Methods in Ecology and Evolution

British Ecological Society

Methods in Ecology and Evolution 2010, **1**, 199–211 doi: 10.1111/j.2041-210X.2010.00023.x

The Primate Life History Database: a unique shared ecological data resource

Karen B. Strier[1]*, Jeanne Altmann[2,11], Diane K. Brockman[3], Anne M. Bronikowski[4], Marina Cords[5], Linda M. Fedigan[6], Hilmar Lapp[7], Xianhua Liu[7], William F. Morris[8], Anne E. Pusey[9], Tara S. Stoinski[10] and Susan C. Alberts[8,11]

[1]Department of Anthropology, University of Wisconsin-Madison, Madison, WI, USA; [2]Department of Ecology and Evolutionary Biology, Princeton University, Princeton, NJ, USA; [3]Department of Anthropology, University of North Carolina-Charlotte, Charlotte, NC, USA; [4]Department of Ecology, Evolution and Organismal Biology, Iowa State University, Ames, IA, USA; [5]Department of Ecology, Evolution and Environmental Biology, Columbia University, New York, NY, USA; [6]Department of Anthropology, University of Calgary, Calgary, Canada; [7]National Evolutionary Synthesis Center, Durham, NC, USA; [8]Department of Biology, Duke University, Durham, NC, USA; [9]Department of Evolutionary Anthropology, Duke University, Durham, NC, USA; [10]The Dian Fossey Gorilla Fund International and Zoo Atlanta, Atlanta, GA, USA; and [11]Institute of Primate Research, National Museums of Kenya, Nairobi, Kenya

Summary

1. The importance of data archiving, data sharing and public access to data has received considerable attention. Awareness is growing among scientists that collaborative databases can facilitate these activities.

2. We provide a detailed description of the collaborative life history database developed by our Working Group at the National Evolutionary Synthesis Center to address questions about life history patterns and the evolution of mortality and demographic variability in wild primates.

3. Examples from each of the seven primate species included in our database illustrate the range of data incorporated and the challenges, decision-making processes, and criteria applied to standardize data across diverse field studies. In addition to the descriptive and structural metadata associated with our database, we also describe the process metadata (how the database was designed and delivered) and the technical specifications of the database.

4. Our database provides a useful model for other researchers interested in developing similar types of databases for other organisms, while our process metadata may be helpful to other groups of researchers interested in developing databases for other types of collaborative analyses.

Key-words: bioinformatics, data archiving, data sharing, database development, evolutionary biology, population ecology

Introduction

The accumulation of long-term ecological data over the past several decades, and increasing recognition of the need for broad collaborative research efforts, present new challenges as well as opportunities for the scientific community. The design and curation of databases that can accommodate the complex, dynamic data sets typical of ecological research pose a practical hurdle because of the enormous range of data that is often involved. In some cases, these data sets are historic and not yet digitized, but of great potential value nonetheless. More frequently, the data sets are contemporary, and are completely or partially digitized, but in non-standardized ways. Inadequacies in the development and maintenance of these data sets may make them largely inaccessible for the kinds of synthetic, collaborative analyses that many ecological questions require (Cook *et al.* 2001; Michener 2006).

Incorporating data from multiple studies that span diverse species and observational conditions into integrated databases for comparative analyses is even more difficult to achieve because differences in data collection and sampling methods across studies must first be reconciled, and a standard vocabulary must be developed based on common criteria. Identifying and standardizing this vocabulary can be an arduous process that relies on the expertise of investigators with sufficient familiarity with the long-term studies to understand and explain the nuances in their data sets (Nelson 2009). Yet, although the benefits of archiving, sharing and increasing public access to

*Correspondence author. E-mail: kbstrier@wisc.edu
Correspondence site: http://www.respond2articles.com/MEE/

biological and ecological data have received considerable attention in the literature (e.g. Arzberger *et al*. 2004; Parr & Cummings 2005; Piwowar *et al*. 2008; Schofield *et al*. 2009; Toronto International Data Release Workshop 2009), there are still only a small number of published examples that describe how synthetic, integrated databases are created and used (e.g. Ellison *et al*. 2006; Even, Shankaranarayanan, & Watts 2006; Jones *et al*. 2008).

Here, we describe the development, design and implementation of the Primate Life History Database (PLHD), a product of a larger collaborative endeavour jointly funded by the National Evolutionary Synthesis Center (NESCent) and the National Center for Ecological Analysis and Synthesis (NCEAS). The PLHD incorporates individual, longitudinal life history data from long-term field studies of wild primates into a synthetic database for comparative analyses. We present the descriptive and structural metadata associated with this database, as well as the process metadata (i.e. a description of how the data set was designed and delivered). In addition to reviewing the content of the database and its underlying rationale, we discuss the criteria we employed to standardize data from seven different field studies, and we provide examples from the database to illustrate its design and technical specifications and the range of data included.

Our scientific motivation for the endeavour lies in our shared interests in comparative analyses of primate life histories, and specifically, in the application of data from wild populations to address questions about the evolutionary ecology of life histories. Age at first reproduction, fertility, longevity and other variables that influence both fitness and population dynamics fluctuate in response to local ecological, social and demographic conditions. Understanding this variability offers insights not only into life history evolution, but also into the conservation and management of endangered species and the ecological impacts of global climate change (Strier *et al*. 2006).

Primates are long-lived compared with most other mammals, and the decades of research required to document their individual life histories make these longitudinal data irreplaceable resources. Yet, despite widespread recognition of the value of these data, primate researchers, like other ecologists, often lack the expertise and resources to construct the kinds of databases that facilitate data sharing and access, and that are needed to protect data from the information loss that can occur over time because of inadequate curation or maintenance (Michener 2006; Jones *et al*. 2008). Our database provides a useful model for other researchers interested in developing similar life history databases for other organisms. At the same time, the description of the process by which we developed our database may be helpful to other groups of researchers interested in developing databases for other types of collaborative analyses.

Materials and methods

Our collaboration involved a multi-stage process. The first stage involved setting the agenda for the development of the PLHD (described in Strier *et al*. 2006). This initial stage led to the formation of the core collaborative group, the Working Group, which included three sets of researchers. (i) Researchers representing seven ongoing field studies of wild primates ranging from 24 to 45 years in duration (Alberts, Altmann, Brockman, Cords, Fedigan, Pusey, Stoinski, Strier). (ii) Two evolutionary ecologists with a particular interest in demography (Bronikowski, Morris). (iii) Two NESCent informatics specialists (Lapp, Liu). The PLHD was developed by our Working Group over the course of three 4- to 5-day meetings held at NESCent in August 2007, January 2008 and August 2008.

The species represented by the seven field studies are taxonomically diverse. They include one indrid (Verreaux's sifaka, *Propithecus verreauxi*, research ongoing for 24 years), two New World monkeys (white-faced capuchin, *Cebus capucinus*, and northern muriqui, *Brachyteles hypoxanthus*, studies ongoing for 25 and 27 years, respectively), two Old World monkeys (yellow baboons, *Papio cynocephalus*, 37 years, and blue monkeys, *Cercopithecus mitis*, 29 years) and two great apes (eastern chimpanzees, *Pan troglodytes schweinfurthii*, 45 years, and mountain gorillas, *Gorilla beringei beringei*, 41 years). All populations are wild, and with a few exceptions, no provisioning or interventions have occurred. Exceptions include veterinary intervention and historical provisioning in the chimpanzee population (i.e. almost daily from 1964 to 1967 and 1990 to 1996 for different communities, after which it was reduced and terminated altogether in 2000; Wrangham 1974; Goodall 1986; Pusey, Wilson, & Collins 2008), veterinary intervention in the gorilla population (Mudakikwa *et al*. 2001), occasional access to human-related foods by the blue monkeys (Cords & Chowdhury, in press) and one instance in which researchers rescued an infant muriqui and returned her to her mother (Nogueira *et al*. 1994).

