide invaluable information, they often use unique domain-specific nomenclatures and formats, which make it difficult to integrate their data across domains.

## **Organisation: Schemas**

Once a project has been “discovered” the next step is to determine means of access to the data. This is challenging, because projects that gather observational data are maintained by a variety of institutions that are dispersed around the world and so their data are stored in various architectures. Maximizing efficient use of observational data for research and analysis requires across-site, interdisciplinary mechanisms to synthesize these disparate resources into a unified entity – that is, the databases must be made *interoperable* (Andelman et al. 2004).

### **Existing schemas for biodiversity specimen and observation data**

Efforts are underway within the observational data community to achieve interoperability of their databases by following standardized data formats. The goal of the community is to facilitate interoperability not only among its own datasets but also with existing metadata standards, external portals and data harvesting structures. Currently, data exchange schemas are used to make data resources interoperable by transforming disparately structured source data onto a standardized target schema (Phokion et al. 2006). Data exchange schemas have been successfully used to organise tens of millions of observations of organisms. In particular, the data exchange schemas known as Access to Biological Collections Data (ABCD), and Darwin Core (DwC), have made important first steps in improving our ability to access biodiversity data. GBIF's index data cache organises observational data that are provided by an ever-growing multitude of sources primarily with DwC, but also includes specific elements of ABCD.

1. *The Darwin Core Schema (<http://www.tdwg.org/activities/darwincore/>)*

The DwC includes a simple set of data element definitions designed to support the sharing and integration of primary biodiversity data. While initially developed to organise specimen collections, it is extensible (additional data elements can be added), and a number of groups have expanded DwC to serve their specific requirements. For example, the Avian Knowledge Network (<http://www.avianknowledge.net>) has extended DwC to allow more complete integration of observational data made on bird populations. The result has been the organisation of over 46 major observational datasets that include more than 51 million records. Regardless of the differences between specimen collection records and observational data sets, there is a commonality in their content that may be exploited to perform ordered search and retrieval from these diverse data sets. The DwC attempts to provide guidelines for utilising this commonality regardless of the underlying mechanism for storing the record content.

2. *The ABCD Schema (<http://www.tdwg.org/activities/abcd/>)*

ABCD is a hierarchical data specification schema developed to support the exchange of biological collections data that include specimens as well as field observations. ABCD is comprehensive, and therefore complex; it uses nearly 1200 concepts. ABCD is highly structured and comprehensive and aimed to provide all of the required variables for the incorporation of any collection (both specimen and observation). The schema is able to integrate data from a variety of sources and allows the inclusion of data that are variously detailed and domain-specific. By taking this full-coverage approach, the goal of ABCD is a complete set of descriptors for natural history collections of any type. There are sufficient variables in ABCD to make it as compatible as possible with other standards, such as DwC. ABCD has been ratified as a TDWG standard and is promoted by GBIF for use globally.

### **Recommended additions to existing schemas for survey data**

The development of DwC and the ABCD was focused on organising specimen collection data, and recent attempts to integrate data from directed survey or broad-scale survey projects has revealed deficiencies that require additional fields (DwC) or fields elevated to a higher level of the organisational hierarchy (ABCD). For example, sufficient variables must be included in a data schema to identify data-gathering protocols, to incorporate both presence and absence data, to deal with multiple organisms observed during single data-collecting events, and other features. These new variables include:

*Collection Event* The data collection event must be hierarchically above that of the individual species record. This is because most observational data are gathered during collecting events in which information on multiple species and individuals is collected. EML provides sufficient variables to fully describe the collection event, and because DwC is a flat schema and extensible, collection event attributes can be added. ABCD is a hierarchical schema and the collection event is at a lower level in the hierarchy making it difficult to adapt in this regard.

