BIOLOGICAL REVIEWS
Cambridge Philosophical Society
Biol. Rev. (2017), pp. 000–000.
doi: 10.1111/brv.12359

Building essential biodiversity variables (EBVs) of species distribution and abundance at a global scale

W. Daniel Kissling[1,*], Jorge A. Ahumada[2], Anne Bowser[3], Miguel Fernandez[4,5,6], Néstor Fernández[4,7], Enrique Alonso García[8], Robert P. Guralnick[9], Nick J. B. Isaac[10], Steve Kelling[11], Wouter Los[1], Louise McRae[12], Jean-Baptiste Mihoub[13,14], Matthias Obst[15,16], Monica Santamaria[17], Andrew K. Skidmore[18], Kristen J. Williams[19], Donat Agosti[20], Daniel Amariles[21,22], Christos Arvanitidis[23], Lucy Bastin[24,25], Francesca De Leo[17], Willi Egloff[20], Jane Elith[26], Donald Hobern[27], David Martin[19], Henrique M. Pereira[4,5], Graziano Pesole[17,28], Johannes Peterseil[29], Hannu Saarenmaa[30], Dmitry Schigel[27], Dirk S. Schmeller[13,31], Nicola Segata[32], Eren Turak[33,34], Paul F. Uhlir[35], Brian Wee[36] and Alex R. Hardisty[37]

[1] *Department Theoretical and Computational Ecology, Institute for Biodiversity and Ecosystem Dynamics (IBED), University of Amsterdam, P.O. Box 94248, 1090 GE Amsterdam, The Netherlands*
[2] *TEAM Network, Moore Center for Science, Conservation International, 2011 Crystal Dr. Suite 500, Arlington, VA 22202, U.S.A.*
[3] *Woodrow Wilson International Center for Scholars, 1300 Pennsylvania Ave, NW Washington, DC 20004, U.S.A.*
[4] *Biodiversity Conservation Group, German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Deutscher Platz 5e, 04103 Leipzig, Germany*
[5] *Institute of Biology, Martin Luther University Halle-Wittenberg, Halle, Germany*
[6] *Instituto de Ecología, Universidad Mayor de San Andrés (UMSA), Campus Universitario, Cota cota, La Paz, Bolivia*
[7] *Estación Biológica de Doñana EBD-CSIC, Américo Vespucio s.n, 41092 Sevilla, Spain*
[8] *Councillor of State of the Kingdom of Spain and Honorary Researcher of the Franklin Institute of the University of Alcalá, Madrid, Spain*
[9] *University of Florida Museum of Natural History, University of Florida at Gainesville, Gainesville, FL 32611-2710, U.S.A.*
[10] *Biological Records Centre, Centre for Ecology & Hydrology, Maclean Building, Benson Lane, Crowmarsh Gifford, OX10 8BB Wallingford, U.K.*
[11] *Cornell Lab of Ornithology, Cornell University, 158 Sapsucker Woods Rd, Ithaca NY 14850, U.S.A.*
[12] *Institute of Zoology, Zoological Society of London, Regent's Park, NW1 4RY London, U.K.*
[13] *UPMC Université Paris 06, Muséum National d'Histoire Naturelle, CNRS, CESCO, UMR 7204, Sorbonne Universités, 61 rue Buffon, 75005, Paris, France*
[14] *Department of Conservation Biology, UFZ-Helmholtz Centre for Environmental Research, Permoserstr. 15, 04318 Leipzig, Germany*
[15] *Department of Marine Sciences, Göteborg University, Box 463, SE-40530 Göteborg, Sweden*
[16] *Gothenburg Global Biodiversity Centre, Box 461, SE-405 30 Göteborg, Sweden*
[17] *CNR-Institute of Biomembranes and Bioenergetics, Amendola 165/A Street, 70126 Bari, Italy*
[18] *Department of Natural Resources, Faculty of Geo-Information Science and Earth Observation (ITC), University of Twente, P.O. Box 217, 7500AE Enschede, The Netherlands*
[19] *Land and Water, Commonwealth Scientific and Industrial Research Organisation (CSIRO), PO Box 1600, Canberra, Australian Capital Territory 2601, Australia*
[20] *Plazi, Zinggstr. 16, 3007 Bern, Switzerland*
[21] *Decision and Policy Analysis (DAPA), International Center for Tropical Agriculture (CIAT), AA6713 Cali, Colombia*
[22] *Instituto Alexander von Humboldt, CALLE 28A # 15-09, Bogota D.C., Colombia*
[23] *Institute of Marine Biology, Biotechnology and Aquaculture, Hellenic Centre for Marine Research, Thalassokosmos, Former US Base at Gournes, 71003 Heraklion, Crete, Greece*
[24] *School of Engineering and Applied Science, Aston University, Aston Triangle, B4 7ET Birmingham, U.K.*

* Address for correspondence (Tel: +31 20 525 8423; E-mail: wdkissling@gmail.com).

[25] *Knowledge Management Unit, Joint Research Centre of the European Commission, Via Enrico Fermi, 21027 Varese, Italy*
[26] *School of BioSciences (Building 143), University of Melbourne, Melbourne, VIC 3010, Australia*
[27] *Global Biodiversity Information Facility Secretariat, Universitetsparken 15, 2100 København Ø, Denmark*
[28] *Department of Biosciences, Biotechnology and Biopharmaceutics, University of Bari "A. Moro", via Orabona 4, 70125 Bari, Italy*
[29] *Department for Ecosystem Research & Environmental Information Management, Umweltbundesamt GmbH, Spittelauer Lände 5, 1090 Vienna, Austria*
[30] *Department of Forest Sciences, University of Eastern Finland, Joensuu Science Park, Länsikatu 15, FI-80110 Joensuu, Finland*
[31] *ECOLAB, Université de Toulouse, CNRS, INPT, UPS, Toulouse, France*
[32] *Centre for Integrative Biology, University of Trento, Via Sommarive 9, 38123 Trento, Italy*
[33] *NSW Office of Environment and Heritage, PO Box A290, Sydney South, NSW 1232, Australia*
[34] *Australian Museum, 6 College Street, Sydney, NSW 2000, Australia*
[35] *Consultant, Data Policy and Management, P.O. Box 305, Callicoon, NY 12723, U.S.A.*
[36] *Massive Connections, 2410 17th St NW, Apt 306, Washington, DC 20009, U.S.A.*
[37] *School of Computer Science & Informatics, Cardiff University, Queens Buildings, 5 The Parade, Cardiff, CF24 3AA, U.K.*

ABSTRACT

Much biodiversity data is collected worldwide, but it remains challenging to assemble the scattered knowledge for assessing biodiversity status and trends. The concept of Essential Biodiversity Variables (EBVs) was introduced to structure biodiversity monitoring globally, and to harmonize and standardize biodiversity data from disparate sources to capture a minimum set of critical variables required to study, report and manage biodiversity change. Here, we assess the challenges of a 'Big Data' approach to building global EBV data products across taxa and spatiotemporal scales, focusing on species distribution and abundance. The majority of currently available data on species distributions derives from incidentally reported observations or from surveys where presence-only or presence–absence data are sampled repeatedly with standardized protocols. Most abundance data come from opportunistic population counts or from population time series using standardized protocols (e.g. repeated surveys of the same population from single or multiple sites). Enormous complexity exists in integrating these heterogeneous, multi-source data sets across space, time, taxa and different sampling methods. Integration of such data into global EBV data products requires correcting biases introduced by imperfect detection and varying sampling effort, dealing with different spatial resolution and extents, harmonizing measurement units from different data sources or sampling methods, applying statistical tools and models for spatial inter- or extrapolation, and quantifying sources of uncertainty and errors in data and models. To support the development of EBVs by the Group on Earth Observations Biodiversity Observation Network (GEO BON), we identify 11 key workflow steps that will operationalize the process of building EBV data products within and across research infrastructures worldwide. These workflow steps take multiple sequential activities into account, including identification and aggregation of various raw data sources, data quality control, taxonomic name matching and statistical modelling of integrated data. We illustrate these steps with concrete examples from existing citizen science and professional monitoring projects, including eBird, the Tropical Ecology Assessment and Monitoring network, the Living Planet Index and the Baltic Sea zooplankton monitoring. The identified workflow steps are applicable to both terrestrial and aquatic systems and a broad range of spatial, temporal and taxonomic scales. They depend on clear, findable and accessible metadata, and we provide an overview of current data and metadata standards. Several challenges remain to be solved for building global EBV data products: (*i*) developing tools and models for combining heterogeneous, multi-source data sets and filling data gaps in geographic, temporal and taxonomic coverage, (*ii*) integrating emerging methods and technologies for data collection such as citizen science, sensor networks, DNA-based techniques and satellite remote sensing, (*iii*) solving major technical issues related to data product structure, data storage, execution of workflows and the production process/cycle as well as approaching technical interoperability among research infrastructures, (*iv*) allowing semantic interoperability by developing and adopting standards and tools for capturing consistent data and metadata, and (*v*) ensuring legal interoperability by endorsing open data or data that are free from restrictions on use, modification and sharing. Addressing these challenges is critical for biodiversity research and for assessing progress towards conservation policy targets and sustainable development goals.

Key words: big data, biodiversity monitoring, data interoperability, ecological sustainability, environmental policy, global change research, indicators, informatics, metadata, research infrastructures.

CONTENTS

I. INTRODUCTION

The diversity of life on Earth is intrinsically and pragmatically essential, and provides vital services to humanity (Millennium Ecosystem Assessment, 2005). Despite recognition of this fact and ongoing conservation efforts, biodiversity continues to be lost globally at an alarming rate (Tittensor *et al.*, 2014). Current extinction rates of species may be 100 times higher than the 'background' rate from fossil records (Pereira, Navarro & Martins, 2012; Ceballos *et al.,* 2015).

Many populations of widespread and threatened species are declining (Butchart *et al.,* 2010; Tittensor *et al.,* 2014) and invasive alien species continue to spread into many parts of the world (van Kleunen *et al.,* 2015). Combined with human exploitation of terrestrial and marine ecosystems, these factors result in an Earth system that is stretched beyond sustainability (Newbold *et al.,* 2016). Reversing such trends is part of the focus of the 20 Aichi Targets developed by Parties to the United Nations (UN) Convention on Biological Diversity (CBD), and of the 17 Sustainable Development Goals (SDGs) identified by the UN 2030 Agenda for Sustainable Development.

Enormous challenges remain for global biodiversity conservation and ecological sustainability, even for simply assessing reliably the progress towards achieving Aichi Targets and SDGs, especially at a global scale. These include finding mechanisms to fill known data gaps (Meyer *et al.,* 2015; Skidmore *et al.,* 2015; Amano, Lamming & Sutherland, 2016), to standardize data and make them available and accessible (Reichman, Jones & Schildhauer, 2011), and to develop the technical tools and sustainable e-infrastructure that supports discovery, analysis, access, dissemination and persistent storage of the increasingly complex data sets needed to quantify biodiversity change at a global scale (Hardisty, Roberts & The Biodiversity Informatics Community, 2013; Hobern *et al.,* 2013; Kissling *et al.,* 2015; Hugo *et al.,* 2017).

To address these challenges, the Group on Earth Observations Biodiversity Observation Network (GEO BON) has introduced the framework of Essential Biodiversity Variables (EBVs) (Pereira *et al.,* 2013). EBVs can be considered to be biological state variables with three key dimensions (time, space, and biological organization) that are critical to document biodiversity change accurately (Schmeller *et al.,* in press). Moreover, EBVs represent harmonized data that are conceptually located on a continuum between primary data observations ('raw data') and synthetic or derived indices ('indicators') (Fig. 1). Similar to Essential Climate Variables – which are designed to provide an empirical basis for understanding past, current and possible future climate variability and change (Bojinski *et al.,* 2014) – the EBV framework has been developed to help in prioritizing a minimum set of essential measurements for the consistent study, reporting and management of the major dimensions of biodiversity change (Pereira *et al.,* 2013).

A total of 22 candidate EBVs are proposed by GEO BON within six EBV classes (i.e. genetic composition, species populations, species traits, community composition, ecosystem functioning and ecosystem structure) (Pereira *et al.,* 2013). These categories of critical biodiversity data provide the motivation and necessary framework for standardizing the global biodiversity data needed for research, management and policy (Geijzendorffer *et al.,* 2016; Proença *et al.,* in press). EBVs thus provide the foundation for consistent derivation of biodiversity indicators that allow repeated assessments of progress against national and global conservation targets and sustainability goals (Turak *et al.,* in press b; Pereira *et al.,* 2013). EBVs can contribute to a range of policy initiatives, such as the global and regional assessments conducted through the Intergovernmental Platform on Biodiversity and Ecosystem Services (IPBES, 2016) or annual reporting by countries to the CBD against their National Biodiversity Strategies and Action Plans (NBSAPs).

A major concern that arises from these efforts is how to build useful EBV data products with global coverage using current technology (Kissling *et al.,* 2015). From several perspectives, this is the 'Big Data' challenge in biodiversity science today (Hampton *et al.,* 2013). It requires dealing with massive volumes of data not readily handled by the usual data tools and practices, establishment of relationality between different data, ensuring data quality, aggregation, cross-referencing and making such data searchable and available (Kelling *et al.,* 2015; Enquist *et al.,* 2016; La Salle, Williams & Moritz, 2016; Wilkinson *et al.,* 2016). Ideally, the primary data required to build EBV data products should be contributed from any research or observation infrastructure, no matter at which spatial or temporal scale they had been collected. Multiple scientific, technical and legal challenges – such as the need for data harmonization and metadata standardization, provision of analytical tools and services for EBV data processing, and open access licenses that allow the interoperability and sharing of relevant data – have to be addressed to produce reliable EBV data products (Kissling *et al.,* 2015).

Quantifying and predicting variations in species distributions and population size is of high importance for biodiversity research, management and policy efforts. For instance, knowledge on the geographic distribution of species and variation in population structure and abundance is central to understanding ecological and biogeographical dynamics (Begon, Townsend & Harper, 2006; Lomolino *et al.,* 2010). Moreover, species distribution and abundance underpins policy indicators to quantify population trends and extinction risk for threat categorization (Butchart *et al.,* 2010), assessments of range dynamics (Schurr *et al.,* 2012), spread of invasive species (McGeoch *et al.,* 2010) and biodiversity responses to climate change (Stephens *et al.,* 2016) and habitat conversion (Newbold *et al.,* 2016). Within the EBV concept, such essential knowledge is captured in the EBV class 'species populations', represented by three candidate EBVs (Pereira *et al.,* 2013, 2017): 'species distribution', 'population abundance' and 'population structure'. Given the societal and scientific relevance of geographic data on the distribution and abundance of species (e.g. Butchart *et al.,* 2010; McGeoch *et al.,* 2010; Jetz, McPherson & Guralnick, 2012; Schmeller *et al.,* 2017) and a high maturity level of the research infrastructure and biodiversity monitoring community (e.g. Constable *et al.,* 2010; Ahumada, Hurtado & Lizcano, 2013; Hobern *et al.,* 2013; Sullivan *et al.,* 2014; La Salle *et al.,* 2016), it is important to explore which barriers and bottlenecks currently prevent the global implementation of species distribution and abundance EBVs.