Development of the database itself occurred in two parallel processes. One process involved intensive discussion among the primate researchers to design a shared terminology of attributes that would capture the relevant life history data for all seven species in a standardized manner. The goal was to permit us to address two central questions; one about the evolution of mortality across species and by sex, and the other about patterns of demographic variation across species. Decisions about which data attributes to include and their relationships to one another were fairly straightforward, but defining the meaning and constraints of each attribute operationally, in ways that made biological sense for all species and that accounted for differences among studies, required painstaking consideration. At the same time, the informaticians conferred with the researchers to design a database that presented the data as a series of three 'views', or virtual tables that represent organized, systematic abstractions of the underlying database. The three main views that resulted were: Biography (Table S1, Supporting information); Fertility Intervals (hereafter, Fertility; Table S2, Supporting information); and Study Population (Table S3, Supporting information).

Biography includes all live-born individuals (females and males) in our study populations, and is the core of our database. In the aggregate, Biography includes data that permit us to calculate the life span for each individual (and to identify both left-truncated and right-censored records, i.e. cases in which individuals were already alive at the onset of observations or still alive when observations ceased, respectively). Because it also includes the identity of the mother for each individual (if known), we can use it to calculate fertility for each mother identified in the database as well. Fertility identifies, for each female, any interruptions in continuous observations that could have resulted in missed live births. Together, Biography and Fertility allow the user to determine the sequence and number of live births that each mother experienced during her lifetime (see 'Database content', for

more details). The STUDY POPULATION view designates a distinctive code for each study, which is used to identify the study for each individual in the BIOGRAPHY and FERTILITY views.

The database was populated in large batch uploads from spreadsheet-formatted data (see below under database design). A web-based graphical user interface was also developed to enable data editing and entry by individual researchers and their collaborators. This interface is flexible yet comprehensive, and allows various levels of access control. For instance, while the database administrators have read and write access to all of the data, other users of the database can be limited to read-only (search) access to one or some or all of the studies, or can be given edit privileges to one or several studies. This arrangement allows the administrators to grant appropriate access to each Working Group member (i.e. each Principal Investigator (PI) has read and write access to their own data, but only read access to others' data). It also allows research personnel from each study to gain access to that study's data (but not to data from other studies), even if they did not participate in the working group directly. Only the administrators can create users, and grant study-specific read or write access.

Because all members of the Working Group would have access to the entire database, we developed a Memorandum of Understanding (MOU) at our first meeting (http://demo.plhdb.org). We agreed that PIs could add their own collaborators as users of the database for the species with which they work, but access to the entire database is currently limited to the Working Group members who have signed our MOU.

Database content

Our intention was to have a fairly simple set of relational tables that reflects the individual-based nature of our data. That is, for each study, the individual animal is the unit of analysis, and hence the set of all individual animals' life history data comprise the data of interest. BIOGRAPHY has one row for each of the 3351 individuals across the seven studies. Each of the 17 columns pertains to a life history variable or estimates or ranges of error of these variables (Table S1, Supporting information).

Measuring individual (as opposed to population-level) fertility represents a special challenge in studies of wild animals, because unlike longevity (measured as the interval between the birth and the death or disappearance of a known individual), individual fertility (measured, in our case, as the interval between recorded live births) will be inaccurately estimated if even short gaps in observation result in missed births. For this reason, we recognized the need to identify, for each female in each study, the periods during which we were reasonably sure that we had captured all births, and equivalently, the periods during which we were not able to rule out the possibility that a birth and death of an infant had occurred during a gap in observations. FERTILITY captures this information (Table S2, Supporting information).

In addition, in STUDY POPULATION, we provide information about each study (location, species, etc.). STUDY POPULATION contains one row for each study ($N = 7$; Table S3, Supporting information), identified by a unique, arbitrarily assigned number and to which all individuals represented in the other views for each study are linked.

Data standardization

We designed the relational database to permit us to address specific questions about the evolution of primate life histories, such as whether species and sexes age at similar rates (A.M. Bronikowski *et al.*, unpublished data) and whether population growth is more sensitive to female fertility vs. infant or adult survival (W.F. Morris *et al.*, unpublished data). This meant identifying the variables most critical to our analyses and then reconciling variability in data coding, observation schedules and confidence intervals for estimated dates to ensure that our criteria for assigning values were uniform and comparable across studies. Differences in the behaviour of the animals and in logistical conditions resulted in variation in data-coding systems both within individual studies and between the studies in the database. These sources of variation necessitated the development of a common vocabulary, established among Working Group members, which would ensure that terms were used the same way across studies. The common vocabulary is encompassed by the terms defined below. These terms fall into three different logical units: one belonging to the biographical properties of an animal (BIOGRAPHY), one belonging to the fertility properties of an animal (FERTILITY) and one belonging to the study population in which the animal lives (STUDY POPULATION).

CONTENTS OF BIOGRAPHY

StudyID

Each of the different study populations, representing different species in our database, was assigned a distinct ID code.

AnimID

Because all of the data were individual-based, and the individual was the unit of analysis in all studies, the ID of each animal (typically an abbreviated code) in each study population was the fundamental unit of information around which all the data were organized. All individuals in each of the studies were unambiguously identifiable by their distinct physical characteristics or, in one case (sifaka), by tagged collars and ear-notches. Habituation to human observers facilitated the recognition of individuals. Within each study, there was a one-to-one relationship between an animal's ID code and the identity of an actual animal in each study population. However, AnimID was not a unique field; animals in different studies might share an AnimID (for instance, study 2 and study 5 both have an animal with AnimID = 'AFR'). Consequently, it is the combination of AnimID and StudyID that produces a unique identifier for each animal in the database.

AnimName

We included a column for the full name of each animal whenever these had been assigned. This was included for completeness, and to enhance the ability of individual researchers to

confirm the accuracy of all records associated with that individual.

BirthGroup and BGCertainty

Our life history analyses were aimed primarily at species-level differences. However, the social nature of primates in general, and the variation in their dispersal patterns in particular, led us to include a column for specifying the social group into which an animal was born (BirthGroup) and the researcher's confidence in this assignment (BGCertainty). These variables will permit us to collectively or individually evaluate whether the group of birth contributes to life history variance.

Sex

Because many life history variables (e.g. age at maturity, dispersal, life span) are known to be sex-specific, we distinguished each individual in our database by sex (M or F). Occasionally, neonates that were born alive have died before researchers were able to assign a sex; hence 'U' (for unknown) was an allowed value in this column.

MomID

This attribute corresponds to the AnimID of an individual's mother, when it was known. No value was assigned for individuals whose mothers were unknown. MomID allows us, in combination with Fertility (see below) to measure fertility for each mother recorded in the database. Also, by associating each individual with his or her mother's AnimID, when known, we can evaluate how individual survivorship (and female fertility) relates to birth sequence and maternal age. As is the case with AnimID, MomID must be used in combination with StudyID to identify a unique mother.

FirstBorn

Because age at first reproduction is a critical life history marker, we distinguished whether individuals were known to be their mother's first offspring.

Birthdate, BDMin, BDMax

Birth dates, and estimates of the range of possible dates in which the birth could have occurred (BDMin and BDMax), are the key to estimating individual ages, and therefore necessary for all analyses of life histories. In some cases, continuity in field personnel and cohesive grouping patterns facilitated daily or near-daily monitoring of all subjects, and the range of days over which Birthdate could be estimated was small. When a mother was observed on sequential days, first without an infant and subsequently with an infant, the Birthdate and BDmax were usually recorded as occurring on the second day, with the BDmin corresponding to either the first or second day depending on the study. In other cases, either gaps in observations or the tendency of individuals to travel widely made it difficult for observers to monitor all individuals regularly. Birthdate estimates in these circumstances were less precise, and the difference between BDMin and BDMax was much greater.