*Protocols* Data schemas used with observational data must allow for the description of the protocols used during data gathering. All observational data contain the same core components: species observed, number counted, location, observer, and date, and these are provided by the schemas under discussion. In addition, elements that describe how the data were gathered must be added because of the variety of data collecting protocols, each of which is designed for the purposes

of collecting a specific type of data. Consequently, any data schema used for observational data must contain sufficient object classes to properly describe the methodology used to amass a dataset.

*Effort* The level of effort required to collect observational data varies based on the goals of the project and the protocols implemented. For example, while most such data are gathered using place- or taxon-based or monitoring protocols (Kutner and Stein 2006), the effort (i.e., time spent or distance travelled) can vary and description of this must be accommodated by the schema. For many biodiversity studies, effort can be categorised in one of the following four ways (J. Bart, personal communication):

1. **Occurrence:** Record when and where an organism was observed. This provides the foundation on which all observational data of organisms can be combined, regardless of whether the data were gathered using broad scale surveys, directed surveys, or by collecting specimens.
2. **Checklist:** Record a list of organisms observed (often including how many individuals of each species were present) at a particular place and time. Most checklist surveys record how much effort (time spent collecting data, distance traveled, or area searched) was expended and whether all the organisms identified were reported.
3. **Repeated Sampling:** Record when and where an organism was observed and how much effort (time, and distance or area covered in order to observe the organism) was invested and how often the effort was made. A well-defined survey protocol is followed and repeated sampling of an area is encouraged.
4. **Formal Sampling-plan:** Record when and where organisms were observed and how much effort (time, distance or area) was invested. A well-defined survey protocol is followed in conducting the surveys and repeated samples of the location are made. The protocol includes a specific description of the data to be collected, the area in which it will be collected, the interval between data collection events, and the specific organisms on which data will be collected.

*Absence Data:* A significant portion of observational data is gathered using either place-based or monitoring protocols (see above) that provide a list of all species and counts of individuals observed, allowing the inference that if a species was not reported it did not occur at the location. While there are caveats (e.g., observer ability is highly variable, and detectability varies from species to species as well as seasonally), observational studies provide data both on where an organism did occur and where it did not. Such absence data are extremely valuable in estimating species distributions and abundance. Unfortunately, ABCD and DwC do not easily handle absence data, and additional work is required to obtain a data set that includes both absence and presence. For example, with DwC, two queries are needed: one for the positive observations and another for all locations with observations within the data and location range.

*Taxonomic:* While organisations that manage observational data take great care in the identification and classification of the organisms being studied, they very often store their data following a taxonomic hierarchy unique to their own institution. Federating these unique taxonomic hierarchies and forcing all data to converge on a single hierarchy is not practical. Therefore, explicit concepts for taxonomic structures have been implemented within the data exchange schemas.

Some projects collect observational data that are not or cannot be identified taxonomically. Additionally, certain projects may use functional groupings or other ecological classifications for their identifications. Consequently, any effort to organise observational data into a unified standard must be able to classify what was observed either as taxonomic rank higher than species, or as a functional grouping, life-history stage, or other ecological classification. Additionally, in ecological datasets, species assemblages (instead of a single species) are often the basis of the identification. This has a major impact on biodiversity informatics because it requires the development of methodologies that accommodate not only a unitary entity such as a specimen as the basis of a record, but also bases of record that are pluralistic entities, such as ecological communities.

*Spatial Attributes:* Properly georeferenced observational data are fundamental to describing and analysing the distribution of organisms and relating that distribution to environmental variables that impact their occurrence. Such georeferencing allows for both repeated sampling at the identical location and the integration of other data types (e.g. land cover or other abiotic data) for the same location. But, because biodiversity data are gathered using a wide variety of techniques, there is much variability in the quality of the location information: Examples include:

- many broad scale survey approaches gather location information with little precision (i.e. postal codes or political boundaries); and
- many observation data protocols (e.g., the United States Breeding Bird Survey) require observers to make their observations along transects, some of which are many tens of miles long, and some such projects only maintain information only at the transect level.