Here, we review the wider scientific, technical and legal issues pertinent to building EBV data products at a global scale. We specifically focus on the candidate EBVs 'species

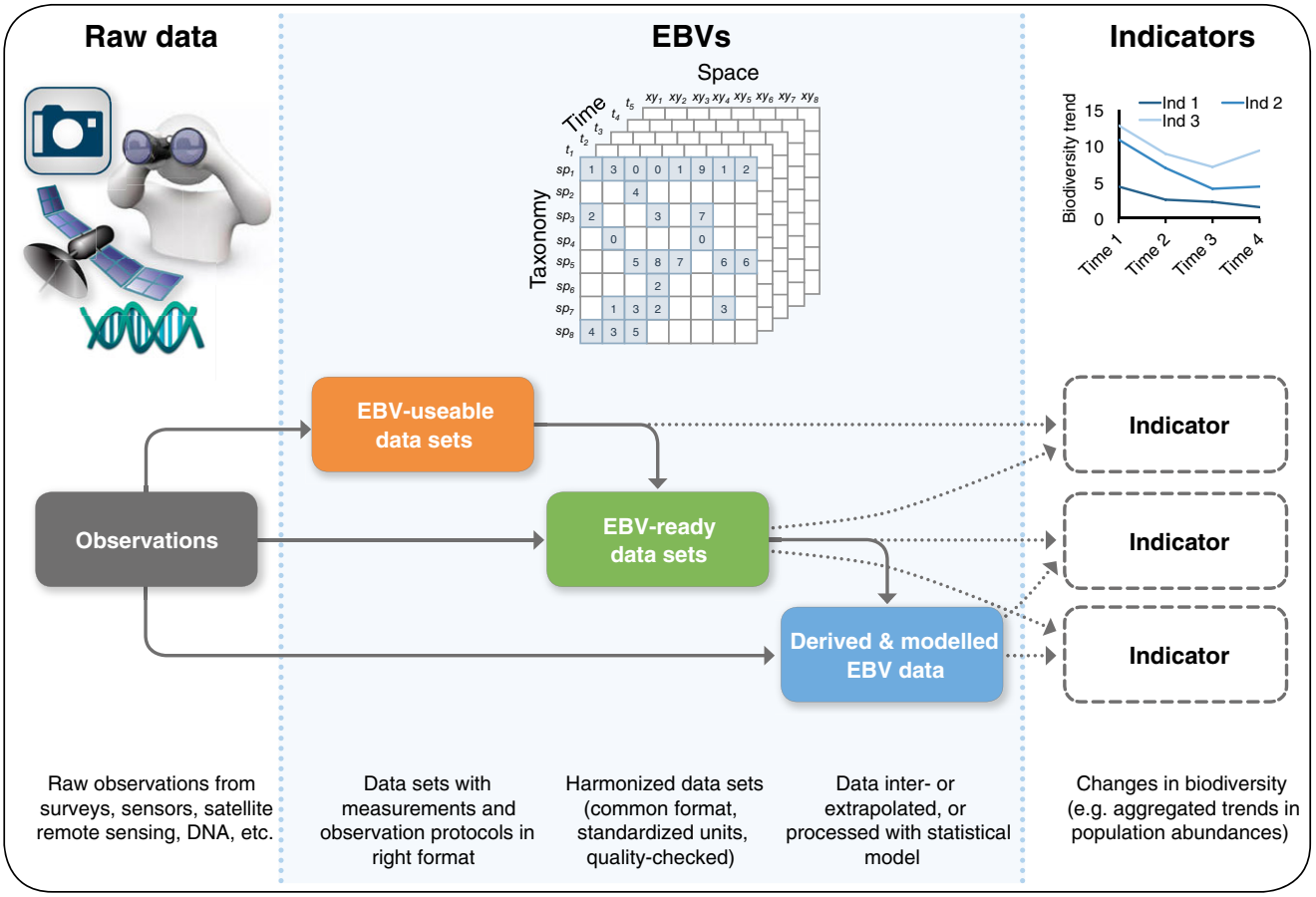

Fig. 1. Essential Biodiversity Variables (EBVs) are part of an information supply chain, conceptually positioned between raw data (i.e. primary data observations) and indicators (i.e. synthetic indices for reporting biodiversity change to policy and management). They can be illustrated as a data cube with three basic dimensions (taxonomy, time and space), covering different species (sp_1, sp_2, . . .) at different points in time (t_1, t_2, . . .) and different locations (xy_1, xy_2, . . .). From the observations (i.e. sampling of raw data), different EBV data products can be obtained with different steps of data processing. We here distinguish EBV-useable data sets, EBV-ready data sets and derived and modelled EBV data. They represent measurements with comparable measurement units or similar observation protocols (EBV-useable data sets), harmonized data sets (EBV-ready data sets) and data products derived from processing data with statistical models (derived and modelled EBV data). These EBV data products can be used in various ways to derive indicators (Ind 1, Ind 2, . . .) that quantify spatiotemporal changes in species distributions and population abundances or other aspects of biodiversity. The four images under raw data are freely available at http://www.clipartpanda.com

distribution' and 'population abundance' in the EBV class 'species populations'. We start by defining the two selected EBVs and their key dimensions (space, time and taxonomy), attributes (extent, resolution, measurement unit) and uncertainties. We then provide a brief overview of existing sources of species distribution and abundance data, discuss key requirements for data harmonization and highlight emerging methods and technologies for data collection. We propose 11 key workflow steps for building EBV data products on species distributions and population abundances and highlight legal and technical barriers for a workflow-oriented production of EBV data products. We summarize continuing efforts to develop relevant metadata standards for EBV workflows and outline how metadata standards for EBV data products could be developed. Finally, we conclude with the current challenges for building global EBV data products and summarize scientific, technical and legal constraints on implementing EBV workflows at a global scale.

II. EBV DEFINITION

(1) The species distribution EBV

The EBV 'species distribution' can be defined as the presence or absence of species, based on observations with specified spatial and temporal dimensions. In most cases, the species distribution EBV is therefore represented through a binary variable that reflects presence–absence of a species across its geographic range. Beyond a binary quantification, species distributions can be estimated by using species distribution models (SDMs) to predict relative likelihoods, probabilities of

observation or probabilities of occupancy, dependent on the available data (MacKenzie *et al.*, 2006; Elith & Leathwick, 2009; Guillera-Arroita *et al.*, 2015).

(2) The population abundance EBV

The candidate EBV 'population abundance' can be defined as population sizes based on observations with specified spatial and temporal dimensions. The population abundance EBV is therefore represented by a continuous variable expressing harmonized quantities for each taxon across space and time. Depending on the available type of raw data, methods mostly estimate relative abundance or relative density, but others may be useful for estimating actual densities (e.g. counts per unit area). Interpolation of localized measurements with geostatistical methods or modelling of abundance across a species' geographic range with environmental covariates provides ways to obtain spatially explicit representations of population abundance (Potts & Elith, 2006).

(3) Relationship between species distribution and population abundance EBVs

The species distribution EBV and the population abundance EBV are closely related to each other. The population abundance EBV contains richer information than the species distribution EBV, i.e. a continuous variable that describes not only the presence of populations of a species but also the population size per unit area throughout its geographic range. Species distribution data are usually much easier to collect than population abundance data because they only require the recorded presence of one individual rather than estimates of the absolute or relative number of individuals. Hence, only a few large-scale and long-term monitoring projects, such as the North American Breeding Bird Survey (Sauer *et al.*, 2013) or the Pan-European Common Birds Monitoring Scheme (Stephens *et al.*, 2016), have measured annual trends in species abundances at a continental scale.

Many data sets that are potentially relevant to producing the population abundance EBV only exist as time series for one or more discrete (local) populations. Such data are useful for the development of biodiversity indicators, like the aggregated population trends among vertebrate species used in the Living Planet Index (Loh *et al.*, 2005; Collen *et al.*, 2009). Even if these data are confined to a few locations, it is crucial to identify such cases clearly because they might be useful for building EBV data products.

In the following section, we explain how the EBV framework can be conceptualized and which dimensions, attributes and uncertainties are relevant for building EBV data products. We further highlight ideal *versus* minimum requirements of EBV data products and show examples of projects that have relevant data for building EBV data products on species distributions and population abundances.

III. OPERATIONALIZING THE EBV FRAMEWORK

(1) From raw data to indicators

A fundamental tenet of the EBV framework is that all EBV data sets lie conceptually between raw data and indicators (Fig. 1). For a given EBV, raw data can come from diverse sources, including field observation campaigns (Dickinson, Zuckerberg & Bonter, 2010; Proença *et al.*, in press), *in situ* sensor networks (Porter *et al.*, 2005), remote sensing (Skidmore *et al.*, 2015; Lausch *et al.*, 2016) and DNA sequencing (Creer *et al.*, 2016). From those raw data, several EBV data products can be built (Fig. 1). First, data that use observation protocols to measure relevant phenomena with comparable units are identified ('EBV-useable data sets'). Multiple data sets can then be combined and harmonized to a common format with standardized units, having been quality-checked and error corrected ('EBV-ready data sets'). Data gaps in space or time (illustrated as empty fields in the data cube of Fig. 1) could be filled by applying statistical techniques for inter- or extrapolation ('derived and modelled EBV data'), as well as through targeted future sampling.

Depending on the nature of the raw data, not all of these processing steps may be needed to build a specific EBV. Nevertheless, the aggregated, harmonized and modelled EBV data products should allow derivation of indicators of the state of biodiversity and estimation of temporal changes in critical aspects of biodiversity (Butchart *et al.*, 2010; McGeoch *et al.*, 2010; Pereira *et al.*, 2013; Tittensor *et al.*, 2014).

(2) Dimensions, attributes and uncertainties of EBVs

Building EBVs with heterogeneous types of data requires identification and clear definition of the key dimensions, attributes and uncertainties of EBV-useable data sets, EBV-ready data sets and derived and modelled EBV data. Three basic dimensions are of particular importance: space, time and taxonomy (Fig. 1). Some of this dimensionality can be represented in more than one axis. For instance, the spatial dimension can be represented by latitude, longitude, water depth and altitude. The dimensions and their axes can form a data cube (Fig. 1). The data cube can be conceptually useful to encapsulate a multidimensional view of a specific EBV (Schmeller *et al.*, in press).

The three dimensions (space, time, taxonomy) can be specified with attributes related to the extent, resolution and measurement unit along which the dimension is expressed (Table 1). For instance, the extent may simply be the spatial and temporal coverage of records across sampling locations, or how many and which species are documented in the data cube (Meyer, Weigelt & Kreft, 2016). Resolution might refer to the spatial and temporal grain size and the taxonomic resolution of the data (Table 1). For instance, the spatial resolution of abundance data refers to a discrete point, a study area, or a volume (e.g. from water samples); the temporal resolution is associated with the periodicity of

Table 1. Examples of key dimensions, attributes and uncertainties related to Essential Biodiversity Variables (EBVs) of species distribution and population abundance

Dimension	EBV attributes			Uncertainties
	Extent	Resolution	Measurement units	
Space	Geographical coverage (e.g. of grid cells, sampling locations, satellites, etc.)	Spatial resolution (e.g. grid cell size, polygons, resolution of satellite sensors, volume, etc.)	Meters, cubic meters, kilometers, degrees, etc.	Precision and accuracy of coordinates and volumes, wrongly recorded coordinates, imprecise sampling locations
Time	Temporal coverage (e.g. length of time series, continuous recording, time period of collection of records, etc.)	Temporal grain (e.g. date or time window of sampling, sampling frequency)	Hours, days, weeks, months, years, decades, etc.	Variation in length of time series, precision of time of collection, etc.
Taxonomy	Taxonomic coverage (e.g. how many and which species are documented)	Species, genus, higher taxonomic level, etc.	Taxonomic entity for which species distribution and abundance data are sampled	Identification and observation uncertainty, ambiguous scientific names, synonyms, differences in taxon concepts, etc.

monitoring; and the taxonomic resolution reflects at which taxonomic level data are collected, e.g. at the species, genus or a higher taxonomic level. Measurement units refer to the quantities that are expressed, such as kilometres, days, number of individuals, or which taxonomic entities are chosen and according to which taxon concept. Such attribute information should ideally be recorded in the metadata associated with the raw data (discussed further in Section VI).

Each of the three attributes (extent, resolution and unit) of the three dimensions (space, time and taxonomy) further comes with uncertainties related to the spatial, temporal and taxonomic information that makes up the EBV data cube (Table 1). For instance, imprecise or faulty geo-referencing of collection localities and an outdated taxonomy or incorrect specimen identification increases uncertainty by decreasing the precision and accuracy of geographical and taxonomic information on species occurrences (Meyer *et al.,* 2016). Similarly, biodiversity monitoring projects across the world vary tremendously in observation efforts per site, sampling frequencies and length of time series, which results in large temporal uncertainties when analysed jointly (Proença *et al.,* in press). For EBVs, uncertainties should be quantified in as much detail as possible. Identified gaps and biases could guide national and international efforts for mobilizing new distribution and population abundance data sets (Meyer *et al.,* 2015; Amano *et al.,* 2016; Proença *et al.,* in press).

(3) Ideal *versus* minimum requirements of EBV data products

Building species distribution and abundance EBVs might require multiple data products (e.g. for specific taxa, regions, or time frames). Ideally, an EBV data product should contain consistent quantitative measurements or estimates across space and time, allowing a genuine comparison of changes in species populations over regional to continental extents and from years to decades. As such, an ideal EBV product would derive from consistent observations at regular intervals collected across an optimally designed configuration of sample locations, allowing conclusions across a range of spatial and temporal scales. However, such ideal data sets do not exist, and given the variation in the design of sampling programs, as well as legislative, social and political constraints, it is unlikely that an equal monitoring effort across taxa, ecoregions or jurisdictional boundaries can ever be achieved (Proença *et al.,* in press; Turak *et al.,* in press a). It is therefore important to define how ideal requirements for an EBV data product differ from what is minimally required (Table 2).