Some Birthdate entries, including those for animals that were already present at the onset of observations of their population or social group, were necessarily estimated on the basis of the individual's visible size or developmental characteristics. These birthdate assignments for adults tended to have larger intervals between BDmin and BDmax estimates than animals first seen as immatures, because there were often few, if any, visible differences between young, middle aged or older adult animals. In these cases, the range of possible birth dates assigned (BDMin and BDMax) depended on the researchers' confidence in their estimates. Birth dates were estimated based on a variety of traits in different species, including dental wear patterns at the time of capture/tagging (used in sifaka), or visible physical similarities with adults of known ages (used in most studies). In the case of females, some species exhibit visible signs of parity, and researchers used these signs to estimate Birthdate and BDmin and BDmax from the average ($\pm$SD) age ranges known for nulliparous and primiparous females in their study populations. Nonetheless, there were often still very large ranges of possible birth dates for some individuals; by having this information in the database, we could decide whether or not to include particular individuals in specific analyses.

BDDist

In a further effort to increase precision in birth date estimates, we assigned a birthdate distribution (BDDist) of Normal (N) when we considered the most likely birthdate to be closer to Birthdate than to BDMin or BDMax, and of Uniform (U) when any birthdate between BDMin and BDMax (including Birthdate) was equally likely. A normal distribution of the estimated birthdate was assigned if BDMin and BDMax represented ± 2 SD from the most likely Birthdate. Uniform distributions were assigned if the probability distribution was truncated at either BDMin or BDMax. For example, a new infant observed after a 30-day gap in observation of its mother would result in the infant's BDMin on the last day its mother had been observed and a BDMax 30 days later when it was first observed. Based on the infant's size and development upon first observation relative to other known infants in the study population, the researcher may have had good reasons to estimate the infant's Birthdate at either the midpoint of BDMin and BDMax, or else closer to either BDMin or BDMax. If estimated at the midpoint, then BDDist could be either Normal or Uniform, but if not at the midpoint, the BDDist would necessarily be assigned as Uniform.

Entrydate and Entrytype

Identifying the date at which individuals entered their respective study populations allowed us to differentiate uncensored observations from left-truncated observations for survival analyses. Individuals were considered to enter their respective

study populations at the time at which they could be individually identified and close observations on them began. In most cases, this corresponded to their birth. However, some individuals immigrated into the study populations some time after their birth, and in these cases Entrydate corresponded to immigration date (or to confirmed AnimID via tagging in sifaka). Finally, all studies included individuals that were present at the onset of the study itself or when close observations were initiated on new groups in the study population; in these cases, Entrydate corresponded to the onset of the study or the individual's inclusion in the study population.

Entrytype specified each of the four possible ways in which a subject could have entered the study population: birth (B); immigration (I); start of confirmed AnimID (C) and initiation of close observation (O). Although births and immigrations were easily assigned, investigators differed in their designations of C and O. For example, in some cases an animal had been recognizable and familiar to the researchers based on occasional or opportunistic sightings before close observation began. In most cases, these animals only entered the database when they immigrated (I) into one of the established study groups. However, in other cases, they entered the database because systematic observations were initiated on their group; in this case, either C or O could have been used. Each researcher described her assignments of C and/or O in their User's documentation, but these entry types are functionally equivalent in terms of analyses, and should be treated as such by database users.

Departdate and DepartdateError

The last date on which an animal was observed in the study population is the Departdate. However, not all animals were equally visible to observers on a daily basis, and observation schedules varied across the different studies and over time and between groups within studies. To capture the variation in the reliability of Departdates between and within the different studies, we calculated the DepartdateError, which reflects the time between Departdate (last date observed) and the first time that an animal was confirmed missing (e.g. when observations resumed and all individuals present could be expected to be re-encountered). DepartdateError was expressed as a fraction of a year (number of days divided by number of days in a year), and was >0 whenever the number of days between Departdate and retrospective confirmed missing date was >15 days. In some studies, members of the study population did not live in cohesive groups, making it difficult to specify an expected lag to re-sighting and a corresponding DepartdateError. In cases when DepartdateError could not be calculated, its value was missing.

Departtype

Similar to Entrytype, we distinguished four Departtypes: death (D); emigration (E); permanent disappearance (P) and the end of observation (O), which means that the individual was still present at the most recent census date. Observations of deaths and the recovery of identifiable corpses in most primate habitats are extremely rare, but we nonetheless required strong circumstantial evidence, such as visibly poor health or other mortality risks, or violations of population-specific behaviour patterns, before assigning D. For example, the sudden permanent disappearance of an animal of the typically non-dispersing sex for that population could have been assigned a Departtype of D, even in the absence of a corpse or other circumstantial evidence. If the animal that disappeared was a member of the dispersing sex, then D was allowed when the disappearance occurred before the youngest known dispersal age in that population, subject to the researchers' expert opinion. Additional information, such as locations associated with high risk, were also considered when assigning D in the absence of observed death or identifiable corpse.

D was never assigned based solely on inferred risks associated with age, and E was never assigned solely on the basis of the disappearance of an individual at the appropriate age and sex for dispersal. To assign E, researchers had to be confident that the individual had emigrated even if its subsequent fate was not known. All disappearances that could not be attributed to D or E were assigned P in the database. In demographic analyses, P, E and O all function to signal right-censored observations. The four types of Departtypes are equivalent to Stoptype in Fertility.

CONTENTS OF Fertility

StudyID

See Biography.

AnimID

See Biography.

Startdate and Stopdate

The most difficult task in constructing Fertility was the development of standardized criteria for what constituted sufficiently continuous observations of a female to merit inclusion of that period, vs. gaps in observations that would be long enough to have possibly resulted in a missed birth record. Each row in Fertility corresponded to one uninterrupted period of observation on a female (an interval during which no possible births would have been missed). Each female for which at least one such uninterrupted period was obtained was represented by one or more rows in Fertility. In addition to the variation in the observation schedules for each study, some of the primates in our database are elusive and impossible to monitor on a daily basis. Each of the primatologists provided detailed documentation about the criteria they applied.

Starttype and Stoptype

See Entrytype and Departype in Biography; these correspond to Starttype and Stoptype in Fertility, where they defined the

beginning and end of each uninterrupted period of observation during which no possible births would have been missed, as described above.

CONTENTS OF STUDY POPULATION

StudyID

See BIOGRAPHY.

Commonname

We used the names most commonly used in the literature for each species, although we recognize that conventions on common names differ regarding the use of English vs. indigenous names in different parts of the world.

Sciname

We provided the scientific names for each species, as their Latin genus and species names. Although taxonomic assignments of some of the species in our database are undergoing reassessment, we used the scientific names identified by each PI as the best current designation. Subspecies designations were included in some cases to distinguish recognized biogeographical variants.

SiteID

Each study population in the database was identified by the name of the park or reserve in which it occurs and by which it is known in the current literature.

Owners

Name(s) of individual(s) and/or organization(s) that control access to the data. Owners were either the principal investigators or the designated representatives of the study populations' research groups. Owners were included to identify the key contact individuals for the life history data from each study population.

Latitude/Longitude

Because SiteIDs have the potential to change with new political or conservation initiatives, we included the latitudinal and longitudinal coordinates for each study site. These coordinates were given in degrees, to the nearest thousandth.

Design and deployment of the database and user interface

We designed the database on the premise that the common vocabulary developed by the Working Group and the three views corresponding to the logical units would serve as a well-defined standard interface, both for accepting and delivering data to and from the database, and for programming user-facing applications. We implemented the standard interface as three database views, one corresponding to each of the three logical units of the terminology (BIOGRAPHY, FERTILITY and STUDY POPULATION) that sit on top of a physical, normalized data model. The attributes of each view are the terms that comprise the respective logical unit of our terminology. Database views, by definition, present the data for browsing, querying and retrieval; we also made them support the full set of CRUD (Create, Retrieve, Update, Delete) data manipulation operations by implementing code that executes within the database server (called 'stored procedures') and translates between the underlying normalized data model and the standard interface views.