Differences such as these in location precision lead to much across-project disparity in the accuracy of the georeferenced information being gathered. Thus the data schema must incorporate an estimate of the level of uncertainty of a project's location information.

### **Future directions: Data ontologies**

On the whole, biodiversity data have been collected and stored in data formats that are unique to the project that has collected the data. The use of data exchange schemas has been an important first step in the gradual improvement of access to biodiversity data. Nonetheless, the organisational structure of exchange schemas is inflexible, and it requires that data be transformed from their source format to the target schema, which leads to potential loss of domain-specific content. This is because data exchange schemas use simple data concepts and store these in static organisational taxonomies.

Efforts are under way to develop alternative approaches to improve project discovery and enhance data interoperability. The goal is to represent observational data using a conceptual model that can represent any type of measurement or research context (Madin et al. 2007). Table 1 provides several examples of data models that are currently in use within particular observational data domains. These ontologies can represent many data types, allow multiple interpretations of a single data set, and provide a semantic approach to the interpretation of data, which is thus allowed to range from cross-disciplinary to discipline-specific.

Ontologies provide more explicit representations of the concepts and relationships that can exist between an organisational structure and the diversity of data types that the structure is capable of

**TABLE 1. Representative technology efforts to model “observations” within the ecological and environmental sciences.** (Thanks to the Matt Jones of the National Center for Ecological Analysis and Synthesis for use of this table.)

<b>Initiative</b>	<b>Short description of observational data modeling approach</b>
<b>SEEK</b>	The Science Environment for Ecological Knowledge extensible observations ontology (OBOE) focuses on capturing the essential information about observations required to comprehensively discover and integrate heterogeneous ecological data.
<b>NatureServe</b>	The NatureServe Observational Data Standard focuses on developing an XML Schema for specimen-oriented survey data to improve data aggregation and sharing within and between organisations.
<b>ALTER-NeT</b>	The European ALTER-NeT Ontology, CEDEX, focuses on developing an object-oriented data system for cataloguing observational ecological data while retaining semantic information to aid data discovery and analysis.
<b>SPIRE</b>	The Spire initiative ( <a href="http://spire.umbc.edu/">http://spire.umbc.edu/</a> ) focuses on developing domain-independent, general-purpose ontologies to enable annotation of the contents and structure of existing ecological databases with an initial focus on taxonomy and food web issues (ETHAN).
<b>OGC</b>	The Open Geospatial Consortium Observation and Measurement Standard focuses on developing a generic conceptual XML Schema for representing all aspects of observation and measurement data.
<b>VSTO</b>	The Virtual Solar-Terrestrial Observatory focuses on building ontologies for interoperating among different existing meteorological and atmospheric metadata standards.
<b>TDWG</b>	Biodiversity Information Standards (TDWG, <a href="http://www.tdwg.org">http://www.tdwg.org</a> ) is developing a “meta-model” to integrate biodiversity observations with specimen data by identifying similarities between these two data types, determining whether existing standards suffice to describe them, and if not, developing the additional concepts needed for clarification
<b>ODM</b>	The Consortium of Universities for the Advancement of Hydrologic Science (CUAHSI) Observations Data Model (ODM) and associated relational database focus on storing hydrologic observations data in a system designed to optimise data retrieval for integrated analysis of information collected by multiple investigators.

housing. Ontologies allow knowledge representation to be more flexible, thereby providing a more comprehensive resource for data discovery and integration (Jones et al. 2006). Ontologies allow rich semantic relationships to be developed among terms and attributes, and provide strict rules for the specification of those relationships.

The aim of these initiatives (Table 1) is to

- provide a common model that facilitates interoperability between observational data sets,
- leverage existing work (i.e. Web Ontology Language, OWL), and
- exploit semantic technologies for data organisation (Doerr 2003).