The aim of defining minimum requirements is to provide guidelines about which data sets are useful in the context of EBVs. For instance, from field observation campaigns one would ideally like to have presence–absence or density estimates from standardized sampling with global coverage at fine resolution derived from continuous long-term time series with the highest adequate temporal resolution (Table 2). However, presence-only or relative abundance estimates might be the only data available across a large spatial and temporal extent (Table 2). Hence, the ideal requirements might only be achievable for a few selected taxa and regions. If ideal requirements are not met, it might be necessary to systematically select a subset of the available data. For example, if data are unavailable across the full geographic range of a species, it may be preferable to track changes in biodiversity at smaller geographic extents, e.g. within particular ecoregions (Turak *et al.,* in press a). More generally, identifying ideal and minimum requirements (e.g. Table 2) can serve as a benchmark to evaluate which existing data sets are appropriate for building EBV data products and to report on the relative value of each product to assess biodiversity change. To date, no guidelines exist for how minimum requirements for EBV data products should be defined. We therefore suggest that GEO BON and the wider

Table 2. Ideal *versus* minimum requirements of Essential Biodiversity Variables (EBVs) data in relation to species distribution and abundance measurements and their spatial, temporal and taxonomic extent and resolution

EBV dimensions and attributes	Ideal requirements	Minimum requirements
Species distribution and abundance measurements	Presence–absence and density of individuals derived from widely accepted, standardized or explicit sampling protocols, including quantification of uncertainty and recording sampling covariates	Presence-only or (relative, qualitative or ordinal) density estimates across space and time based on raw observations or modelling but only if derived from widely accepted protocols with consistency across space/time
Spatial extent	Global coverage, with the capacity to provide high-quality information for global assessments (e.g. Aichi targets, Sustainable Development Goals)	Adequate spatial coverage to provide reliable information on biodiversity trends for policy decision making (e.g. at regional or national level)
Spatial resolution	Fine-scale estimates of population abundance across subnational (e.g. a protected area), national, regional/continental and global extent	Statistically driven design that allows combining scattered, high-quality information at the scale of policy or management interest (e.g. national extent)
Temporal extent	Continuous long-term time series of abundance or occupancy over several decades suitable to assess potential biodiversity change	Repeated measurements at policy-relevant time intervals to differentiate between fluctuations and trends, including a baseline
Temporal resolution	Temporal resolution (hours, days, weeks, months, years) that is adequate to detect population dynamics for a specific taxon	Reliable species distribution and abundance estimates for at least two time slices at the same spatio-temporal resolution, with relevance to policy and/or management
Taxonomic extent	Maximum possible number of species covering a wide variety of taxa and life forms, and providing information on various dimensions of global change and different ecosystem functions and services	Selected species representing particular taxonomic or functional groups, representative of overall diversity and environments within spatial extent
Taxonomic resolution	Updating compilations of taxonomic names and associated concepts of all species and their synonyms	Clearly defined taxonomic units following known taxonomic authorities

scientific community further develops our recommendations on minimum requirements of EBV data products.

(4) Examples of projects with EBV-relevant data products

Due to the vast variability in spatial, temporal and taxonomic resolution and extent, as well as measurement units considered by different species distribution and abundance data sets, there is currently no comprehensive global database that fulfils all ideal requirements for EBV data products (see Table 2). As an alternative, different species distribution and abundance EBV data products could be built based on consistent global or regional data sets that are available for particular taxa. Herein, we selected as examples four projects that have compiled such data sets, covering both the terrestrial and marine realms (Table 3). We use these projects to identify how EBV-useable data sets, EBV-ready data sets and derived and modelled EBV data are produced from raw observations and how projects implement the data processing in a workflow environment.

The four projects were (*i*) eBird, a citizen-science program collecting massive information on bird species distributions, abundances and trends; (*ii*) the TEAM network, wildlife monitoring surveys of ground-dwelling mammals and birds using camera traps in tropical forests; (*iii*) the Living Planet Index (LPI) data set, a collection of over 18500 time series records of more than 3700 vertebrate species worldwide; and (*iv*) the Baltic Sea zooplankton monitoring (BALTIC) data set, about 60000 abundance measurements of marine zooplankton collected at 26 stations from the national plankton monitoring programs in the Baltic region (Table 3). These projects represent data sets covering a range of spatial and temporal extents and resolutions, different measures of species distribution or abundance, and various statistical modelling and data analysis tools. An overview is provided in Table 3 and a detailed description of these projects is provided as online Supporting Information in Appendix S1 Tables S1–S5.

IV. DATA AND TOOLS FOR BUILDING EBV DATA PRODUCTS

In principle, a single point measurement of species distribution or abundance could be incorporated into

Table 3. Characteristics of four Essential Biodiversity Variable (EBV)-relevant projects, including eBird, the Tropical Ecology Assessment and Monitoring (TEAM) network, the Living Planet Index (LPI) and national plankton monitoring programs in the Baltic region (BALTIC)

Characteristics	eBird	TEAM	LPI	BALTIC
Spatial extent	Global (predominately Western Hemisphere)	Tropical forests worldwide	Global	Baltic Sea
Spatial resolution	Three million local sites, model resolution is 3 km^2	23 tropical forest sites (120–200 km^2 resolution)	5598 sites with varying resolution, not stratified	26 marine stations in the Baltic Sea
Temporal extent	2000–present	2007–present	1970–present	2006–present
Temporal resolution	Hourly and daily, weekly after modelling	7-day time periods, annual after modelling	Varies among locations	Monthly
Taxonomic extent	Birds	Ground-dwelling mammals and birds	Vertebrates	Zooplankton
Taxonomic resolution	Species	Species	Species	Species
Measure of species distribution or abundance	Checklists (counts of individuals of a species during a search)	Presence/absence derived from camera-trap records	Population size, density, catch per unit effort, or abundance indices	Number of individuals per m^3
Statistical model or data analysis	Spatiotemporal exploratory model (STEM)	Bayesian dynamic occupancy model	Generalized Additive Model (GAM)	Summary statistics
Data after modelling or analysis	Predicted relative abundance, trends, habitat use	Geometric mean of relative occupancies	Geometric mean of average change in abundance or average annual rates of changes	Mean abundance
Key references	Fink *et al.* (2010); Kelling *et al.* (2015); Sullivan *et al.* (2014)	Ahumada *et al.* (2013); Beaudrot *et al.* (2016); Jansen *et al.* (2014)	Collen *et al.* (2009); Loh *et al.* (2005)	http://sharkdata.se/; Appendix S1

an EBV data product. However, the attributes and uncertainties of all dimensions of the point measurement or a set of consistent point measurements should ideally be expressed in standardized metadata to allow integration with other data (see Section VI). Additionally, information about the sampling protocol needs to be accessible. With data from many diverse sources, a set of algorithms would be required to convert multiple data points into common measurement units comparable across space, time and taxonomy (EBV-ready data sets). However, there is enormous complexity in harmonizing distribution and abundance estimates (Chave, 2013; Azaele *et al.*, 2015; Proença *et al.*, in press) and no effort has yet been undertaken to combine all available data into one EBV data product at a global scale. Below, we highlight and summarize different types of distribution and abundance data, key aspects to consider when combining multi-source data sets, and emerging methods and technologies for data collection.

(1) Distribution data

A diverse set of species distribution data types is available. Here, we summarize them as opportunistic incidence records, presence–absence data and repeated surveys (Table 4).

The most common type of observation data is opportunistic incidence records, which are often incidentally reported or aggregated without a specific sampling protocol. Opportunistic incidence records generally refer to presence-only observations (Peterson *et al.*, 2011). Such presence-only records – e.g. derived from museum or herbarium collections and unstructured citizen observations – contain vast amounts of information about where and when organisms have been observed, but do not report searches that did not find the species (i.e. absences). Presence-only data are often subject to bias in space and time, such as uneven sampling and variation in detectability among species and habitats (Isaac & Pocock, 2015). These biases can severely impact the potential usefulness of these data for EBV data products.

A second type of data is presence–absence data such as those available from checklists or atlas projects (Table 4). These are often produced through surveys where sites are visited to record whether they are 'occupied' (species presence) or not (species absence) (MacKenzie *et al.*, 2006). For these data, the main issue is whether the species has been reliably detected (true *versus* false absence). For instance, occupied sites may be visited and yet no individuals may be detected (= false absence). Hence, measuring absences is more time consuming and subject to bias even in rigorously controlled field assessments (Isaac & Pocock, 2015).

Table 4. Examples of data types considered candidates for building Essential Biodiversity Variable (EBV) data products on species distribution and abundance, including their advantages and disadvantages

Type of data	Examples	Advantages	Disadvantages
Species distribution data			
Opportunistic incidence records	Presence-only data from museum or herbarium collections	Vast amounts of data available, easily aggregated across infrastructures, common minimum data set	Mostly opportunistically collected, often without details of survey effort or method, usually no true absences, hard to estimate detection probabilities, wide variation in data quality
Presence–absence data	Checklists, atlas or camera-trap data	More information content (absences) than opportunistic incidence records	Measuring absences is time consuming and depends on species and habitats
Repeated surveys	Monitoring schemes, repeated atlas projects	Standardized protocols for sampling, occurrences from multiple points in time	Often restricted geographically to Europe and North America, temporal extent varies among surveys
Abundance data			
Opportunistic population counts	Large-scale citizen science projects, eBird, aerial surveys of wide-ranging or aggregating fauna, some vegetation surveys	Massive amount of data	Not sampled repeatedly at fixed sites, sometimes sampled without standardized protocols
Population time series	Soay sheep on St. Kilda, capture histories, North American Breeding Bird Survey, UK Butterfly Monitoring Scheme, LTER Network, TEAM	Repeated population surveys with standardized protocols at fixed sites over multiple years	Available for few species, spatial and temporal resolution depends on organism size and life history, geographic bias towards Europe and North America, variation in sampling protocols or their applicability, some methods are resource-intensive

LTER, Long-Term Ecological Research; TEAM, Tropical Ecology Assessment and Monitoring.

A third source of data comes from repeated surveys that use a standardized protocol for sampling occurrences (and sometimes absences) at multiple points in time (Guillera-Arroita, 2017), for example monitoring schemes (Proença *et al.*, in press) or repeated atlas projects (Jetz *et al.*, 2012). This provides occurrence information from multiple points in time (Table 4), but data are often geographically restricted to wealthier countries (Proença *et al.*, in press).

Another source for distribution data are expert range maps which are expert-drawn outlines of species distributions (Jetz *et al.*, 2012). Examples are distribution maps of birds provided by BirdLife International (http://datazone.birdlife.org/home) or those of mammals and amphibians provided by the International Union for Conservation of Nature and Natural Resources (IUCN, http://www.iucnredlist.org/). Expert range maps (Jetz *et al.*, 2012) provide rough estimates of the outer boundaries of areas within which species are likely to occur, albeit patchily. However, range maps contain large spatial and temporal uncertainties which limit their applicability for measuring changes in species distributions across time.

(2) Abundance data

For population abundance data, two major data types can be distinguished: opportunistic population counts and population time series (Table 4).

Opportunistic population counts are often derived from initiatives and projects that do not sample repeatedly at fixed sites (Table 4). Even when collected with standardized protocols and metadata documentation (regarding survey effort, sampling method, etc.), such data are difficult to analyse due to various sources of bias that need to be accounted for (Kelling *et al.*, 2011; Hochachka & Fink, 2012).

A second type of abundance data is population time series (Table 4). These can result from repeated and consistent population surveys from single-species or multi-species monitoring schemes. Single-species population time series are often produced in conservation monitoring programs, e.g. of threatened or invasive species and for populations of economically important species with commercial or recreational value. Another valuable type of data is time series of populations of multiple species that are repeatedly recorded with standardized protocols at networks of sites.

Examples of population time-series are complete counts of all individuals (Coulson *et al.*, 2001), capture histories of marked individuals using capture–recapture methods (Nichols, 1992), citizen science monitoring schemes such as the North American Breeding Bird Survey (Sauer *et al.*, 2013) and the UK Butterfly Monitoring Scheme (Pollard & Yates, 1993), or projects such as the Long Term Ecological Research (LTER) Network (www.lternet.edu), the US National Ecological Observatory Network (NEON) (http://www.neonscience.org/) and the Tropical

Ecology Assessment and Monitoring (TEAM) Network (www.teamnetwork.org). Data availability is biased towards Europe and North America (McRae, Deinet & Freeman, 2017; Proença *et al.*, in press). Moreover, many of the large-scale monitoring schemes provide indices of relative abundance rather than actual population sizes.

(3) Key aspects for building EBV data products

The typical data characteristics outlined above suggest that EBV data products will usually need to be built from multiple sources, e.g. mixing data from various opportunistic counts, repeated surveys or from observations that were sampled with different protocols or at different spatial and temporal resolutions. Hence, several aspects must be considered when building EBV-ready data sets or when producing derived and modelled EBV data products.

(a) *Harmonizing measurement units from different data sources*

Measurement units present a multi-layered consideration for building EBV-ready data sets, particularly those relating to population abundance. First, abundance is not easily derived from population density when the surveyed area is not reported. Second, measurements of abundance are different for different types of organisms: most vertebrates are usually counted as individuals, plants and fish are usually measured as biomass or percentage cover, and aquatic meiofauna in terms of unit volume. Third, constraints in the way data are collected produce data sets with a great variety of units for measuring abundance. Some of these issues could be overcome by developing a set of well-validated correction factors, or by combining abundance data from different sources (e.g. using geometric means) (Buckland *et al.*, 2011). However, this requires at least that metadata are sufficiently informative (see Section VI).

(b) *Dealing with different spatial scales*

Spatial scale, i.e. resolution (grain size) and extent, is a pervasive issue in ecology (Chave, 2013). Most metrics in ecology depend on the spatial resolution at which they are measured, including species abundance and distribution. For building EBV-ready data sets, it is necessary that the spatial units of sampling are well defined and that data can be converted and standardized to consistent spatial extents and resolutions. Population data sets collected at high spatial resolutions can provide sensitive metrics of occurrence and abundance patterns across scales, but high-resolution information rarely covers broad spatial extents. More often, EBV data products will need some type of rescaling to combine observations collected at different spatial scales. The rescaling procedure can have important consequences for the consistency of the EBV data product because few ecological measurements increase in direct proportion with spatial scale (e.g. probability of occupancy does not scale linearly with sampled area). There is a growing theoretical and methodological literature that explores scaling relationships of distribution and abundance information across resolutions and extents (e.g. Storch, Marquet & Brown, 2007; Nichols *et al.*, 2008; Keil, Wilson & Jetz, 2014; Pagel *et al.*, 2014; Azaele *et al.*, 2015). However, such methods have not yet been applied in the wider context of building global data products from multi-source observational data. It therefore remains unclear whether such approaches can be generalized and how they can be implemented within EBV workflows.