The physical data model powering the three database views essentially follows a third normal form model of the logical entities represented by the common vocabulary in the database (for a description and discussion of relational database normal forms, see Date 2004; Kent 1983). Specifically, the normalized entities include: Individual (the individual animal); Individual_Relationship (parent-child links between individuals); Study (a population of individuals); Site (the geographic site of a study); Observation (biographical events of a given type at a specified time, such as birth, death or emigration) and RecordingPeriod (a period of time that starts and ends with an observation, such as the observed fertility interval of a female individual). Furthermore, the normalized model includes two additional tables, CVTerm and CVTerm_Relationship, which hold a controlled vocabulary designating the specific type and meaning of the data stored in each row of various tables, such as the type of observation (in table Observation), or the type of relationship (in table Individual_Relationship). The CVTerm_Relationship table relates terms to super-classes that allow a client application, such as the web application we developed, to obtain the allowable terms (types) for a particular super-class. This in turn allows the application to validate input data for allowable types, for instance, the types of start dates and end dates of biographies and fertility intervals. For example, 'permanent_disappearance' is a term designating the type of an observation, and is one of the subtypes of the super-class 'end_of_recording', but not a subtype of 'start_of_recording'. The web application uses this to offer 'permanent_disappearance' as an option only for end dates but not for start dates.

The database was implemented in the PostgreSQL relational database engine, which is an open-source and freely available. We wrote stored procedures in the PL/pgSQL database programming language, natively supported by PostgreSQL, to encapsulate the necessary business logic behind look-up, create and update data operations in an application programming interface (API). The API enables client code to obtain a record by supplying identifying attributes (such as AnimID and StudyID for an individual), and a parameter that dictates whether the API will simply try to locate and return the record, to create it first if it cannot find it, or to update it first if it can find it. The database views combine data from several tables in the underlying model, making the data from some tables appear multiple times (called 'de-normalization' in database theory). Thus,

the views cannot directly allow data manipulation operations (create, update and delete). We therefore used a combination of the PostgreSQL-specific capability of 'rules' and standard database triggers to support these operations. The PostgreSQL rules re-route data manipulation operations executed against the database views to PL/pgSQL stored procedures that use the PL/pgSQL API described above. To facilitate bulk import of spreadsheet-formatted data, we implemented tables specifically designated for this purpose; these tables mirror the spreadsheet templates agreed upon by the scientists of the Working Group. The COPY FROM command, which is built into PostgreSQL, populates a destination table from the rows of an input data file in spreadsheet comma seperated values (CSV) format. We then wrote PL/pgSQL database triggers to intercept incoming rows of data and re-route them through the PL/pgSQL API to the physical normalized relational tables.

Together, these components allowed us to present the data within the database so that it was fully compliant with the terminology and the logical units that were agreed upon and understood by the Working Group scientists. At the same time, our approach utilizes the database engine's capabilities of enforcing basic constraints on the data and preventing data redundancy through a normalized relational model. It also provides an API for client application programmers, which encapsulates the business logic of data look-up and manipulation within the database itself, guaranteeing that all client access passes through the same business logic.

Our approach also allows those participants of the Working Group who are comfortable with desktop database tools, such as MS Access, to connect directly to the database and browse, query, and manipulate the data in the tool of their choice. In addition to supporting direct access, we constructed a web-based user interface application that supports browsing all rows of all tables to which the user is granted access. It also supports adding to and editing biography and fertility observation records, as well as building custom queries and downloading the results as reports for subsequent import into analysis programs. Viewing and editing privileges are controlled by the user authorization module described above, and granted on a per-study basis. Only users with administrative privileges can add users and change user privileges.

The web application was implemented in the Java programming language using the Spring application framework and Hibernate for mapping application data objects to records in the relational database (for links, see http://demo.plhdb.org/jsp/about.jsp). This mapping fully utilizes the database view API described above, and hence is agnostic to the actual physical data model being used. The web application also utilizes the controlled vocabulary term component of the database to provide values for the input fields about biographical events (such as Entrytype, or DepartType, as described above), and at the same time restricts input to those terms that are valid for a given event type (e.g. Immigration is not permitted for a DepartType, whereas Emigration is permitted).

The schema definition and all other source code, including the PL/PgSQL database code, and the source code for the web application are freely available under GNU Public License (see http://www.gnu.org/licenses/gpl.html). In addition, we have created a live instance of the database and application at http://demo.plhdb.org, which is populated with representative sample records from each study represented among the Working Group participants. All source code can be downloaded or browsed at http://github.com/plhdb/plhdb or http://sf.net/projects/plhdb.

Use of the database

One of the immediate analytical goals for the database was the construction of life tables and the analysis of ages at death. An ideal individual was one for whom Birthdate was known to within a fraction of a year, Entrytype was birth, Sex was determined, and Departdate was known to within a fraction of a year. This was true for the majority of data. Individuals of unknown sex were excluded from these analyses, but because most of these were animals that died within a few days of birth, their exclusion would impact our calculations of infant mortality. Similarly, individuals who were present as adults when observations began could have a several year difference between their Birthdate minimum and maximum, resulting in a wide age range at death even if Departdate was known precisely and Departtype was death.

Below, we provide a few examples of how the database, as shown in Tables 1 and 2 and live at http://demo.plhdb.org, can be used to calculate three standard life history parameters: Age at first reproduction; Lifetime reproductive success of females; and Interbirth interval.

AGE AT FIRST REPRODUCTION

Determining a female's age at first reproduction is facilitated by the 'FirstBorn', 'MomID', and 'StudyID' columns in Biography (Table 1). We first pull out the first-born offspring (indicated by 'Y' in 'FirstBorn') for each mother for whom this record exists, then find the mother's date of birth by searching for her entry in Biography, and finally compute her age as the difference between the birth date of her firstborn offspring and her own birth date. For example, we can see that AnimID = BRAH in StudyID = 1 (born on 14 August 1993) was 3292 days (9·01 years) old when her firstborn offspring, AnimID = BRL-J, was born (on 19 August 2002). Note that age at first reproduction can be determined entirely from data in Biography, without reference to Fertility. In this example, the birthdates of both BRAH and BRL-J were known to be within 0–2 days, as indicated by BDMin and BDMax. Including the ranges of possible birthdates in the database permits researchers to evaluate whether or not to include all individuals in particular analyses.

LIFETIME REPRODUCTIVE SUCCESS

One life history component we would often like to know is a female's lifetime reproductive success. We can easily determine all known live-born offspring of a female by searching for all entries for that female in the 'MomID' column of Biography.

Table 1. Sample data from PLHD BIOGRAPHY[1]