For the organisation of biodiversity data, achieving this aim begins with a common data conceptualisation that applies across all biodiversity data types (i.e., the results from directed surveys, broad-scale surveys, and specimen collections). This foundation provides a model for the description and identification of aggregated data resources, but is extensible to allow for specialisation within specific observational data gathering processes.

A single “core” observational ontology data model should focus on the fundamental concepts of an observation, but also allow extensions for domain-specific entities. In so doing, these “core” components of the model can enhance knowledge discovery, improve data organization, and increase data interpretations. For example, the development of the “core” of the model will provide greater data interoperability from observations from widely varying projects (Madin et al. 2007). At the same time, the domain-specific extensions will ensure that there is no data loss.

## Conclusions

Gaining a solid understanding of the patterns by which organisms are distributed across a broad range of spatial and temporal scales requires vast amounts of species occurrence data. The foundation for such an understanding is the organisation of these resources in such a way that they can be discovered and accessed. The initiatives that are undertaking this organisation and the provision of discovery tools are now recognising that the information that is available regarding species occurrence is much richer than previously estimated.

The addition of observational data that have been gathered using directed or broad scale survey techniques to specimen-based datasets not only increases data volume but also provides more detailed contextual information about how the data were gathered. In turn, this provides much more detailed information for analysis and modelling, such as more accurate mapping of species distributions, estimates of relative abundance, or population trends over time. The result is a much more detailed view of the distribution and abundance of organisms and the factors that might influence this.

Existing descriptive metadata standards and data exchange schemas are an excellent first step for organising biodiversity data, but more work must be done. Recent advances in semantic technologies, particularly observational ontologies, will increase the extensibility of data organisation by providing greater opportunities for data synthesis, and incorporating the specialisation of particular biodiversity domains (Madin et al. 2007, Lagoze 2001). The use of ontologies for observational data will make possible the development of a general observational data model that describes species occurrence data. Because such data are the foundation of biodiversity studies and conservation, such a model will provide both the description of and access to the aggregated resources of the biodiversity community.

Creation, implementation and sustained management of an integrated and comprehensive data curation strategy is essential to best meet the grand challenges of biodiversity conservation and provide the resource for sound decision making. This task will not be easy, and will require coordination and cooperation across diverse disciplines by domain scientists, informatics and computational specialists, application developers, database managers, and visualisation specialists.

The new data framework that must be developed as part of this strategy must overcome the data dependencies that disciplines or individuals within a discipline often embrace, and address the needs of inherently diverse research cultures. It should employ a varied array of concepts, practices, and terminologies for gathering, managing, and providing access to data. Specifically, a successful data framework for biodiversity must provide an overlay of interoperability and accommodate the diversity of resources available. It must provide sufficient bridging of data sets so researchers can investigate complex ecological and environmental issues at broad geographic or temporal scales. Most importantly, the data framework must provide an obvious improvement over current practices for a broad range of users that includes scientists, educators, students, land managers, and the interested public.

## Acknowledgements

This material is based on work supported by the National Science Foundation under Grant Numbers ITR- 0427914, DBI- 0542868, and IIS- 0612031. Additional support was provided by the Leon Levy Foundation, The Wolf Creek Foundation, and the Global Biodiversity Information Facility. The author would like to thank Jon Bart, Wesley Hochachka, Brian Sullivan, Daniel Fink, and Mirek Riedewald for critical comments, and Matt Jones for permission to use Table 1. Finally, the author would like to thank Larry Speers and Vishwas Chavan for their encouragement and Meredith Lane for editing the manuscript.