(c) *Correcting for imperfect detection*

Imperfect detection needs to be accounted for when estimating species distributions and population parameters (Kéry & Schaub, 2012), and is particularly important in monitoring applications. A species that is not detected in a survey might be either not there (true absence) or undetected (false absence). A range of statistical models and tools has been developed to correct for imperfect detection and to address different data types (Guillera-Arroita, 2017). In general, accounting for imperfect detection requires information about the observation process (e.g. number of visits, observers, or detection methods), meaning that sampling detection covariates and accurate metadata on data collection and processing is crucial for building EBV-ready data sets. For example, distance sampling corrects for detection probability in transect counts as a function of distance and other covariates (Buckland *et al.*, 2015). If data structures allow separate estimation of detection and occupancy parameters, the detection probability can be estimated statistically (MacKenzie *et al.*, 2006; Royle & Dorazio, 2009). Hence, methods for imperfect detection play a crucial role in building EBV data products.

(d) *Interpolation and extrapolation*

Original observations from which species distribution and abundance EBVs can be derived are by necessity sparse and heterogeneous. Moreover, for most biodiversity data sets the distribution of search effort across space and time is irregular, and the data available in a particular area or time of year may be limited (Kelling *et al.*, 2015). This means that in addition to aggregating and harmonizing data (i.e. producing EBV-ready data sets), a significant amount of spatial and temporal interpolation (e.g. gap filling) and extrapolation (i.e. prediction beyond sampled space or time) might be required for producing derived and modelled EBV data products. One possibility is to apply geostatistical models for spatial interpolation (e.g. kriging or co-kriging) to estimate abundance or occupancy in places where sample data is limited (Meng, Liu & Borders, 2013). Another possibility is to use SDMs for filling in the geographical or temporal gaps between locations with data (Elith & Leathwick, 2009; Fink *et al.*, 2010). Whilst most of these methods are commonly used and well suited for interpolation, their use for extrapolation is prone to error and high uncertainty (Elith & Leathwick, 2009).

(*e*) *Quantifying uncertainties*

Uncertainties in EBV data products might derive from uncertainties in the underlying raw data (e.g. when producing EBV-ready data sets) or from the applied models (e.g. when producing derived and modelled EBV data). Data uncertainties are evident in all key dimensions of species distribution and abundance EBVs, e.g. due to inaccuracy and imprecision in raw data collection (Table 1). This can relate to the accuracy of geographic coordinates, the precision of time of collection, or spelling errors and orthographic variants in taxon names. Several tools have already been applied to quantify uncertainties of geographic coordinates, sampling dates and taxon names (Guralnick *et al.*, 2006; Belbin *et al.*, 2013; Lepage, Vaidya & Guralnick, 2014; Enquist *et al.*, 2016; Meyer *et al.*, 2016). However, assessments of data uncertainties are often lacking and high-throughput processing tools for quantifying uncertainties must be developed for EBV data products.

Data uncertainties can be further exacerbated by uncertainties in models, e.g. through covariates, model fitting and parameter estimation (Beale & Lennon, 2012). Different methods can account for uncertainty from different sources when deriving indicators. The most common way is to use the data to generate pseudo-replicates of the parameters from bootstrapping and then to propagate this distribution into aggregated indicators (O'Brien *et al.*, 2010). Another is to use hierarchical models (Gelman *et al.*, 2013) or to employ Monte Carlo Markov Chains to solve simultaneously for the parameters and calculate the derived indicators as a by-product under a Bayesian framework (Royle & Dorazio, 2009; Ahumada *et al.*, 2013). Such methods are important when building EBV data products.

(4) Emerging methods and technologies for data collection

Although large amounts of species distribution and abundance data are already available from traditional surveys, huge gaps exist in their geographic, temporal and taxonomic coverage (Fernández *et al.*, 2015; Meyer *et al.*, 2015, 2016; Amano *et al.*, 2016; Proença *et al.*, in press). A number of emerging methods and technologies could potentially help to fill these gaps (Fig. 2).

(*a*) *Citizen science*

One opportunistic way to improve data coverage is to mobilize more data from species monitoring projects and field observation campaigns (Stephenson *et al.*, 2017; Proença *et al.*, in press). This includes citizen science projects (Chandler *et al.*, in press; Dickinson *et al.*, 2010) which can have several advantages over traditional field surveys (Fig. 2). Assuring the usefulness of citizen-science data for EBV data products requires careful design of data-input and management procedures and recording of associated information such as sampling effort, species absence and other data-collection variables (Sullivan *et al.*, 2014; Isaac & Pocock, 2015; Kelling *et al.*, 2015). Citizen science data may also require additional cleaning to protect the privacy of volunteers, and additional metadata documentation to meet conditions of attribution (Bowser, Wiggins & Stevenson, 2013).

(*b*) *Sensor networks*

The *in situ* collection of species distribution and abundance information can be accelerated by building networked sensor instruments (Porter *et al.*, 2005). For instance, networks of camera traps allow monitoring the spatiotemporal dynamics of terrestrial birds and mammals in tropical forests (Kays *et al.*, 2009; Jansen *et al.*, 2014). Such camera-trap networks have been successfully applied in many terrestrial systems to measure species occupancy and abundance (Burton *et al.*, 2015), but only recently have they been deployed in marine systems (Bicknell *et al.*, 2016). Extending the use of camera traps within marine, terrestrial and freshwater research domains can provide valuable new data and insightful images on species distributions and abundances. Lessons from terrestrial systems (e.g. imperfect detection, effective sampling area, multi-species inference) will help to facilitate a successful transition of such methods to other environments (Bicknell *et al.*, 2016).

Beyond camera traps, the development of new networked sensors for automatic species recognition based on sound detection (Jeliazkov *et al.*, 2016) also offers novel possibilities to calculate large-scale temporal trends of species distribution and abundances. Comprehensive tracking of animal movements from space may further enable the distributed monitoring of species occurrence (Kays *et al.*, 2015), although this approach might largely be limited to terrestrial and marine birds and mammals (Kissling, 2015). Overall, sensor networks have many advantages over traditional surveys (Fig. 2). They allow a less costly and much more comprehensive sampling than field observations.

(*c*) *DNA-based techniques*

Recent progress in molecular techniques related to high-throughput DNA sequencing has disclosed unprecedented perspectives for monitoring biodiversity (Creer *et al.*, 2016). In particular, DNA (meta-)barcoding (i.e. analysis of one or a few orthologous but variable DNA regions) or metagenomics (i.e. shotgun sequencing of genomic fragments) are rapid and cost-effective means for taxonomic identification of hundreds or thousands of organisms in both terrestrial and aquatic environments (Bourlat *et al.*, 2013; Segata *et al.*, 2013; Creer *et al.*, 2016; Leray & Knowlton, 2016). These approaches can provide presence/absence data for macro- and micro-organisms from given locations at a particular time. However, the DNA-based techniques currently do not allow reliable estimation of absolute population abundances because the number of sequence reads is relative to the sum of the abundances of the other species in a given sample (Aylagas, Borja & Rodríguez-Ezpeleta, 2014).

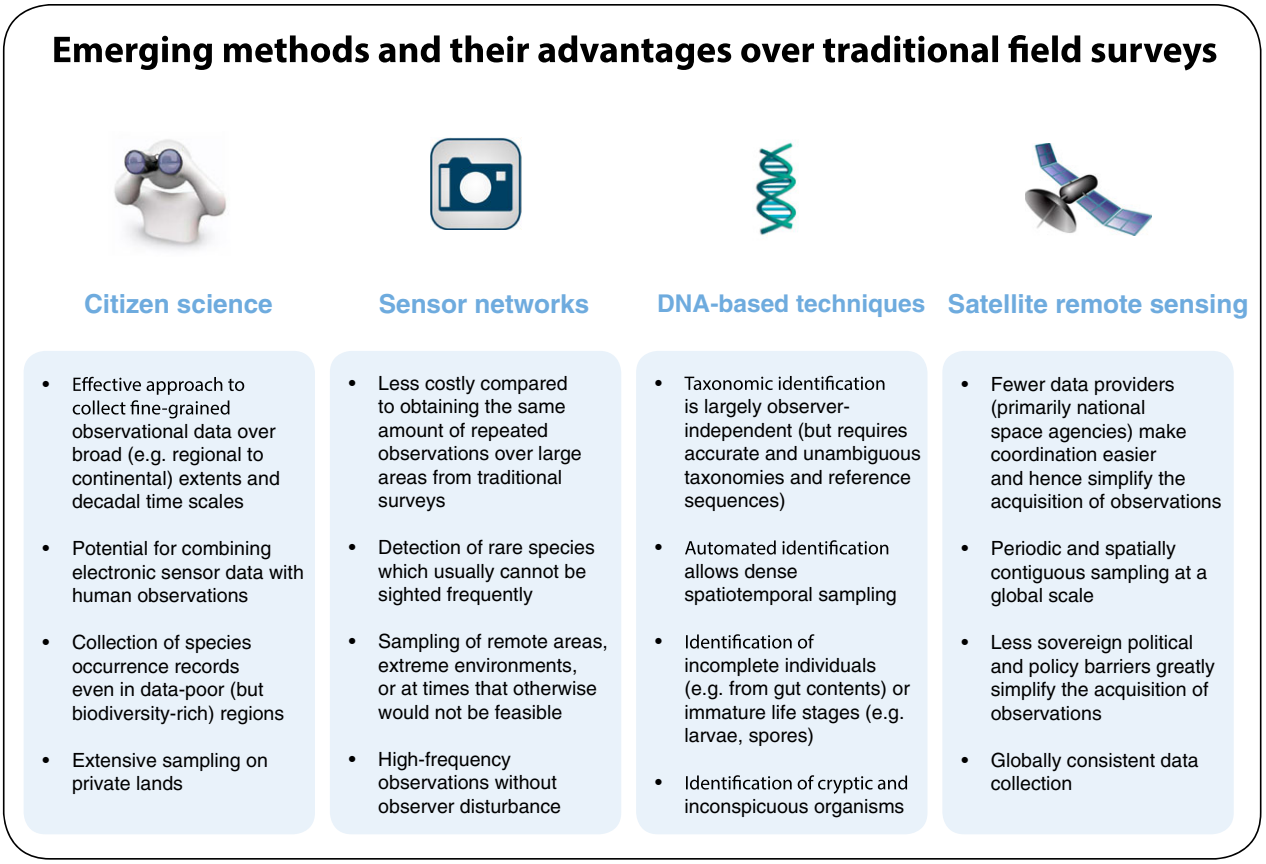

Fig. 2. Emerging methods and technologies for data collection include citizen science, sensor networks, DNA-based techniques and satellite remote sensing. They have several advantages over traditional *in situ* field surveys for collecting species distribution and abundance data. The images are freely available at http://www.clipartpanda.com

Despite advantages, several problems exist with DNA-based surveys. Common marker sequences (e.g. 18S, 28S and 16S rRNA genes, or the ribosomal internal transcribed spacer) may not yield sufficient variation for species identification. Mitochondrial genes (e.g. cytochrome oxidase I) can identify species, but their use in large-scale and long-term projects is constrained by tremendous variation in primer use, amplification steps and sequencing platforms (Bucklin *et al.*, 2016). Most species also have insufficient reference sequences in public archives such as GenBank, and a universal taxonomic system that allows combining the operational taxonomic units (OTUs) with the traditional Linnaean names is still lacking (Kõljalg *et al.*, 2016). Nevertheless, DNA-based techniques have many advantages for assessing species distributions when compared to *in situ* field observations, including identification of cryptic organisms or incomplete individuals (Fig. 2).

(*d*) *Satellite remote sensing*

Satellite remote sensing can play a crucial role in building EBV data products, including those on species distributions and population abundances (Pereira *et al.*, 2013; Skidmore *et al.*, 2015; Lausch *et al.*, 2016; Pettorelli *et al.*, 2016). Nevertheless, measuring species distributions and population abundances from satellite remote sensing has various restrictions when identifying individual plants or animals. One key bottleneck is spatial resolution. For instance, the accurate identification of individual animals such as large wildlife in open savannah habitats (Yang *et al.*, 2015) or penguins on ice (Witharana & Lynch, 2016) requires very high-resolution satellite images with a spatial resolution preferably below 1 m. However, such global remote-sensing products with very high resolution are currently only available through commercial satellite operators, and are costly. As new spaceborne hyperspectral instruments become available, species distribution monitoring from space will become increasingly common and viable, especially for monitoring plant species that characteristically dominate specific vegetation types. The planned launches of next-generation satellites such as EnMAP (http://www.enmap.org/) will allow scaling up towards global monitoring.

Future research should widen the applications for mapping species distributions and abundances from both airborne and spaceborne spatial imagery to a larger number of animal and plant species from diverse habitats and biomes. This has several advantages compared to *in situ* observations, including a much more consistent and contiguous data collection at broad spatial scales (Fig. 2). With appropriate

ground-truthing, this makes satellite remote sensing an ideal method for understanding biodiversity change at national, continental and global scales (Schimel, Asner & Moorcroft, 2013).

V. WORKFLOWS FOR BUILDING EBV DATA PRODUCTS

(1) Importance of workflows for building EBV data products

The overview in the previous section shows that building EBV data products requires a substantial level of integration of data from a large and dispersed number of data providers, as well as complex preparation and processing steps. Historically, similar work has involved sets of computer 'scripts'. Often, these are implemented in a variety of programming languages and executed *via* different user interfaces, with non-automated tasks interspersed throughout the process. However, for consistently producing and replicating EBV data products it would be advantageous to develop and preserve all data access, integration and processing steps as an open-source and freely available service in a workflow-oriented e-infrastructure that supports curation, sharing and collaboration (Gärdenfors *et al.*, 2014; Enquist *et al.*, 2016; Hardisty *et al.*, 2016; La Salle *et al.*, 2016; Hugo *et al.*, 2017). The development of this kind of workflow in such an environment not only supports the automation of routine tasks, but also allows formation of analytical protocols that are robust, transparent and reusable, thereby improving reproducibility of ecological research (Borregaard & Hart, 2016).

Some individual workflows with relevance to abundance and distribution data have already been developed. Examples include occurrence retrieval and taxonomic data cleaning and integration (Mathew *et al.*, 2014) and creating, applying, projecting and visualizing models for species distributions and range shifts (De Giovanni *et al.*, 2016). So far, these individual workflows have not been exploited for EBV production but they illustrate many of the important steps. Here, we identify key workflow steps that are needed to build EBV data products on species distributions and population abundances from multi-source data sets. We then show how the previously mentioned projects (i.e. eBird, TEAM, LPI and BALTIC) relate to these workflow steps. Finally, we highlight legal and technical aspects that are important for a workflow-oriented production of EBV data products at a global scale.