Study ID	Anim ID	Anim Name	BirthGroup	BG Certainty	Sex	Mom ID	First Born	Birthdate	BD Min	BD Max	BDDist	Entrydate	Entrytype	Departdate	Departdate Error	Depart type
1	BRIS	Brisa	Matão	C	F	BR	Y	2-Sep-89	15-Aug-89	20-Sep-89	U	20-Sep-89	B	27-Sep-07	Null	D
1	BRAH	Brahma	Matão	C	F	BR	N	14-Aug-93	14-Aug-93	14-Aug-93	N	14-Aug-93	B	14-Dec-08	0	O
1	BRS	Brasa	Matão	C	F	BR	N	4-Jul-96	4-Jul-96	4-Jul-96	N	5-Jul-96	B	17-Feb-99	0	D
1	BRE	Brenda	Matão	C	F	BR	N	29-May-99	29-May-99	29-May-99	N	29-May-99	B	10-Apr-02	0	D
1	BRN	Branca	Matão	C	F	BR	N	20-Jul-02	20-Jul-02	20-Jul-02	N	20-Jul-02	B	20-Apr-04	0	D
1	BN	Breno	Matão	C	M	BR	N	2-Aug-04	22-Jul-04	14-Aug-04	U	14-Aug-04	B	14-Jun-05	0	D
1	BM	Bruma	Matão	C	F	BR	N	14-Apr-06	12-Apr-06	17-Apr-06	U	17-Apr-06	B	17-Dec-08	0	O
1	BER	Bera	Matão	C	F	BR	N	16-Jun-08	15-Jun-08	16-Jun-08	U	16-Jun-08	B	17-Dec-08	0	O
1	BRL-J	Brasil	Jaó	C	M	BRAH	Y	19-Aug-02	18-Aug-02	20-Aug-02	N	20-Aug-02	B	17-Dec-08	0	O
1	BR	Bruna	Matão	U	F	BS	U	15-Mar-82	13-Dec-81	14-Jun-82	U	25-Jun-83	O	17-Dec-08	0	O
1	NI	Nilo	Matão	U	M	NY	U	15-Mar-82	13-Dec-81	14-Jun-82	U	25-Jun-83	O	17-Dec-08	0	O
2	DOT	DOTTY	1	C	F	ALT	N	21-Jun-73	21-Jun-73	21-Jun-73	N	21-Jun-73	B	25-Feb-01	0	D
2	DUD	DUDU	1	C	F	DOT	N	5-Jul-83	5-Jul-83	5-Jul-83	N	5-Jul-83	B	18-Jun-05	0	D
2	DRO	DRONGO	1·1	C	F	DUD	Y	13-Sep-89	13-Sep-89	13-Sep-89	Null	13-Sep-89	B	31-Dec-08	0	O
2	DEL	DELTA	1·1	C	M	DUD	N	2-Aug-91	2-Aug-91	2-Aug-91	Null	2-Aug-91	B	2-Apr-96	0·027	P
2	LIB	LIBERTY	2	C	M	LEL	N	28-Sep-92	28-Sep-92	28-Sep-92	N	28-Sep-92	B	9-Nov-06	0	P
2	NUT	NUTTY	1·1	C	F	NAD	Y	17-Aug-95	17-Aug-95	17-Aug-95	N	17-Aug-95	B	26-Dec-06	0	O
2	NJU	NJUGU	2·1	C	M	NIG	N	29-Oct-97	29-Oct-97	29-Oct-97	N	29-Oct-97	B	27-Dec-06	0	O
2	ALT	ALTO	1	C	F	Null	U	23-Feb-59	24-Feb-55	22-Feb-63	N	1-Aug-71	O	21-May-76	0	D
2	STU	STU	1	C	M	Null	U	1-Jul-67	1-Jul-66	30-Jun-68	N	1-Aug-71	O	9-Nov-87	0	P
2	KUS	KUSH	4	C	M	Null	U	21-Aug-69	22-Aug-67	21-Aug-71	N	25-Dec-78	I	30-Jun-90	0	D
2	ROC	ROCKY	9	C	M	Null	U	19-Oct-84	20-Oct-81	19-Oct-87	N	21-Dec-93	I	5-Jul-03	0·022	P
2	PRU	PRUDY	2	C	F	PIN	N	31-Jan-82	31-Jan-82	31-Jan-82	N	31-Jan-82	B	3-Jul-04	0	D
3	Ambo	Amboseli	Twn	C	F	Ambe	N	8-Feb-05	8-Feb-05	8-Feb-05	U	8-Feb-05	B	28-Nov-06	0	D
3	Diam	Diamond	Tw	C	F	Dd	N	29-Aug-95	29-Aug-95	29-Aug-95	U	29-Aug-95	B	31-Dec-08	0	O
3	Worm	Worm	G	C	F	HookF	N	15-Feb-94	31-Jan-94	1-Mar-94	U	19-Aug-94	B	28-Oct-95	0·63	O
3	Linc	Lincoln	Gs	C	M	Lily	N	30-Apr-04	30-Apr-04	30-Apr-04	U	30-Apr-04	B	9-Jan-06	0	D
3	Mick	Mickey	Gn	C	M	Mixu	N	23-Apr-01	22-Apr-01	24-Apr-01	U	24-Apr-01	B	19-Jun-08	0	E
3	Clow	Clownface	T	C	M	Nit	Y	15-Feb-82	31-Jan-82	1-Mar-82	U	1-Mar-82	B	16-Aug-01	0	P
3	Rose	Rose	T	U	F	Null	U	15-Feb-65	18-Feb-55	13-Feb-75	U	25-Jan-80	C	21-Jul-92	0·17	D
3	Plat	Plato	Tw	C	M	Pans	Y	20-Mar-02	20-Mar-02	21-Mar-02	U	21-Mar-02	B	2-Mar-06	0	D
3	Blaz	Blaze	Tw	C	F	Ray	N	1-Mar-94	1-Feb-94	31-Mar-94	U	1-Aug-94	B	31-Dec-08	0	O
3	Tip	Tip	Gn	C	M	Tap	N	28-Jan-98	27-Jan-98	30-Jan-98	U	30-Jan-98	B	27-Jul-08	0	P
4	AQ	AQUA	MT	C	F	AP	N	4-Apr-92	16-Mar-92	22-Apr-92	N	21-Apr-92	B	25-Jun-03	Null	D
4	AL	ATLAS	KK	C	M	AT	Y	25-Sep-67	17-Sep-67	2-Oct-67	N	2-Oct-67	B	14-Jan-99	0	D
4	CAD	CADBURY	KK	C	M	CD	N	29-Oct-98	14-Oct-98	12-Nov-98	N	13-Nov-98	B	1-Jul-01	0·07	D
4	DM	DOMINIE	KK	C	F	DO	Y	1-Nov-72	17-Oct-72	15-Nov-72	N	15-Nov-72	B	12-Sep-95	0	E
4	FB	FABEN	KK	U	M	FLO	U	2-Jul-47	2-Jul-45	1-Jul-49	N	26-Jul-63	C	6-Sep-75	Null	D
4	FG	FIGAN	KK	U	M	FLO	N	2-Jul-53	2-Jul-52	2-Jul-53	U	17-May-63	C	20-Jul-82	Null	D
4	FF	FIFI	KK	U	F	FLO	N	2-Jul-58	2-Jul-57	2-Jul-59	N	2-Jul-63	C	1-Sep-04	0·04	D
4	FT	FLINT	KK	C	M	FLO	N	1-Mar-64	1-Mar-64	13-Apr-64	N	15-Apr-64	B	17-Sep-72	0	D
4	FM	FLAME	KK	C	F	FLO	N	24-Aug-68	24-Aug-68	24-Aug-68	N	25-Aug-68	B	2-Feb-69	0·02	D