## Literature Cited

- Andelman, S. J., C. M. Bowles, M. R. Willig, and R. B. Waide. 2004. Understanding environmental complexity through a distributed Knowledge Network. *BioScience* **54**:240-246.
- Beaman, R., J. Wieczorek, and S. Blum. 2004. Determining Space from Place for Natural History Collections In a Distributed Digital Library Environment. *D-Lib Magazine* 10 (5). ISSN 1082-9873. <http://dlib.org/dlib/may04/beaman/05beaman.html>
- Becker, B.-H., and S.-R. Beissinger. 2006. Centennial decline in the trophic level of an endangered seabird after fisheries decline. *Conservation Biology* **20**:470-479.
- Berendsohn, W. G. 2007. Access to Biological Collections Data- ABCD, <http://www.tdwg.org/activities/abcd/charter/>.
- Chapman, A. D. 2005. Principles of Data Quality. Report for the Global Biodiversity Information Facility, Copenhagen.
- Cooper, J.-H., and F.-D. Steinheimer. 2003. Why museums matter: Report from the workshops 14-15 November 1999 'Increased cooperation between bird collections'. *Bulletin of the British Ornithologists' Club* **123A**:355-360.

- Dhondt, A. A., D.L. Tessaglia, and R.L. Slothower. 1998. Epidemic mycoplasmal conjunctivitis in House Finches from Eastern North America. *Journal of Wildlife Diseases* **34**:265-280.
- Dhondt, A. A., W.M. Hochachka, S.M. Altizer, and B.K. Hartup. 2001. The house finch hot zone: Citizen science on the trail of an epidemic. *Living Bird* **20**:24-30.
- Doerr, M., J. Hunter, and C. Lagoze. 2003. Toward a Core Ontology for Information Integration. *Journal of Digital Information* **4**:169-194.
- Ellison, A. M., Leon J. Osterweil, Lori Clarke, Julian L. Hadley, Alexander Wise, Emery Boose, David R. Foster, Allen Hanson, David Jensen, Paul Kuzeja, Edward Riseman, and Howard Schultz. 2006. Analytic Webs Support the Synthesis of Ecological Data Sets. *Ecology* **87**:1345-1358.
- FGDC. 2003. The FGDC Biological Metadata Profile. <http://www.fgdc.gov/library/factsheets/documents/metaprof.pdf>.
- Fischer, J. R., D.E. Stallknecht, M.P. Luttrell, A.A. Dhondt, and K A. Converse. 1997. Mycoplasmal conjunctivitis in wild songbirds: The spread of a new contagious disease in a mobile host population. *Emerging Infectious Diseases* **3**:69-72.
- Friedman, T. 2005. *The World Is Flat*. Farrar, Straus and Giroux.
- GBIF. 2006. GBIF Plans 2007-2011: from prototype towards full operation ([http://www.gbif.org/GBIF\\_org/GBIF\\_Documents/strategic\\_plans](http://www.gbif.org/GBIF_org/GBIF_Documents/strategic_plans)).
- Hamilton, M. P., E. A. Graham, P. W. Rundel, M. F. Allen, W. Kaiser, M. H. Hansen, and D. L. Estrin. 2007. New Approaches in Embedded Networked Sensing for Terrestrial Ecological Observatories. *Environmental Engineering Science* **24**:192-204.
- Hochachka, W. M., and A.A. Dhondt. 2000. Density-dependent decline of host abundance resulting from a new infectious disease. *Proceedings of the National Academy of Sciences of the United States of America* **97**:5303-5306.
- Hochachka, W., R. Caruana, D. Fink, A. Munson, M. Riedewald, D. Sorokina, and S. Kelling. 2007. Data mining for discovery of pattern and process in ecological systems. *Journal of Wildlife Management* **71**:2427-2437.
- Jones, M.-B., M.-P. Schildhauer, O. J. Reichman, and S. Bowers. 2006. The new bioinformatics: Integrating ecological data from the gene to the biosphere. *Annual Review of Ecology Evolution and Systematics* **37**:519-544.
- Kelling, S., W. M. Hochachka, D. Fink, M. Riedewald, R. Caruana, G. Ballard and G. Hooker. 2008 (submitted). Data intensive science: A new paradigm for biodiversity studies.
- Kutner, L., and B. Stein. 2006. *Observational Data Standard, Version 1.0*.
- Lagoze, C., J. Hunter 2001. The ABC Ontology and Model. *Journal of Digital Information*. <http://oasis-open.org:1-18>.
- Laihonen, P., and R. Kalliola. 2004. Biodiversity information clearing-house mechanism (CHM) as a global effort. *Environmental Science and Policy* **7**:99-108.
- Ley, D. H., J.E. Berkhoff, and S. Levisohn. 1997. Molecular epidemiologic investigations of *Mycoplasma gallisepticum* conjunctivitis in songbirds by random amplified polymorphic DNA analyses. *Emerging Infectious Diseases* **3**:375-380.