(2) Workflow for building species distribution and abundance EBV data products

It is possible to identify the most generic key steps of a workflow system that allows the building of EBV data products related to species distributions and abundances (Fig. 3). This includes multiple sequential activities, such as identification and aggregation of various raw data sources, data quality control, including duplicate data checks and taxonomic name matching, and statistical modelling of integrated data. We identified 11 workflow steps of key relevance (Fig. 3) in relation to the three major types of EBV data sets (Fig. 1). We discuss each step briefly below.

(*a*) *EBV-useable data sets*

The first series of workflow steps is aimed at gathering EBV-useable data sets (orange in Fig. 3). This process begins with identifying and importing the relevant raw data (step 1, Fig. 3). For *in situ* species distribution and abundance data, raw observations are currently mostly derived from natural history collections, national and international monitoring programs, or research surveys including those utilizing citizen science. For some of these, the Global Biodiversity Information Facility (GBIF) serves as a global aggregation infrastructure that speeds up the discovery and retrieval process. For a workflow implementation the available raw data need to contain not only relevant measurements on species distributions or abundances (e.g. opportunistic incidence records, checklists, density estimates), but also accessible metadata (e.g. in relation to spatial and temporal extent and resolution, measurement units, sampling strategy, taxonomy) that facilitate an automatic and standardized process of data extraction and processing (e.g. Fegraus *et al.*, 2011).

In a second step, data-sharing agreements and licenses (if available) must be checked and enforced (step 2, Fig. 3). Not all data are freely available or useable, and some uses may be restricted in certain contexts (e.g. commercial use) or through certain conditions (e.g. requiring citation or another form of attribution). While including standardized and machine-readable data licenses would allow this step to be integrated with step 1, machine-readable licenses are not currently consistently applied to biodiversity data. Licenses and other key considerations related to legal interoperability are discussed below.

In a third step, a basic data consistency and completeness check must be performed to identify whether the data are complete and whether they have time formats, measurement units, spatial information and species identity descriptors that can be mapped to known standards (step 3, Fig. 3). This is important because available data sets are often incomplete or archived in a way that partially or entirely prevents their reuse (Roche *et al.*, 2015). These first three steps allow to compile data into EBV-useable data sets for subsequent processing in a standardized way.

(*b*) *EBV-ready data sets*

In the next sequence of workflow steps, harmonized (EBV-ready) data sets can be produced that are standardized. They will have undergone rudimentary quality control, with corrections applied where relevant (green in Fig. 3). This requires aggregation of data sets from different sources (step 4, Fig. 3). Combining different types of data (e.g. opportunistic population counts and population time series, or population time series from different sources) into the

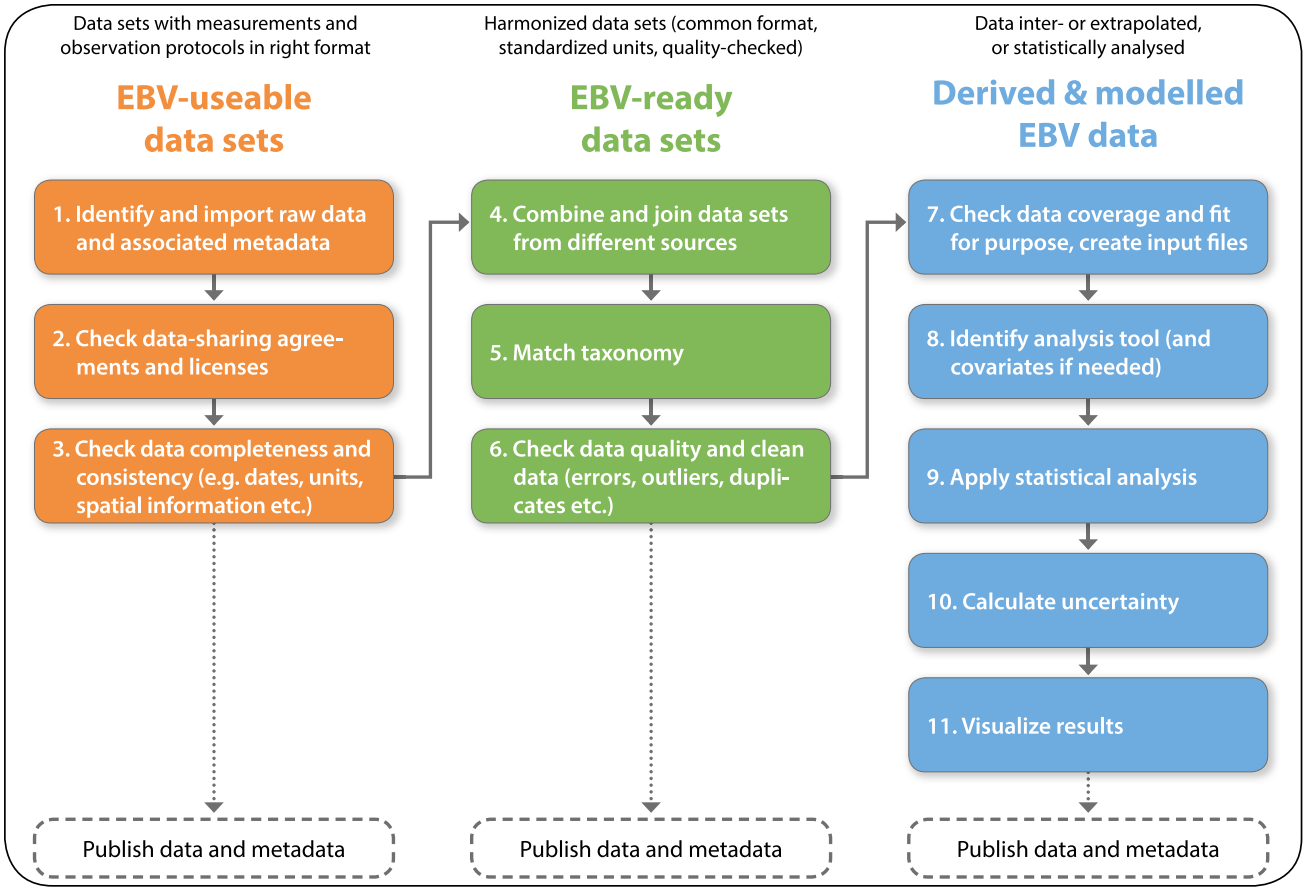

Fig. 3. Workflow steps of key relevance for building Essential Biodiversity Variable (EBV) data sets for species distributions and abundances. The workflow steps are grouped into three major types of EBV data sets (see Fig. 1): EBV-useable data sets (orange), EBV-ready data sets (green) and derived and modelled EBV data (blue). Each of the EBV data sets should ideally be published with relevant metadata.

same EBV is not trivial because each data type represents different sampling schemes, measurement units and spatial and temporal resolutions. It might therefore be most feasible to combine different data sets that have the same data type or those collected for similar purposes.

Another key step is to match taxonomic names (Boyle *et al.*, 2013), e.g. to relate synonyms to accepted species names if they derive from different taxonomic treatments (step 5, Fig. 3). As a minimum, this requires standardized lists or backbone taxonomies such as those from the Catalogue of Life, the World Register of Marine Species, or those directly maintained by specific research communities (e.g. AmphibiaWeb, ReptileBase, Mammal Species of the World, iPlant). Nevertheless, such a process can be complicated (requiring additional tools) because of numerous taxonomic revisions and inconsistencies over time. Tools such as Avibase (Lepage *et al.*, 2014), which provides taxon concept mapping across different bird taxonomic resources, may be particularly relevant, but are rare outside well-studied taxonomic groups or geographic regions.

A data quality check and cleaning of data is then needed (step 6, Fig. 3). This includes identifying and annotating errors, outliers, wrong identifications and duplicates which may arise from merging heterogeneous data sets (Fernández *et al.*, 2015; Meyer *et al.*, 2016). Many e-Science infrastructures have already developed tools for cleaning and correcting outliers, errors and duplicates, e.g. in relation to taxonomy and geo-referencing (Constable *et al.*, 2010; Kelling *et al.*, 2011; Belbin *et al.*, 2013; Enquist *et al.*, 2016; La Salle *et al.*, 2016). Of course, not all issues can be automatically detected or corrected, and community assistance is needed to help improve the quality and cleaning of the data (e.g. Constable *et al.*, 2010; Kelling *et al.*, 2011).

The six steps described above are fundamental for producing EBV-useable and EBV-ready data sets (Fig. 3).

(c) *Derived and modelled EBV data products*

A further set of workflow steps must be performed if EBVs are being interpolated or extrapolated (e.g. for gap filling) or further processed with specific analysis tools (blue in Fig. 3). This requires checking data coverage and fitness for purpose, and then creating input files for subsequent modelling (step 7, Fig. 3). This process should allow for the custom selection of specific taxa or groups of taxa required for a particular purpose, as well as the ability flexibly to aggregate data

spatially into grid cells or temporally into time bands of a specific resolution and extent.

The next workflow step is to choose an appropriate analytical tool (e.g. a specific statistical model) and, if required, relevant covariates (step 8, Fig. 3). This depends on the type of data (e.g. presence-only, presence–absence, abundance) and the specific purpose (e.g. inter- or extrapolation, spatial prediction, temporal modelling of occupancy). The choice can include various species distribution, occupancy and abundance models (Fink *et al.*, 2010; Conn *et al.*, 2015; Beaudrot *et al.*, 2016). For some (but not all) model applications, various detection and habitat covariates might be needed (Fink *et al.*, 2010; Beaudrot *et al.*, 2016).

The analytical tools can then be applied (step 9, Fig. 3). A next step is to calculate and report uncertainty of the analysis (step 10, Fig. 3). A final step of the workflow is to generate visualizations that can be provided together with the underlying data (step 11, Fig. 3). The last two steps might not always be needed to produce a derived and modelled EBV data product.

(*d*) *Publishing EBV data products*

A relevant and important aspect of the workflow is to publish the EBV data products and associated metadata (Fig. 3). This should include not only the EBV data products but also the raw data sets used (or links to them), the scripts, analytical tools and software applied, machine-readable workflow metadata, and licensing information. In addition, dashboards with numbers and summary statistics, geospatial layers, maps and animations for occurrence and abundance predictions, predictive performance metrics and predictor importance information for the covariates that were used in the models, and conditions for reuse should also be published. Each data product should have a persistent identifier, such as a Digital Object Identifier (DOI), so that it is traceable and provides appropriate credit (Hugo *et al.*, 2017). The content and structure of metadata should be in a standardized form to allow accessibility *via* machine-to-machine interactions, but this must still be developed for EBV data products (see Section VI).

(3) Application of workflow to empirical examples

To demonstrate the application of the above-described workflow, we investigated how the various workflow steps have been implemented by eBird, TEAM, LPI and BALTIC (Table 3) when applied to building EBV-useable and EBV-ready data sets (Fig. 4) as well as derived and modelled EBV data products (Fig. 5). A detailed description of the workflow steps used in each of the projects is provided in Appendix S1 Tables S1–S5.

The comparison of the generic workflow steps with those realized in specific projects (Figs 4 & 5) largely demonstrated congruency. All workflow steps for derived and modelled EBV data (Fig. 5) and most steps for building EBV-useable and EBV-ready data sets were realized by these projects (Fig. 4). However, some of the generic workflow steps for building EBV-useable and EBV-ready data sets – e.g. the matching of taxonomic names (step 5) and data quality-control procedures (step 6) – were implemented at different points in the workflow (Fig. 4). Moreover, workflow steps such as checking data-sharing agreements and licenses (step 2) or combining data sets from different sources (step 4) are often done by these projects before the EBV workflow starts. For instance, data-sharing agreements and licenses (step 2) were already arranged between data aggregators and key data providers before step 1. This allowed the raw data to enter the workflow, e.g. by agreeing to use data for all non-commercial purposes or by signing specific data-sharing agreements (Fig. 4).

Of key importance is the publication of data products (Fig. 3). The eBird data (i.e. EBV-useable and EBV-ready data sets) can be requested from the eBird portal (http://ebird.org/ebird/data/download; this requires registration and login). Additional bar charts, maps, graphs, tables and visualization tools can also be explored (see http://ebird.org/ebird/explore). For TEAM, the camera trap data (i.e. EBV-useable and EBV-ready data sets) are publicly available (http://www.teamnetwork.org/data/query) and the modelled occupancy time series for each population at each site (i.e. derived & modelled EBV data) can be downloaded (http://wpi.teamnetwork.org/wpi/dashboard). The LPI provides the individual records for each time series (i.e. the EBV-useable data set) – excluding about 3000 time series with confidential data – through its data portal (http://livingplanetindex.org/data_portal) as well as with the latest publication (McRae *et al.*, 2017) and as part of the Living Planet Report (http://www.livingplanetindex.org). All data used in BALTIC were downloaded from the Swedish LifeWatch portal (Gärdenfors *et al.*, 2014) for marine monitoring data (EBV-useable data set), while processed data were stored in the data archives of Environment Climate Data Sweden (https://ecds.se/ under file identifier: ccc84507-49c1-43df-9887-97d2232bcb89), including the harmonized data (EBV-ready data set) and the statistically analysed data (derived & modelled EBV data). Despite these publishing efforts by all four projects, their data sets are published in different ways, and limited consistency and adoption of data and metadata standards is apparent.

(4) Legal interoperability in EBV workflows

The above-mentioned workflow assumes legal interoperability, i.e. a condition where: (*i*) the legal use conditions can be clearly and readily determined for each data set; (*ii*) the legal use conditions for each data set allow for both creation and use of combined, or derivative products; and (*iii*) users can legally access and use each data set without seeking authorization from data rights holders on a case-by-case basis (RDA-CODATA Legal Interoperability Interest Group, 2016). Legal interoperability is therefore an important requirement for automated workflows and for the successful development of EBV data products. However, there are many cases where raw data and data sets might not

Workflow steps	eBird	TEAM	LPI	BALTIC
1. Identify and import raw data and associated metadata	eBird checklists with metadata from web portal	Camera trap images with metadata from memory cards, initially identified to species	Population time series with metadata from literature, websites, and online databases	Marine monitoring data with basic metadata from web portal
2. Check data-sharing agreements and licenses	Project's data is publicly available for non-commercial use	Project's data is publicly available for non-commercial use	Most raw data publically available, otherwise data-sharing agreements were signed	Data are freely available from the data aggregator, including commercial use
3. Check data completeness and consistency (e.g. dates, units, spatial information etc.)	All data passed through data-quality filters based on historic eBird data and expert opinion. Unusual records are flagged for further review	Each data set (site) is checked for spatial and temporal consistency and continuity, and whether species are within their geographic ranges	Basic quality checks (e.g. consistency of method and survey effort across time)	Basic cleaning and formatting, e.g. regarding latitude/longitude, time format, tabular structure with same headers, metadata templates, etc.
4. Combine and join datasets from different sources	Data come from the same web portal. No additional data joining is needed at this stage	Data come from the same web portal. No additional data joining is needed at this stage	Data come from different studies and sites, and are combined in one database	Data come from the same web portal. No additional data joining is needed at this stage
5. Match taxonomy	Taxonomy matches the annual updates of the Clements checklist (a globally standardized bird taxonomy), done in step 1	Taxonomy follows the IUCN Red List and taxonomic matching is done in step 1	Taxonomy follows IUCN Red List except for mammals, taxonomic matching is done in step 1	Taxonomic names from Swedish Taxonomic backbone (before step 1), matching with European Marine Taxonomic backbone
6. Check data quality and clean data (errors, outliers, duplicates etc.)	Data quality checks done in step 3	Images of difficult-to-identify species are shared with experts	Data quality checks and removal of duplicates done in step 3	Correction of location names and mismatches in spatial coordinates

Fig. 4. Workflow steps 1–6 in relation to four projects with Essential Biodiversity Variable (EBV)-relevant data products. The workflow steps follow Fig. 3. Additional details about these projects are summarized in Table 3 and Appendix S1.

be 'findable, accessible, interoperable and reusable' (FAIR principles; Wilkinson *et al.*, 2016), and this can constrain legal interoperability.