Table 1. Continued

Study ID	Anim ID	Anim Name	BirthGroup	BG Certainty	Sex	Mom ID	First Born	Birthdate	BD Min	BD Max	BDDist	Entrydate	Entrytype	Departdate	Departdate Error	Depart type
4	JM	JIMI	KK	C	M	JO	N	30-Jun-81	11-Feb-81	15-Nov-81	N	17-Nov-81	B	23-Sep-85	0	P
4	GB	GOBLIN	KK	C	M	ML	Y	6-Sep-64	6-Sep-64	6-Sep-64	N	7-Sep-64	B	24-Aug-04	0	D
4	FLO	FLO	Null	C	F	Null	U	2-Jul-19	3-Jul-14	30-Jun-24	N	15-May-63	C	21-Aug-72	0	D
4	MB	MADAM B	Null	C	F	Null	U	2-Jul-45	3-Jul-40	1-Jul-47	U	15-Aug-63	C	19-Sep-75	0	D
4	JO	JOANNE	Null	C	F	Null	U	2-Jul-56	3-Jul-54	2-Jul-58	N	16-Nov-78	I	23-Sep-85	0	P
4	SW	SPARROW	Null	C	F	Null	U	2-Jul-58	2-Jul-56	1-Jul-60	N	15-Sep-71	I	31-Dec-08	0	O
5	MAG	Maggie	Group 5	C	F	EFF	N	15-Jun-80	31-May-80	30-Jun-80	U	30-Jun-80	B	31-Dec-08	0	O
5	SAM	Samvura	Group 5	C	M	MAG	Y	20-May-89	20-May-89	20-May-89	U	20-May-89	B	12-Oct-91	0	D
5	UMO	Umubano2	Group 5	C	M	MAG	N	15-Feb-93	31-Jan-93	1-Mar-93	U	1-Mar-93	B	27-Dec-93	0	D
5	RUK	Rukundo	Pablo	C	M	MAG	N	7-Nov-94	23-Oct-94	22-Nov-94	U	22-Nov-94	B	31-Dec-08	0	O
5	TMR	Twemere	Pablo	C	F	MAG	N	1-Dec-98	1-Dec-98	1-Dec-98	U	1-Dec-98	B	18-Mar-99	0·12	D
5	AFR	Afrika	Pablo	C	F	MAG	N	20-Apr-00	20-Apr-00	20-Apr-00	U	20-Apr-00	B	31-Dec-08	0	O
5	TBA	Turibamwe	Pablo	C	M	MAG	N	25-Sep-03	25-Sep-03	25-Sep-03	U	25-Sep-03	B	31-Dec-08	0	O
5	I61	Infant 61	Pablo	C	M	MAG	N	8-Oct-06	8-Oct-06	8-Oct-06	U	8-Oct-06	B	13-Nov-06	0	D
5	I71	Infant 71	Bwenge	C	U	MAG	N	23-Dec-07	23-Dec-07	23-Dec-07	U	23-Dec-07	B	24-Dec-07	0	D
5	I84	Infant 84	Bwenge	C	F	MAG	N	9-Dec-08	9-Dec-08	9-Dec-08	U	9-Dec-08	B	24-Dec-08	0	D
6	80	Null	VV	C	F	20	U	15-Jul-85	15-Jul-85	15-Jul-85	U	15-Jul-85	B	15-Sep-08	0·08	D
6	80–1	Null	VV	C	M	80	Y	15-Jul-90	15-Jul-90	15-Jul-90	U	15-Jul-90	B	14-Nov-90	0	D
6	199	Null	VV	C	M	80	N	15-Jul-92	15-Jul-92	15-Jul-92	U	15-Jul-92	B	11-Dec-97	0	D
6	80–2	Null	VV	C	U	80	N	15-Jul-94	15-Jul-94	15-Jul-94	U	15-Jul-94	B	15-Jul-94	0·08	D
6	80–3	Null	VV	C	U	80	N	15-Jul-95	15-Jul-95	15-Jul-95	U	15-Jul-95	B	22-Dec-95	0·08	D
6	80–4	Null	VV	C	U	80	N	3-Jul-96	3-Jul-96	3-Jul-96	U	3-Jul-96	B	3-Jul-96	0·08	D
6	314	Null	VV	C	F	80	N	1-Aug-97	1-Aug-97	1-Aug-97	U	1-Aug-97	B	15-Dec-08	0	O
6	467	Null	VV	C	M	80	N	7-Jul-01	7-Jul-01	7-Jul-01	U	7-Jul-01	B	15-Oct-08	0	O
6	483	Null	VV	C	M	80	N	23-Jul-04	23-Jul-04	23-Jul-04	U	23-Jul-04	B	15-Jul-07	0·08	D
6	86–1	Null	CS	C	U	86	Y	15-Jul-93	15-Jul-93	15-Jul-93	U	15-Jul-93	B	15-Jul-93	0·08	D
6	86–2	Null	PP	C	U	86	N	15-Jul-94	15-Jul-94	15-Jul-94	U	15-Jul-94	B	15-Jul-94	0·08	D
6	9146	Null	PP	C	M	86	N	19-Jul-95	19-Jul-95	19-Jul-95	U	19-Jul-95	B	15-Dec-08	0	O
6	86–3	Null	PP	C	U	86	N	30-Jul-96	30-Jul-96	30-Jul-96	U	30-Jul-96	B	30-Jul-96	0·08	D
6	86–4	Null	PP	C	U	86	N	15-Jul-97	15-Jul-97	15-Jul-97	U	15-Jul-97	B	8-Dec-97	0·08	D
6	86–5	Null	PP	C	U	86	N	9-Jul-99	9-Jul-99	9-Jul-99	U	9-Jul-99	B	9-Jul-99	0·08	D
6	403	Null	PP	C	F	86	N	17-Jul-00	17-Jul-00	17-Jul-00	U	17-Jul-00	B	15-Dec-08	0	O
6	86–6	Null	PP	C	U	86	N	11-Jul-01	11-Jul-01	11-Jul-01	U	11-Jul-01	B	31-Jan-02	0·08	D
6	497	Null	PP	C	F	86	N	15-Jul-02	15-Jul-02	15-Jul-02	U	15-Jul-02	B	15-Dec-08	0	O
6	86–7	Null	PP	C	U	86	N	31-Jul-03	31-Jul-03	31-Jul-03	U	31-Jul-03	B	31-Jul-03	0·08	D
6	576	Null	PP	C	M	86	N	10-Jul-05	10-Jul-05	10-Jul-05	U	10-Jul-05	B	15-Dec-08	0	O
6	86–8	Null	PP	C	U	86	N	3-Aug-06	3-Aug-06	3-Aug-06	U	3-Aug-06	B	2-Oct-06	0·08	D
7	CT–	Chutney	LV	C	F	BLAN	N	2-Aug-99	2-Aug-99	2-Aug-99	N	2-Aug-99	B	31-Dec-08	0	O
7	KATH	Kathy Lee	LV	C	F	GR–	U	1-Apr-89	30-Dec-88	1-Jul-89	N	25-Mar-90	O	31-Dec-08	0	O
7	ALIE	Alien	LV	C	M	KATH	Y	2-Jan-96	2-Jan-96	2-Jan-96	N	15-Jan-96	B	20-Feb-04	0·04	P
7	MAYO	Mayo	LV	C	F	KATH	N	2-May-98	2-May-98	2-May-98	N	2-May-98	B	15-Jul-04	0	D
7	NOSE	Nose	Null	Null	M	Null	U	1-Jan-83	1-Jan-81	31-Dec-84	U	15-Jan-93	I	16-Sep-07	0·04	P

Table 1. Continued

Study ID	Anim ID	Anim Name	BirthGroup	BG Certainty	Sex	Mom ID	First Born	Birthdate	BD Min	BD Max	BDDist	Entrydate	Entrytype	Departdate	Departdate Error	Depart type
7	SUND	Sundance	Null	Null	M	Null	U	1-Jan-83	1-Jan-81	31-Dec-84	U	15-Feb-93	I	15-Sep-99	0·08	E
7	DARK	Dark	Null	Null	F	Null	U	1-Jan-87	1-Jan-86	1-Jan-88	N	15-Jan-97	I	27-May-98	0·08	P
7	TRIC	Trickle	Null	Null	M	Null	U	1-Jan-89	2-Jan-88	1-Jan-90	N	15-Dec-96	O	9-Nov-02	0·12	P
7	WEIR	Weirdo	Null	Null	M	Null	U	1-Jan-94	2-Jan-92	1-Jan-96	U	15-Mar-04	I	31-Dec-08	0	O
7	CARM	Carmen	LV	C	F	SPLO	N	1-Jan-89	1-Oct-88	2-Apr-89	N	25-Mar-90	O	15-Jul-96	0·21	D

[1]The rows in this table represent a subset of data from the Primate Life History Database (PLHD) Biography, and were selected to illustrate the range of variation in entries. For example, we show cases in which Birthdate is known exactly (e.g. StudyID = 1, AnimID = BRAH) or has a wide span due to different factors such as the researcher's inability to estimate Birthdate more precisely for an adult female who was present when the study began (e.g. StudyID = 1, AnimID = BS) or because the mother was rarely seen (e.g. StudyID = 4, AnimID = JIMI). Data are sorted here by StudyID, then MomID, then Birthdate, as described in the text. BG, BirthGroup; BD, Birthdate.