- Luttrell M.P., J. R. F., D.E. Stallknecht, and S.H. Kleven. 1996. Field investigation of *Mycoplasma gallisepticum* infections in house finches (*Carpodacus mexicanus*) from Maryland and Georgia. *Avian Diseases* **40**:335-341.
- MacArthur, R. H. 1972. *Geographical Ecology: Patterns in the Distribution of Species*. Harper and Row.
- Madin, J., S. Bowers, M. Schildhauer, S. Krivov, D. Pennington, and F. Villa. 2007. An ontology for describing and synthesizing ecological observation data. *Ecological Informatics* **2**:18.
- Maschinski, J., J. E. Baggs, P. F. Quintana-Ascencio, and E. S. Menges. 2006. Using Population Viability Analysis to Predict the Effects of Climate Change on the Extinction Risk of an Endangered Limestone Endemic Shrub, Arizona Cliffrose. *Conservation Biology* **20**: 218-228.
- Michener, W. K. 2006. Meta-information Concepts for Ecological Data Management. *Ecological Informatics* **1**:3-7.
- Nichols, J.-D., and B.-K. Williams. 2006. Monitoring for conservation. *Trends in Ecology and Evolution* **21**:668-673.
- Phokion, G. K., P. Jonathan, and T. Wang-Chiew. 2006. The complexity of data exchange. Proceedings of the twenty-fifth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems. ACM, Chicago, IL, USA.
- Platt, J. R. 1964. Strong Inference. *Science* **146**:347-353.
- Ponder, W. F., G. A. Carter, P. Flemons, and R. R. Chapman. 2001. Evaluation of museum collection data for use in biodiversity assessment. *Conservation Biology* **15**:648-657.
- Ralph, C. J., S. Droege, and J. R. Sauer. 1995. *Managing and Monitoring Birds Using Point Counts: Standards and Applications*. Pacific Southwest Research Station, Forest Service, U.S. Department of Agriculture, Albany CA.
- Vieno, M., M. Sarvala, I. Saaksjarvi, and T. Toivonen. 2005. Observations on Observational Data. Pages 5-12 *in* M. Sarvala, M. Vieno, and T. Toivonen, editors. *Observations on Observational Biodiversity Data*. University of Turku, Finland.
- UNEP. 2005. UNEP Annual Report 2004: 76.
- University of Washington Health Services. 2000. Definitions of Commonly Used Research Terms (<http://www.washington.edu/healthresearch/definitions.html>).
- Van Horne, B., P. Schmidt, B. Andres, L. Barnhill, J. Bart, R. Bishop, S. Brown, C. Francis, D. Hahn, D. Humburg, M. Koneff, B. Peterjohn, K. Rosenberg, J. Sauer, R. Szaro, and C. Vojta. 2007. Opportunities for Improving Avian Monitoring. U.S. North American Bird Conservation Initiative Committee.
- Wiens, J. A. 1992. *Cambridge Studies in Ecology: The ecology of bird communities*, Vols. 1 and 2. Foundations and patterns, Vol. 1; Processes and variations, Vol. 2.
- Winker, K. 1996. The crumbling infrastructure of biodiversity: The avian example. *Conservation Biology* **10**:703-707.
- Winker, K. 2004. Natural history museums in a postbiodiversity era. *BioScience* **54**:455-459.