(*a*) *Constraints on legal interoperability*

Key constraints on legal interoperability emerge from restrictions on data use, modification and sharing. For instance, legal interoperability is constrained when data sets are protected by different intellectual property rights (IPRs), e.g. U.S. copyright law or E.U. Database protection. Other restrictions on data access and re-use can come from national security regulations, protection of endangered species, other types of confidentiality, individual use agreements (e.g. contracts, licenses and disclaimers) as well as incompatible data policies of different data sets for the same species. In many situations, individual researchers and research teams act as if data were proprietary, and the use of data is not necessarily designated as open. When data sets are made accessible only on a case-by-case basis or if they are kept secret, legal interoperability is not achievable.

Restrictions on data access, use and sharing therefore have severe consequences for building EBV data products. For instance, including copyrighted or otherwise restricted raw data in EBV data products can impede quality control, limit data aggregation and restrict re-usability. Building EBV data products that involve proprietary data or sensitive data (e.g. because of threatened or endangered species) could lead to situations where different results are produced by different researchers. Moreover, legal interoperability becomes a key issue when combining data from multiple sources because the most restrictive access rights of an included data set will dictate the access rights of the whole EBV data product. Existing legal mechanisms therefore need to be promoted and enforced to ensure open access and legal interoperability (RDA-CODATA Legal Interoperability Interest Group, 2016).

(*b*) *The need for common-use licenses*

One of the most efficient approaches that help to assure legal interoperability are common-use licenses. The best-known set of common-use licenses are the Creative Commons (CC) licenses (Carroll, 2006). Six CC licenses describe different conditions for re-use, and one designation (CC0) supports a full waiver of copyright in favour of placing work in the public domain (Table 5). When combining data from multiple sources with different CC licenses, the resulting EBV

Workflow steps	eBird	TEAM	LPI	BALTIC
7. Check data coverage and fit for purpose, create input files	Only data across the Western Hemisphere are currently used for modelling	Data cubes (camera trap point, time in days, and year) are created for populations and species at each site	Data subsets are created for different kinds of analyses	Data filtering (annual, seasonal, diurnal, depth slices, adults, frequent species), averaging of replicates
8. Identify analysis tool (and covariates if needed)	Two-stage species distribution model called Spatio-Temporal Exploratory Model (STEM); elevation and MODIS land cover classes as covariates	Bayesian-formulated dynamic occupancy model; elevation, temperature, rainfall, people presence–absence and distance to border as covariates	Chain method (CM), generalized additive model (GAM), mixed-effects model; geographic, ecological and life-history traits as covariates	Summary statistics (mean, min, max abundance as well as abundance range), no covariates
9. Apply statistical analysis	Species distribution models are fitted for small regions and time periods, then are combined into a single STEM 'ensemble' model that covers the entire region and calendar year	A Bayesian occupancy model is fitted using Monte Carlo Markov Chain (MCMC) algorithms using JAGS (Just Another Gibbs Sampler)	First, a GAM or CM is fitted to each time series, then either an aggregated indicator is calculated or mixed-effects models are fitted with covariates and annual rates of change	Calculation of summary statistics
10. Calculate uncertainty	Uncertainty is estimated by randomly resampling training data within local regions. Mean, 10th and 90th percentiles are calculated across ensembles of 100 local models	The Bayesian model estimates distributions of all parameters	Bootstrapping is used to produce 95% confidence intervals for calculated indices of abundance, uncertainty is propagated through the time series	Uncertainty is not calculated
11. Visualize results	Geospatial layers for occurrence and abundance predictions, maps and animations	Occupancy trend for each population at each site, or as combined occupancy trend using the Wildlife Picture Index	Indices of abundance over time	Boxplots showing abundance trends over years

Fig. 5. Workflow steps 7–11 in relation to four projects with Essential Biodiversity Variable (EBV)-relevant data products. The workflow steps follow Fig. 3. Additional details about these projects are summarized in Table 3 and Appendix S1.

data product will incorporate the accumulated restrictions imposed by each source of data. Since not all licenses are compatible with each other (Table 5), this can have severe consequences for aggregating data from multiple sources. For instance, if one data set licensed with CC0 (i.e. placed in the public domain) is combined with data licensed as CC BY (i.e. attribution only) and also combined with data licensed as CC BY-NC (i.e. with attribution and only for non-commercial uses), then the resulting EBV data product will contain the most stringent license type (here, CC BY-NC). This is problematic because it limits the re-use of the EBV data product by imposing conditions of attribution and non-commercial use. Moreover, two of the six CC license types are incompatible with each other because they do not permit modification (Table 5).

Hence, the ideal data sets for building EBV products are those in the public domain, with no restrictions on re-use and attribution (i.e. no need to specify source and license). If conditions for re-use are limited to attribution, the only legal requirement will be to include sufficient metadata to indicate the data source and the appropriate standardized license, if one is needed. Two Creative Commons designations (CC0 and CC BY) are therefore recommended for denoting open data (Table 5), while the remaining five impose restrictions on re-use. International data-sharing principles, such as those from the Group on Earth Observations (GEOSS, http://www.earthobservations.org/dswg.php), endorse CC0 and CC BY. We recommend following this endorsement when building EBV data products.

(5) Technical requirements for a workflow-oriented production

There are multiple technical requirements related to a workflow-oriented production of global EBV data products. Interoperable computing infrastructure and promotion of the development, sharing and use of workflows is needed (Hardisty *et al.*, 2013; Gärdenfors *et al.*, 2014; Kissling *et al.*, 2015; Enquist *et al.*, 2016). This includes agreeing and adopting appropriate structural formats for EBV data products, for means of data storage and for the execution and implementation of EBV workflows.

(a) *Structural formats of EBV data products*

The structure and format of the three types of EBV data (EBV-useable data sets, EBV-ready data sets and derived and modelled EBV data) must be further investigated. The format and specification should facilitate interdisciplinary research

Table 5. Overview of Creative Commons (CC) licenses and designations (https://creativecommons.org/) and their compatibility when combining data from multiple sources. Combining data from multiple sources with different licenses will result in the most stringent license type. Several licenses are compatible with each other, but those that do not permit modification are not compatible with others. The most flexible is the CC0 designation, which represents a full waiver of copyright in favour of placing work in the public domain. For the building of Essential Biodiversity Variable data products, CC0 and CC BY are recommended because they either impose no restrictions on re-use (CC0) or only require attribution (CC BY)

License symbol	License name	Description	Compatible with
	No Rights Reserved [CC0]	Copyright holder chooses to opt out of copyright, placing work in the public domain	Any
	Attribution [CC BY]	Permits access and use; including modification; for any purpose; with attribution	Any
	Attribution-ShareAlike [CC BY-SA]	Permits access and use; including modification; for any purposes; with attribution. All derivative works must use this license	CC0, CC BY
	Attribution-NonCommercial [CC BY-NC]	Permits access and use; including modification; with attribution. Does not permit commercial use	CC0, CC BY
	Attribution-NonCommercial-ShareAlike [CC BY-NC-SA]	Permits access and use; including modification; with attribution. Does not permit commercial use. All derivative works must use this license	CC0, CC BY, CC BY-NC
	Attribution-NoDerivs [CC BY-ND]	Permits access and use; for any purpose; with attribution. Does not permit modification	None
	Attribution-NonCommercial-NoDerivs [CC BY-NC-ND]	Permits access and use; with attribution. Does not permit modification; does not permit commercial use	None

and other uses. It should allow manipulation by a wide range of software tools and be accessible to a range of end users. Many established file formats currently exist, including the Darwin Core Archive (DwC-A) for raw data, the Network Common Data Format (NetCDF) widely used in Earth sciences and the Web Map Service (WMS) for geospatial data (e.g. raster and vector) from the Open Geospatial Consortium (OGC) (Hugo *et al.*, 2017). GEO BON and the wider scientific community (e.g. data producers and product developers) should therefore develop recommendations and agreements on the most appropriate data models and file formats for EBV data products.

(*b*) *Data storage*

EBV data products must be stored and managed on a long-term, semi-permanent basis. It is presently unclear who can take the burden of the storage commitment, and how this will be funded. The means of investment will depend on whether the semi-permanent storage of EBV data products will be done centrally (i.e. in a single repository, perhaps with duplicated mirror sites) or as a network of distributed, independent but interoperable storage services. The GEO BON Secretariat may host a selection of derived and modelled EBV data, but rely on partners such as GBIF nodes to host other EBV data products, including EBV-useable and EBV-ready data sets. Any large-scale trusted data repository for EBV data products must offer quality assurance as well as publishing and archival services so that different Biodiversity Observation Networks (BONs) and other actors can contribute to, manage and publish high-quality EBV data products.

(*c*) *Execution and implementation of workflows*

Implementing the large-scale execution of workflows for EBV data production requires appropriate computing infrastructure. This probably requires workflow management systems that allow an automated execution of data-intensive scientific workflows on distributed computing infrastructures by multiple users (Deelman *et al.*, 2009; Liu *et al.*, 2015; Hardisty *et al.*, 2016). Since infrastructure providers want to be flexible in terms of which computing environment and technologies they use, a workflow solution neutral to specific hardware architectures and execution models is needed. The particular software for managing these workflows is less critical than the use of a standard execution-independent mechanism of workflow representation. An open question is who can take on the responsibility and cost of implementing specific EBV workflows. It is probably not the role of data publishers, and few existing biodiversity research infrastructures are currently well funded, sustained and able to do this.

Two basic options for the technical production cycle exist. The first option is a 'create-on-demand' process that requires

ready access to relevant raw data, the workflow and processing capacity at any time of interest. It requires that the source data for computing EBV data products are compliant with technical and semantic interoperability standards to allow on-demand machine translation and mapping of algorithms. The second option is a periodic and systematic production of EBV data products (e.g. annually). This reflects that EBV data products are published/archived as an ever-extending data archive that can be consulted for a specific place or area of interest (local, regional, national) at the time of interest. The differences between create-on-demand and periodic systematic production are fundamental for how the technical production processes are defined, how infrastructures are organized and how access and use permissions for raw data are handled. A mixture of both approaches may be needed, e.g. with the GEO BON Data Portal hosting periodic EBV data products, but on-demand EBV data products being hosted/offered by GEO BON partners.

VI. METADATA AND DATA-SHARING STANDARDS

(1) The need for standardized metadata to describe EBV data products

The development of EBV data products from multi-source observation data in a workflow-oriented e-infrastructure depends on standardized metadata to make data findable, accessible, interoperable and reusable (FAIR principles; Wilkinson *et al.*, 2016). When a community agrees to such standardization efforts, the result is consistently presented information that can be searched using known terms. Together with controlled vocabularies (i.e. carefully selected lists of words and phrases) and ontologies (i.e. formal statements of relationships among concepts represented by vocabulary terms) this facilitates sharing and discovery of biodiversity data because it allows consistency for interoperability and machine reasoning (Thessen & Patterson, 2011; Michener & Jones, 2012). The ideal metadata for assessing fitness-for-purpose of potential EBV-useable data sets (*cf*. Fig. 3) would provide information about the extent, resolution, measurement units and uncertainties of spatial, temporal and taxonomic data dimensions. This would allow machine-readable discovery and aggregation of large numbers of candidate data sets (Wilkinson *et al.*, 2016). For a data set to be considered EBV-ready (see Fig. 3), its structure needs to be transformed and harmonized into interoperable formats and units. Furthermore, the often-complex workflows required to consistently reproduce derived and modelled EBV data products (see Fig. 3) must document data provenance, i.e. a record of the data's origin and what has been done with it. This includes recording the modelling and processing steps in a consistent manner. Below, we provide a brief assessment of the current state of the art by highlighting current standards and common formats for sharing biodiversity data.

(2) Current standards for sharing biodiversity data

To date, there is no agreed schema for documenting EBV data products, but there are many ongoing efforts that can facilitate the capture of metadata for ecological data (e.g. Michener *et al.*, 1997; Fegraus *et al.*, 2005; Wieczorek *et al.*, 2012; Walls *et al.*, 2014). Here, we summarize the most relevant existing standards related to species distribution and abundance EBVs (Table 6).

(a) *The Darwin Core standard and the Event Core*

The Darwin Core (DwC) provides a set of terms that facilitate the exchange of information about the occurrence of organisms in nature and the resulting specimens in biological collections (Wieczorek *et al.*, 2012). As originally conceived, it is a data and metadata standard for reporting incidental observations or specimens (summarized under 'opportunistic incidence records' in Section IV.1). To help organize the 169 terms currently in the standard, these are grouped into several categories or classes. A list of terms and descriptions is available at http://rs.tdwg.org/dwc/terms/.

Until recently, the DwC has been insufficient to provide detailed reporting on many aspects of inventories because no DwC terms existed to report the scope, effort and completeness of surveys (which is critical information for assessing potential absence of species). New DwC terms have begun to close these gaps, including better capture of quantities found within sampling units. These include biomass or number of individuals as proxies of abundance, capture of sample effort reporting and identifiers relating parent–child relationships between events to represent hierarchical sampling designs.

In addition, a new organization of the data in the DwC has recently been proposed that is particularly suitable for EBV data products. The vast majority of records in DwC format are 'occurrence core' records that have broad use for reporting opportunistic incidence records in that profile format. However, there is now the possibility to make the 'event' the core, and to place occurrence records as related to that event (Wieczorek *et al.*, 2014). This enables data holders to share structured survey data such as population time series or presence–absence data (GEO BON, 2016), for instance through GBIF's Integrated Publishing Toolkit (IPT) (http://www.gbif.org/sites/default/files/gbif_IPT-sample-data-primer_en.pdf). However, standard details about sampling method and effort are not required in these 'Event Core' fields. One possible option is to utilize extensions for various kinds of inventories, with the caveat that this may create new challenges with heterogeneous data and interpretation. In sum, new models of publishing biodiversity data extend the DwC approach from its basis in reporting opportunistic incidence records to more structured surveys with detailed methods and abundance data.