Table 2. Sample data from PLHD Fertility[1]

StudyID	AnimID	Startdate	Starttype	Stopdate	Stoptype
1	BM	17-Apr-06	B	17-Dec-08	O
1	BR	25-Jun-83	O	16-Jul-84	O
1	BR	28-Jun-86	C	17-Dec-08	O
1	BRAH	14-Aug-93	B	26-Jan-99	E
1	BRE	29-May-99	B	10-Apr-02	D
1	BRIS	20-Sep-89	B	13-Nov-95	E
1	BRN	20-Jul-02	B	20-Apr-04	D
1	NY	25-Jun-83	O	16-Jul-84	O
1	NY	28-Jun-86	C	17-Dec-08	O
2	ALT	1-Aug-71	O	21-May-76	D
2	DOT	21-Jun-73	B	31-Dec-89	O
2	DOT	1-Jan-96	O	25-Feb-01	D
2	DUD	5-Jul-83	B	31-Dec-89	O
2	DUD	8-Jun-90	O	5-Jul-90	O
2	DUD	2-Jan-95	O	17-Jan-95	O
2	DUD	1-Jan-96	O	18-Jun-05	D
2	NUT	17-Aug-95	B	26-Dec-06	O
2	PRU	31-Jan-82	B	3-Jul-04	D
3	Ambo	8-Feb-05	B	28-Nov-06	D
3	Blaz	8-Jun-97	O	31-Dec-08	O
3	Diam	8-Jun-97	O	31-Dec-08	O
3	Rose	25-Jan-80	C	11-Apr-81	O
4	AQ	21-Apr-92	B	25-Jun-03	D
4	DM	15-Nov-72	B	31-Jan-92	E
4	FLO	15-May-63	C	21-Aug-72	D
4	JO	23-Jul-80	I	11-Feb-81	O
4	JO	17-Nov-81	C	25-Mar-84	O
4	JO	11-Sep-85	C	23-Sep-85	P
4	MB	15-Aug-63	C	19-Sep-75	D
4	SW	15-Sep-71	I	31-Dec-08	O
5	AFR	30-Apr-00	B	31-Dec-08	O
5	MAG	30-Jun-80	B	17-Jun-97	O
5	MAG	29-Sep-98	O	31-Dec-08	O
6	80	15-Jul-85	B	15-Sep-08	D
6	86	15-Jul-85	B	15-Dec-08	O
7	CARM	25-Mar-90	O	1-Jul-91	O
7	CARM	1-Feb-92	O	1-Oct-95	O
7	CARM	1-Dec-95	O	15-Jul-96	D
7	CT–	2-Aug-99	B	1-Oct-99	O
7	CT–	1-Jan-00	O	1-Aug-00	O
7	CT–	1-Jan-01	O	1-Sep-04	O
7	CT–	1-Jan-04	O	31-Dec-08	O
7	DARK	15-Jan-97	I	1-Aug-97	O
7	DARK	1-Jan-98	O	27-May-98	P
7	KATH	25-Mar-90	O	1-Jul-91	O
7	KATH	1-Feb-92	O	1-Oct-95	O
7	KATH	1-Dec-95	O	1-Aug-96	O
7	KATH	1-Dec-96	O	1-Aug-97	O
7	KATH	1-Jan-98	O	1-Oct-99	O
7	KATH	1-Jan-00	O	1-Aug-00	O
7	KATH	1-Jan-01	O	1-Sep-04	O
7	KATH	1-Jan-05	O	31-Dec-08	O
7	MAYO	2-May-98	B	1-Oct-99	O
7	MAYO	1-Jan-00	O	1-Aug-00	O
7	MAYO	1-Jan-01	O	15-Jul-04	D

[1]The rows in this table represent a subset of data from the Primate Life History Database (PLHD) Fertility, and were selected to illustrate the range of variation in entries. For example, we show cases in which a female's fertility was monitored from her birth to death without interruptions (e.g. StudyID = 2, AnimID = PRU) or with interruptions (e.g. StudyID = 7, AnimID = MAYO).

But for those offspring to represent all of the female's offspring requires that she be born and die during the study and that her fertility be observed throughout her entire lifetime. This will not be true if the female's reproductive period: (i) was right-censored (i.e. observation of her reproduction ended before she died); (ii) was left-censored (i.e. the beginning portion of her reproductively mature period was not observed) or (iii) includes intermediate gaps.

Right-censoring of reproduction occurs for animals still alive at the most recent census date (indicated by a Departtype of 'O' and a Departdate at the most recent census date for a study), such as AnimID = BR in Study 1 or MAG in Study 5, who may still give birth to more offspring in the future. Left-censoring of reproduction occurs for females who were present at the beginning of the study and were judged to have been old enough at that time that they might have given birth to offspring before the study began. One example is AnimID = Rose in Study 3, whose 'Start Type' in FERTILITY is 'C'. Another example is AnimID = FLO in Study 4. None of FLO's five offspring in BIOGRAPHY are listed as her firstborn because it could not be determined whether she had given birth prior to her entry into the study. Note that FLO has a wide range of possible birth dates, as is frequently true of individuals such as FLO, who entered the study as an adult with confirmed AnimID and Entrytype = C, and of individuals, such as AnimID = DARK in Study 7, who entered the study by immigration (Entrytype = I).

Finally, lifetime reproductive success cannot be determined with certainty for females with gaps during which their fertility was not observed, as indicated by multiple entries for a female in FERTILITY (e.g. AnimID = KATH in Study 7). An example of a female whose entire reproductive output is captured in the database is AnimID = 80 in Study 6, who was born and died during the study, has an identified firstborn (80–1), and has a single entry in FERTILITY that begins with her birth and ends with her death (Table 2).

INTERBIRTH INTERVAL

Like lifetime reproductive success, determining the interval between successive births requires information from both BIOGRAPHY and FERTILITY. We first sort BIOGRAPHY by 'StudyID', 'MomID', and 'BirthDate', in that order. All of the recorded offspring of a female will now appear in birth order as a group of successive rows in BIOGRAPHY with the same entry in the 'MomID' column. But to be sure that adjacent rows represent successive siblings, we must first check to see that the two offspring were born in the same fertility interval for the mother (otherwise, additional offspring might have been born during a period when the mother's fertility was not being observed, and thus be missing from BIOGRAPHY). We determine this by searching over all rows for the mother (by her AnimID) in FERTILITY, checking to see if the birthdates of both offspring are after the 'Startdate' and before the 'Stopdate' of the same fertility interval. If so, then the interbirth interval is simply the difference between the birth dates of the two (now confirmed to be successive) offspring. An example of two successive offspring are BRS and BRE from Study 1, who were born during their mother's (BR's) second fertility interval (Table 2). The interval between their births is 1056 days or 2·89 years. In contrast, we cannot reliably compute an interbirth interval using the birth dates of DRO and DEL from Study 2, even though they are the two offspring of DUD with the closest birth dates in BIOGRAPHY, because they were not born during the same fertility interval for DUD (DRO was born in DUD's first interval in FERTILITY, but DEL was born during a period when DUD's fertility was not being intensively monitored; that is, it does not fall into any of DUD's fertility intervals in FERTILITY).

Discussion

The major strength of the PLHD resides in the high quality data that result from the development of the common vocabulary for the various data types, and the consequent standardization of variable data from multiple long-term field studies. The result is a truly comparative database, in which terms are defined by common criteria across disparate species and studies. Comparative databases with well-defined common vocabularies are now well-developed and heavily used in genomics research, and these often represent large-scale collaborative efforts. However, such collaborative efforts to develop common vocabularies and shared databases are relatively unusual for studies of life history and behaviour (Nelson 2009). A common vocabulary and standardization are necessary for comparative analyses that focus on questions pertaining to the actual variation in primate life histories instead of on the potential sources of error that can arise when divergent data sets without such stringent standardization are compared.

The design of the PLHD facilitates routine updates and management of the data that populate the current database. Currently, our Working Group members have agreed that we will notify the rest of the group whenever we update our data, which we can do at any time although annual updates are probably most useful. The Departdate associated with any animal for which Departtype = O corresponds to the last date on which data for that individual were updated, making it easy to keep track of updates.

The PLHD also has the potential to expand with comparable life history data from additional studies over time, provided that the data are individual-based and can conform to the common vocabulary that we have developed. For example, the criteria we established for the continuity of observations in FERTILITY may be prohibitive for researchers whose data rely almost exclusively on annual censuses, unless the study species are exclusively and narrowly seasonal breeders, as is the case for the sifaka in our database. Nonetheless, researchers working on other populations or taxa may find our BIOGRAPHY view to be a model for organizing data. The minimal criteria for conforming to the model are identifiable individuals censused repeatedly in a study population. Even data collected during single annual censuses can conform to this model.

The structure of both FERTILITY and BIOGRAPHY can be adapted with some minor adjustments to address different kinds of questions for different kinds of animals. For example,

FERTILITY is designed to identify gaps in observations or other circumstances that affect the probability of detecting an event during a particular time period for a particular animal. In demographic studies, these will usually be fertility events, but the general approach could be extended to other types of events. Similarly, although the structure of BIOGRAPHY is best suited for species that produce a single young with each birth, species that routinely produce litters can be accommodated by assigning each individual a separate row, similar to our treatment of the occasional cases of twins in the PLHD. Among animals that rear their young in crèches or dens, the number of live births may not be known and the closest equivalent to our criterion of 'live birth' for inclusion in the database might be 'emerging young'. Communal rearing could present a greater challenge if mother–offspring relationships cannot be assigned as reliably as they can for animals that typically produce single offspring. Additional columns could be included to differentiate individuals within and between litters in these cases.

The creation of the PLHD represents an essential positive response to the increasingly important challenge of developing collaborative models of data sharing. Such models must encourage researchers to share their data with others in a manner that enhances scientific progress, but at the same time protects the researchers' interests and their ability to continue to obtain funding for their research. This is particularly important in the case of long-term field data that require decades to collect. For these types of data, continued funding (typically awarded in 3–5 year increments at most) depends heavily upon the researchers' ability to argue for the novelty and uniqueness of the hypotheses they will be able to test (and hence on their ability to restrict access to their unpublished data). Thus, two important components of promoting a 'culture' of collaboration between data producers and data users (e.g. The Toronto Statement; Toronto International Data Release Workshop 2009) will be (i) recognizing that it will be important for data producers to maintain some control over access to the data they produce, and (ii) commitments on the part of funding agencies to both data producers and data users to provide long-term support of the databases themselves.

The longitudinal ecological data that comprise the PLHD have taken decades to accumulate. The PLHD provides a secure repository for the preservation and management of these irreplaceable data. Preserving these data for posterity is not only important for endangered species, but also for all populations whose ecosystems are under increasing pressures from encroaching human activities and are being altered by global climate change.

Acknowledgements

We are grateful to both the National Evolutionary Synthesis Center (NESCent) and the National Center for Ecological Analysis and Synthesis (NCEAS) for jointly funding our project on the Evolutionary Ecology of Primate Life Histories, and for hosting additional meetings of our working group to refine and work with the PLHD in January 2009 and November 2009, respectively. We also thank all of the government and funding agencies that have provided permissions and financial support for our field studies, and the many students, colleagues and field assistants and data technicians who have contributed to the collection and management of the long-term data now preserved in the PLHD. Study-specific acknowledgments can be found at http://demo.plhdb.org. We appreciate the comments of anonymous reviewers on the first version of this manuscript.

References

Arzberger, P., Schroeder, P., Beaulieu, A., Bowker, G., Casey, K., Laaksonen, L., Moorman, D., Uhlir, P. & Wouters, P. (2004) An international framework to promote access to data. *Science*, **303**, 1777–1778.

Cook, R.B., Olson, R.J., Kanciruk, P. & Hook, L.A. (2001) Best practices for preparing ecological data sets to share and archive. *Bulletin of the Ecological Society of America*, **82**, 138–141.

Cords, M. & Chowdhury, S. (in press) Life history of blue monkeys (*Cercopithecus mitis stuhlmanni*) in the Kakamega Forest, Kenya. *International Journal of Primatology*.

Date, C.J. (2004) *An Introduction to Database Systems*, 8th edn. Addison-Wesley, Reading, MA.

Ellison, A.M., Osterweil, L.J., Clarke, L., Hadley, J.L., Wise, A., Boose, E., Foster, D.R., Hanson, A., Jensen, D., Kuzeja, P., Riseman, E. & Schultz, H. (2006) Analytic webs support the synthesis of ecological data sets. *Ecology*, **87**, 1345–1358.

Even, A., Shankaranarayanan, G. & Watts, S. (2006) Enhancing decision making with process metadata: theoretical framework, research tool, and exploratory examination, pp. 209a. Proceedings of the 39th Annual Hawaii International Conference on System Sciences (HICSS'06) Track 8. Available at: http://www2.computer.org/portal/web/csdl/doi/10.1109/HICSS.2006.152 (accessed 27 December 2009).

Goodall, J. (1986) *The Chimpanzees of Gombe: Patterns of Behavior*. Belknap Press, Cambridge, MA.

Jones, O.R., Clutton-Brock, T., Coulson, T. & Godfray, H.C.J. (2008) A web resource for the UK's long-term individual based time-series (LITS) data. *Journal of Animal Ecology*, **77**, 612–615.

Kent, W. (1983) A simple guide to five normal forms in relational database theory. *Communications of the Association for Computing Machinery*, **26**, 120–125.

Michener, W.K. (2006) Meta-information concepts for ecological data management. *Ecological Informatics*, **1**, 3–7.

Mudakikwa, A.B., Cranfield, M.R., Sleeman, J.M. & Eilenberger, U. (2001) Clinical medicine, preventive health care and research on mountain gorillas in the Virunga Volcanoes region. *Mountain Gorillas: Three Decades of Research at Karisoke* (eds M.M. Robbins, P. Sicotte & K.J. Stewart), pp. 341–360. Cambridge University Press, Cambridge.

Nelson, B. (2009) Data sharing: empty archives. *Nature*, **461**, 160–163.

Nogueira, C., Carvalho, A.R., Oliveira, L., Veado, E.M. & Strier, K.B. (1994) Recovery and release of an infant muriqui, *Brachyteles arachnoides*, at the Caratinga Biological Station, Minas Gerais, Brazil. *Neotropical Primates*, **2**, 3–5.

Parr, C.S. & Cummings, M.P. (2005) Data sharing in ecology and evolution. *Trends in Ecology and Evolution*, **20**, 362–363.

Piwowar, H.A., Becich, M.J., Bilofsky, H. & Crowley, R.S. (2008) Towards a data sharing culture: recommendations for leadership from academic health centers. *PLoS Medicine*, **5**(9), e183. doi:10.1371/journal.pmed.0050183.

Pusey, A.E., Wilson, M.W. & Collins, D.A. (2008) Human impacts, disease risk, and population dynamics in the chimpanzees of Gombe National Park, Tanzania. *American Journal of Primatology*, **70**, 738–744.

Schofield, P.N., Bubela, T., Weaver, T., Portilla, L., Brown, S.D., Hancock, J.M., Einhorn, D., Tocchini-Valentini, G., Hrabe de Angelis, M., Rosenthal, N. & CASIMIR Rome Meeting participants. (2009) Post-publication sharing of data and tools. *Nature*, **461**, 171–173.

Strier, K.B., Alberts, S., Wright, P.C., Altmann, J. & Zeitlyn, D. (2006) Primate life history databank: setting the agenda. *Evolutionary Anthropology*, **15**, 44–46.

Toronto International Data Release Workshop. (2009) Prepublication data sharing. *Nature*, **461**, 168–170.

Wrangham, R.W. (1974) Artificial feeding of chimpanzees and baboons in their natural habitat. *Animal Behaviour*, **22**, 83–93.

Received 4 March 2010; accepted 8 March 2010
Handling Editor: Robert P. Freckleton

Supporting Information

Additional Supporting Information may be found in the online version of this article.

Table S1. Biography structure, with each variable, content, and coding options.

Table S2. Fertility structure, with each variable, content, and coding options.

Table S3. Study Population structure, with each variable, content, and coding options.

As a service to our authors and readers, this journal provides supporting information supplied by the authors. Such materials may be re-organized for online delivery, but are not copy-edited or typeset. Technical support issues arising from supporting information (other than missing files) should be addressed to the authors.