(b) *The Ecological Metadata Language*

The Ecological Metadata Language (EML) is an extensive metadata standard for environmental data sources

Table 6. A list of the most relevant existing standards related to species distribution and abundance Essential Biodiversity Variables (EBVs). The relevance of each standard is highlighted for different types of EBV data sets (see EBV workflow in Fig. 3). The first three standards are more focal to species distribution and abundance data and do not cover derived and modelled data.

Standard	Explanation	Relevance for EBV data sets	
		EBV-useable and EBV-ready data sets	Derived and modelled EBV data
Simple Darwin Core (DwC)	Used to standardize occurrence records (e.g. opportunistic incidence records) or taxon checklists	X	—
Darwin Core 'event'	Used to report sampling events and associated taxa or recovered specimens	X	—
Humboldt Core	A detailed specification for reporting inventory process and type, including scope, method and completeness assessment	X	—
Biocollections Ontology (BCO)	An ontology for representing sampling processes for biological data, including inventory processes	X	X
Ecological Metadata Language (EML)	A broad standard and language for reporting information about data sets coming from ecological studies	X	X
Extensible Observation Ontology (OBOE) and Observations and Measurements (O&M)	OBOE is a broad ontology for reporting observations and measurements of entities from evidence, along with context O&M has a similar remit with application focus for sensor networks	X	X
ISO 19115	A standard for describing spatial and temporal distribution of digital geographic data	X	X
ISO 19157	A standard for reporting geographic data quality including outputs from processing steps of geographic input data	X	X
PROV	A broad family of recommendations to report and exchange provenance information of digital data objects	X	X

(Michener *et al.*, 1997; Jones *et al.*, 2001; Michener & Jones, 2012). It uses controlled vocabularies (i.e. predefined, authorized terms) and specifically targets long-term observation data. Some EML modules describe the data set's spatial, temporal and taxonomic coverage, and since these metadata elements can be accessed at the data discovery stage, powerful filtering becomes possible (e.g. based on taxonomic criteria without downloading and inspecting large data sets). Other EML modules describe methods and protocols of sampling or processing or contain information on responsible persons and organizations, software resources, or access rules. EML metadata and DwC data files can be bundled into DwC archives, which are self-describing zip files produced during data publication. These are ultimately harvested by data aggregators such as GBIF. For this, data publishers use EML standard fields to describe ownership, creation and licensing (Robertson *et al.*, 2014).

(*c*) *Other specifications and standards*

A number of other specifications and standards can serve the need for metadata associated with EBV workflows. Several ontologies have been developed with the aim of improving data aggregation and integration across the biodiversity domain, including the Biological Collections Ontology (BCO) (Walls *et al.*, 2014), the Population and Community Ontology (PCO) (Walls *et al.*, 2014), the Environmental Ontology (ENVO) (Buttigieg *et al.*, 2016) and genomic standards such as the Minimum Information about any (x) Sequence (MIxS) (Yilmaz *et al.*, 2011). For *in situ* biotic inventory processes, the Humboldt Core provides a vocabulary for describing spatial, temporal, environmental and taxonomic information (https://mol.org/humboldtcore/). It includes methodology and effort reporting and a way to assess how a survey and inventory was performed and what was collected (e.g. abundance information and how it was recorded). Other ontologies for describing observations and measurements based on digital or material sample evidence include the Observation and Measurements ontology (O&M) and the Extensible Observation Ontology (OBOE) (Table 6). For Earth observation data, the ISO 19115/19157 metadata standard allows detailed documentation of spatial and temporal data characteristics, measurement units, legal and licensing restrictions, data quality and provenance. The ISO standard is widely used to describe satellite products and other environmental data, and is a core component of the Global Earth Observing System of Systems (GEOSS). These

standard reporting mechanisms may themselves be linked to general models for provenance such as PROV (Missier, Belhajjame & Cheney, 2013) to assure both human and machine-readable information about provenance.

(3) Metadata standards for EBV data products

Practical integration of the above-mentioned standards must still be achieved in the context of the species distribution and abundance EBVs. Lack of such integration was evident when investigating the four example projects (see Section V). However, aggregators such as GBIF already provide workflow services that utilize community-developed standards for expression of occurrence data, e.g. by creating a centralized resource in DwC format with additional complementary metadata files in EML (Robertson *et al.*, 2014). This includes the steps needed to produce EBV-ready data sets (workflow steps 4–6 in Fig. 3), e.g. to combine and join data sets of opportunistic incidence records using a shared taxonomic backbone, and with capacity to perform geographic quality checks (e.g. Otegui & Guralnick, 2016). The recent extension of DwC terms and new publishing mechanisms for an 'event core' further extend the DwC ability to serve multi-species distribution and abundance data from surveys, monitoring networks and other inventories.

There has been no focus so far on standard metadata approaches for derived and modelled EBV data products. ISO 19115/19157 can provide a suitable starting point for these modelling steps as it has flexible and extensive provenance options for describing (in a structured, machine-readable format) how a data set was generated (e.g. algorithms, inputs, outputs and processing steps). Such an approach allows development of reproducible workflows for EBV products, especially if standardized identifiers and controlled vocabularies are used (Guralnick *et al.*, 2015). EML can also provide a detailed description of methods applied to data, but this is currently done *via* an unstructured textual description and not *via* controlled vocabularies. Semantically rich approaches such as the BCO and the PCO (see Section VI.2*c*) that focus on inputs and outputs as part of workflow steps could further be useful for the EBV production chain.

In an EBV workflow, uncertainties in terms of data, model algorithms and parameters must be effectively characterized and reported. These can come from heterogeneous data that vary in spatial, temporal and taxonomic dimensions (Table 1). Such uncertainties can propagate and reverberate through the EBV production chain. They must be clearly quantified, described and ultimately controlled. For traditional geospatial metadata, ISO 19157 explicitly allows various facets of data uncertainty to be captured. This could be particularly relevant if metadata documents are to be linked to modelled and derived EBV data products (see Fig. 3). When the ISO standard is combined with other controlled vocabularies (e.g. Yang *et al.*, 2013), detailed quantitative reports of errors can be constructed in a machine-readable form. This would allow the propagation of statistical uncertainty information throughout the EBV workflow to be recorded.

VII. CONCLUSIONS

(1) In this review, we have provided a first overview about how to operationalize the EBV concept for species distribution and abundance data at a global scale. We discussed (*i*) important data and tools for building EBV data products, (*ii*) the potential for a workflow-oriented production of EBVs, and (*iii*) relevant standards for capturing consistent machine-readable metadata to drive interoperability. We also addressed several challenges associated with building global EBV data products from multi-source data sets in a workflow-oriented e-infrastructure. Many of these topics reflect aspects of the 'Big Data' challenge in biodiversity science today.

(2) Building global EBV data products on species distributions and abundances requires multiple data sets on presence (and absence) or population size of species to be combined and harmonized. This can be achieved by developing workflows that take multiple sequential activities into account, including identification and aggregation of various raw data sources, data quality control, taxonomic name matching and statistical modelling of integrated data. All data access, integration and processing steps should be provided as an open and free service in a workflow-oriented e-infrastructure that supports curation, sharing and collaboration. We urge funding agencies to provide financial resources that support the building of EBV data products, the implementation and development of EBV workflows, and the coordination and cooperation of research infrastructures to achieve these goals.

(3) Harmonizing data from opportunistic records and counts, presence–absence data, repeated surveys and population time series will be a major step towards building global EBV data products of species distribution and abundance. Combining such heterogeneous, multi-source data sets across space, time, taxonomy and different sampling methods requires the development of tools and models for data and model integration. Key scientific issues include correcting for imperfect detection, dealing with different spatial resolution and extents, harmonizing measurement units from different data sources, applying methods for spatial inter- or extrapolation and developing tools for quantifying and propagating sources of uncertainty in data and models. We recommend that innovation in this field is promoted by developing methods and tools that support harmonization and integration of disparate raw observations into EBV-useable, EBV-ready and derived and modelled EBV data products.

(4) The identification of key workflow steps is highly relevant for building global EBV data products. This helps to establish analytical protocols that are robust, transparent and reusable, thereby improving reproducibility of ecological research. Existing projects, research infrastructures and citizen science efforts already operationalize many generic workflow steps. We demonstrate that the identified workflow steps are applicable to both the terrestrial and aquatic systems and a broad range of spatial, temporal and taxonomic scales. Nevertheless, these workflow steps still must be integrated

or combined for global EBV data production. It is therefore imperative to develop and implement such analytical protocols and workflows in a sustained e-infrastructure that allows global EBV data production.

(5) In line with the FAIR principles (Wilkinson *et al.*, 2016), it is of vital importance to document metadata with controlled vocabularies and ontologies and to improve (meta-)data standards and procedures for building global EBV data products. Metadata should capture information about the extent, resolution, measurement units and uncertainties of the spatial, temporal and taxonomic data dimensions, as well as conditions for data access and use. They should further describe the provenance of EBV data products with metadata on modelling and processing steps. Data standards already provide means to share and discover data sets and their properties. Although no specific metadata standards and information models have yet been developed specifically for EBVs, recent developments on biodiversity data standards are particularly suitable for building EBV data products on species distribution and abundance. Engagement with the existing bio-geospatial standards communities can catalyse progress towards an information model that allows semantic interoperability for integrating multi-source data sets when building global EBV data products.

(6) Building reliable and representative global EBV data products requires filling of data gaps in geographic, temporal and taxonomic coverage. This necessitates a renewed effort in data mobilization and expanding existing biodiversity-monitoring initiatives worldwide, including citizen science projects. In addition, field observations on species distributions and population abundances need to be supplemented with data from sensor networks, DNA-based techniques and satellite remote sensing. New computing infrastructure as well as storage capacity is needed for processing, building and storing such EBV data products. Agreement is also needed on the most effective and efficient data structures and representation formats.

(7) Building global EBV data products from multiple sources will benefit from open data or data that are free from restrictions on use, modification and sharing. Governmental mechanisms, such as intergovernmental agreements, as well as national legislation, regulations, or policies and non-governmental mechanisms such as 'common-use' licenses or simple normative (*versus* legal) agreements need to be promoted and enforced to ensure open access and legal interoperability. An important step is the endorsement of the CC0 designation and CC BY license when building EBV data products. Any restrictions on re-use and attribution will affect the processing of multi-source data sets in a workflow-oriented e-infrastructure and constrain the usefulness of EBV data products for science and policy advice.

VIII. ACKNOWLEDGEMENTS

This paper emerged from the first two workshops of the Horizon 2020 project GLOBIS-B (GLOBal Infrastructures for Supporting Biodiversity research; http://www.globis-b.eu/). We thank Carsten Meyer and one anonymous reviewer for constructive comments on a previous version, Jacco Konijn for administrative support, Jan van Arkel for graphical support, and Jörg Freyhof, Renato De Giovanni, Liqiang Ji, Francisco Hernandez, Dimitris Koureas, Jesus Marco de Lucas, David Manset, Jeffrey Manuel, Eise van Maanen and John W. Watkins for discussions during the workshops. We are grateful to Jörg Freyhof, Helen Matthey, the Group on Earth Observations Biodiversity Observation Network (GEO BON) and the German Centre for Integrative Biodiversity Research (iDiv) for hosting our first workshop in Leipzig, Germany, from 29 February to 2 March 2016. We further thank Antonio Torralba Silgado, Juan Miguel González Aranda, Jesús Miguel Santamaría, Clara Luján, Alfonso Herrera, the University of Seville, and the Joint Research Unit LifeWatch Spain-JRU LW.ES for supporting our second workshop in Sevilla, Spain, from 13–15 June 2016. Financial support came from the European Commission (grant 654003). C. A. additionally received funding from the LifeWatchGreece infrastructure (MIS 384676), funded by the Greek Government under the General Secretariat of Research and Technology (GSRT), ESFRI Projects and National Strategic Reference Framework (NSRF). M. O. was supported by the Swedish LifeWatch project funded by the Swedish Research Council (Grant no. 829-2009-6278), and J.E. by the Australian Research Council (grant FT0991640).

IX. REFERENCES

References marked with asterisk have been cited within the supporting information.

Ahumada, J. A., Hurtado, J. & Lizcano, D. (2013). Monitoring the status and trends of tropical forest terrestrial vertebrate communities from camera trap data: a tool for conservation. *PLoS ONE* **8**, e73707.

Amano, T., Lamming, J. D. L. & Sutherland, W. J. (2016). Spatial gaps in global biodiversity information and the role of citizen science. *Bioscience* **66**, 393–400.

Aylagas, E., Borja, Á. & Rodríguez-Ezpeleta, N. (2014). Environmental status assessment using DNA metabarcoding: towards a genetics based marine biotic index (gAMBI). *PloS One* **9**, e90529.

Azaele, S., Maritan, A., Cornell, S. J., Suweis, S., Banavar, J. R., Gabriel, D. & Kunin, W. E. (2015). Towards a unified descriptive theory for spatial ecology: predicting biodiversity patterns across spatial scales. *Methods in Ecology and Evolution* **6**, 324–332.

Beale, C. M. & Lennon, J. J. (2012). Incorporating uncertainty in predictive species distribution modelling. *Philosophical Transactions of the Royal Society B: Biological Sciences* **367**, 247–258.

Beaudrot, L., Ahumada, J. A., O'Brien, T., Alvarez-Loayza, P., Boekee, K., Campos-Arceiz, A., Eichberg, D., Espinosa, S., Fegraus, E., Fletcher, C., Gajapersad, K., Hallam, C., Hurtado, J., Jansen, P. A., Kumar, A., et al. (2016). Standardized assessment of biodiversity trends in tropical forest protected areas: the end is not in sight. *PLoS Biology* **14**, e1002357.

Begon, M., Townsend, C. R. & Harper, J. L. (2006). *Ecology: From Individuals to Ecosystems*, Fourth Edition. Blackwell Publishing, Malden.

Belbin, L., Daly, J., Hirsch, T., Hobern, D. & LaSalle, J. (2013). A specialist's audit of aggregated occurrence records: an 'aggregator's' perspective. *ZooKeys* **305**, 67–76.

Bicknell, A. W. J., Godley, B. J., Sheehan, E. V., Votier, S. C. & Witt, M. J. (2016). Camera technology for monitoring marine biodiversity and human impact. *Frontiers in Ecology and the Environment* **14**, 424–432.

Bojinski, S., Verstraete, M., Peterson, T. C., Richter, C., Simmons, A. & Zemp, M. (2014). The concept of essential climate variables in support of climate research, applications, and policy. *Bulletin of the American Meteorological Society* **95**, 1431–1443.

Borregaard, M. K. & Hart, E. M. (2016). Towards a more reproducible ecology. *Ecography* **39**, 349–353.

Bourlat, S. J., Borja, A., Gilbert, J., Taylor, M. I., Davies, N., Weisberg, S. B., Griffith, J. F., Lettieri, T., Field, D., Benzie, J., Glöckner, F. O., Rodríguez-Ezpeleta, N., Faith, D. P., Bean, T. P. & Obst, M. (2013). Genomics in marine monitoring: new opportunities for assessing marine health status. *Marine Pollution Bulletin* **74**, 19–31.

Bowser, A., Wiggins, A. & Stevenson, R. (2013). *Data Policies for Public Participation in Scientific Research: A Primer*. DataONE, Albuquerque.

Boyle, B., Hopkins, N., Lu, Z., Raygoza Garay, J. A., Mozzherin, D., Rees, T., Matasci, N., Narro, M. L., Piel, W. H., Mckay, S. J., Lowry, S., Freeland, C., Peet, R. K. & Enquist, B. J. (2013). The taxonomic name resolution service: an online tool for automated standardization of plant names. *BMC Bioinformatics* **14**, 16.

Buckland, S. T., Rexstad, E. A., Marques, T. A. & Oedekoven, C. S. (2015). *Distance Sampling: Methods and Applications*. Springer, Cham.

Buckland, S. T., Studeny, A. C., Magurran, A. E., Illian, J. B. & Newson, S. E. (2011). The geometric mean of relative abundance indices: a biodiversity measure with a difference. *Ecosphere* **2**, 1–15.

Bucklin, A., Lindeque, P. K., Rodriguez-Ezpeleta, N., Albaina, A. & Lehtiniemi, M. (2016). Metabarcoding of marine zooplankton: prospects, progress and pitfalls. *Journal of Plankton Research* **38**, 393–400.

Burton, A. C., Neilson, E., Moreira, D., Ladle, A., Steenweg, R., Fisher, J. T., Bayne, E. & Boutin, S. (2015). Wildlife camera trapping: a review and recommendations for linking surveys to ecological processes. *Journal of Applied Ecology* **52**, 675–685.

Butchart, S. H. M., Walpole, M., Collen, B., van Strien, A., Scharlemann, J. P. W., Almond, R. E. A., Baillie, J. E. M., Bomhard, B., Brown, C., Bruno, J., Carpenter, K. E., Carr, G. M., Chanson, J., Chenery, A. M., Csirke, J., et al. (2010). Global biodiversity: indicators of recent declines. *Science* **328**, 1164–1168.

Buttigieg, P. L., Pafilis, E., Lewis, E. S., Schildhauer, M. P., Walls, R. L. & Mungall, C. J. (2016). The environment ontology in 2016: bridging domains with increased scope, semantic density, and interoperation. *Journal of Biomedical Semantics* **7**, 1–12.

Carroll, M. W. (2006). Creative commons and the new intermediaries. *Michigan State Law Review* **45**, 45–65.

Ceballos, G., Ehrlich, P. R., Barnosky, A. D., García, A., Pringle, R. M. & Palmer, T. M. (2015). Accelerated modern human–induced species losses: entering the sixth mass extinction. *Science Advances* **1**, e1400253.

Chandler, M., See, L., Copas, K., Bonde, A. M. Z., López, B. C., Danielsen, F., Legind, J. K., Masinde, S., Miller-Rushing, A. J., Newman, G., Rosemartin, A. & Turak, E. (in press). Contribution of citizen science towards international biodiversity monitoring. *Biological Conservation* https://doi.org/10.1016/j.biocon.2016.09.004.

Chave, J. (2013). The problem of pattern and scale in ecology: what have we learned in 20 years? *Ecology Letters* **16**, 4–16.

Collen, B. E. N., Loh, J., Whitmee, S., McRae, L., Amin, R. & Baillie, J. E. M. (2009). Monitoring change in vertebrate abundance: the Living Planet Index. *Conservation Biology* **23**, 317–327.

Conn, P. B., Johnson, D. S., Hoef, J. M. V., Hooten, M. B., London, J. M. & Boveng, P. L. (2015). Using spatiotemporal statistical models to estimate animal abundance and infer ecological dynamics from survey counts. *Ecological Monographs* **85**, 235–252.

Constable, H., Guralnick, R., Wieczorek, J., Spencer, C., Peterson, A. T. & The VertNet Steering Committee (2010). VertNet: a new model for biodiversity data sharing. *PLoS Biology* **8**, e1000309.

Coulson, T., Catchpole, E. A., Albon, S. D., Morgan, B. J. T., Pemberton, J. M., Clutton-Brock, T. H., Crawley, M. J. & Grenfell, B. T. (2001). Age, sex, density, winter weather, and population crashes in Soay sheep. *Science* **292**, 1528–1531.

Creer, S., Deiner, K., Frey, S., Porazinska, D., Taberlet, P., Thomas, W. K., Potter, C. & Bik, H. M. (2016). The ecologist's field guide to sequence-based identification of biodiversity. *Methods in Ecology and Evolution* **7**, 1008–1018.

Deelman, E., Gannon, D., Shields, M. & Taylor, I. (2009). Workflows and e-science: an overview of workflow system features and capabilities. *Future Generation Computer Systems* **25**, 528–540.

De Giovanni, R., Williams, A. R., Ernst, V. H., Kulawik, R., Fernandez, F. Q. & Hardisty, A. R. (2016). ENM components: a new set of web service-based workflow components for ecological niche modelling. *Ecography* **39**, 376–383.

Dickinson, J. L., Zuckerberg, B. & Bonter, D. N. (2010). Citizen science as an ecological research tool: challenges and benefits. *Annual Review of Ecology, Evolution, and Systematics* **41**, 149–172.

Elith, J. & Leathwick, J. R. (2009). Species distribution models: ecological explanation and prediction across space and time. *Annual Review of Ecology, Evolution, and Systematics* **40**, 677–697.

Enquist, B. J., Condit, R., Peet, R. K., Schildhauer, M. & Thiers, B. M. (2016). Cyberinfrastructure for an integrated botanical information network to investigate the ecological impacts of global climate change on plant biodiversity. *PeerJ Preprints* **4**, e2615v2.

Fegraus, E. H., Andelman, S., Jones, M. B. & Schildhauer, M. (2005). Maximizing the value of ecological data with structured metadata: an introduction to Ecological Metadata Language (EML) and principles for metadata creation. *Bulletin of the Ecological Society of America* **86**, 158–168.

Fegraus, E. H., Lin, K., Ahumada, J. A., Baru, C., Chandra, S. & Youn, C. (2011). Data acquisition and management software for camera trap data: a case study from the TEAM Network. *Ecological Informatics* **6**, 345–353.

Fernández, M., Navarro, L. M., Apaza-Quevedo, A., Gallegos, S. C., Marques, A., Zambrana-Torrelio, C., Wolf, F., Hamilton, H., Aguilar-Kirigin, A. J., Aguirre, L. F., Alvear, M., Aparicio, J., Apaza-Vargas, L., Arellano, G., Armijo, E., et al. (2015). Challenges and opportunities for the Bolivian Biodiversity Observation Network. *Biodiversity* **16**, 86–98.

*Fink, D., Damoulas, T., Bruns, N. E., La Sorte, F. A., Hochachka, W. M., Gomes, C. P. & Kelling, S. (2014). Crowdsourcing meets ecology: hemisphere-wide spatiotemporal species distribution models. *AI Magazine* **35**, 19–30.

Fink, D., Hochachka, W. M., Zuckerberg, B., Winkler, D. W., Shaby, B., Munson, M. A., Hooker, G., Riedewald, M., Sheldon, D. & Kelling, S. (2010). Spatiotemporal exploratory models for broad-scale survey data. *Ecological Applications* **20**, 2131–2147.

Gärdenfors, U., Jönsson, M., Obst, M., Wremp, A. M., Kindvall, O. & Nilsson, J. (2014). Swedish LifeWatch — a biodiversity infrastructure integrating and reusing data from citizen science, monitoring and research. *Human Computation* **1**, 147–163.

Geijzendorffer, I. R., Regan, E. C., Pereira, H. M., Brotons, L., Brummitt, N., Gavish, Y., Haase, P., Martin, C. S., Mihoub, J.-B., Secades, C., Schmeller, D. S., Stoll, S., Wetzel, F. T. & Walters, M. (2016). Bridging the gap between biodiversity data and policy reporting needs: an essential biodiversity variables perspective. *Journal of Applied Ecology* **53**, 1341–1350.

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, K. & Rubin, D. (2013). *Bayesian Data Analysis*, Third Edition. Chapman and Hall, Boca Raton.

GEO BON (2016). *The Event Core: Moving Beyond Presence-only Data*. Group on Earth Observations Biodiversity Observation Network (GEO BON) Secretariat, Leipzig.

*Gorsky, G., Ohman, M. D., Picheral, M., Gasparini, S., Stemmann, L., Romagnan, J.-B., Cawood, A., Pesant, S., García-Comas, C. & Prejger, F. (2010). Digital zooplankton image analysis using the ZooScan integrated system. *Journal of Plankton Research* **32**, 285–303.

Guillera-Arroita, G. (2017). Modelling of species distributions, range dynamics and communities under imperfect detection: advances, challenges and opportunities. *Ecography* **40**, 281–295.

Guillera-Arroita, G., Lahoz-Monfort, J. J., Elith, J., Gordon, A., Kujala, H., Lentini, P. E., McCarthy, M. A., Tingley, R. & Wintle, B. A. (2015). Is my species distribution model fit for purpose? Matching data and models to applications. *Global Ecology and Biogeography* **24**, 276–292.

Guralnick, R. P., Cellinese, N., Deck, J., Pyle, R. L., Kunze, J., Penev, L., Walls, R., Hagedorn, G., Agosti, D., Wieczorek, J., Catapano, T. & Page, R. (2015). Community next steps for making globally unique identifiers work for biocollections data. *ZooKeys* **494**, 133–154.

Guralnick, R. P., Wieczorek, J., Beaman, R., Hijmans, R. J. & The BioGeomancer Working Group (2006). BioGeomancer: automated georeferencing to map the world's biodiversity data. *PLoS Biology* **4**, e381.

Hampton, S. E., Strasser, C. A., Tewksbury, J. J., Gram, W. K., Budden, A. E. & Batcheller, A. L. (2013). Big data and the future of ecology. *Frontiers in Ecology and the Environment* **11**, 156–162.

Hardisty, A., Roberts, D. & The Biodiversity Informatics Community (2013). A decadal view of biodiversity informatics: challenges and priorities. *BMC Ecology* **13**, 16.

Hardisty, A. R., Bacall, F., Beard, N., Balcázar-Vargas, M.-P., Balech, B., Barcza, Z., Bourlat, S. J., De Giovanni, R., de Jong, Y., De Leo, F., Dobor, L., Donvito, G., Fellows, D., Guerra, A. F., Ferreira, N., et al. (2016). BioVeL: a virtual laboratory for data analysis and modelling in biodiversity science and ecology. *BMC Ecology* **16**, 49.

Hobern, D., Apostolico, A., Arnaud, E., Bello, J. C., Canhos, D., Dubois, G., Field, D., Garcia, E. A., Hardisty, A., Harrison, J., Heidorn, B., Krishtalka, L., Mata, E., Page, R., Parr, C., Price, J. & Willoughby, S. (2013). *Global Biodiversity Informatics Outlook: Delivering Biodiversity Knowledge in the Information Age*. GBIF Secretariat, Copenhagen.

Hochachka, W. & Fink, D. (2012). Broad-scale citizen science data from checklists: prospects and challenges for macroecology. *Frontiers of Biogeography* **4**, 150–154.

Hugo, W., Hobern, D., Kõljalg, U., Tuama, É. Ó. & Saarenmaa, H. (2017). Global infrastructures for biodiversity data and services. In *The GEO Handbook on Biodiversity Observation Networks* (eds M. Walters and R. J. Scholes), pp. 259–291. Springer International Publishing, Cham.

IPBES (2016). In *Summary for policymakers of the methodological assessment of scenarios and models of biodiversity and ecosystem services of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services*. (eds S. Ferrier, K. N. Ninan, P. Leadley, R. Alkemade, L.

A. Acosta, H. R. Akçakaya, L. Brotons, W. W. L. Cheung, V. Christensen, K. A. Harhash, J. Kabubo-Mariara, C. Lundquist, M. Obersteiner, H. Pereira, G. Peterson, R. Pichs-Madruga, N. Ravindranath, C. Rondinini and B. A. Wintle). Secretariat of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services, Bonn.

Isaac, N. J. B. & Pocock, M. J. O. (2015). Bias and information in biological records. *Biological Journal of the Linnean Society* **115**, 522–531.

Jansen, P. A., Ahumada, J. A., Fegraus, E. & O'Brien, T. (2014). TEAM: a standardised camera trap survey to monitor terrestrial vertebrate communities in tropical forests. In *Camera Trapping: Wildlife Research and Management* (eds P. Meek and P. Fleming), pp. 263–270. CSIRO Publishing, Collingwood.

Jeliazkov, A., Bas, Y., Kerbiriou, C., Julien, J.-F., Penone, C. & Le Viol, I. (2016). Large-scale semi-automated acoustic monitoring allows to detect temporal decline of bush-crickets. *Global Ecology and Conservation* **6**, 208–218.

Jetz, W., McPherson, J. M. & Guralnick, R. P. (2012). Integrating biodiversity distribution knowledge: toward a global map of life. *Trends in Ecology & Evolution* **27**, 151–159.

Jones, M. B., Berkley, C., Bojilova, J. & Schildhauer, M. (2001). Managing scientific metadata. *IEEE Internet Computing* **5**, 59–68.

Kays, R., Crofoot, M. C., Jetz, W. & Wikelski, M. (2015). Terrestrial animal tracking as an eye on life and planet. *Science* **348**, aaa2478.

Kays, R., Kranstauber, B., Jansen, P., Carbone, C., Rowcliffe, M., Fountain, T. & Tilak, S. (2009). Camera traps as sensor networks for monitoring animal communities. In *2009 IEEE 34th Conference on Local Computer Networks*, pp. 811–818. Zürich, Switzerland.

Keil, P., Wilson, A. M. & Jetz, W. (2014). Uncertainty, priors, autocorrelation and disparate data in downscaling of species distributions. *Diversity and Distributions* **20**, 797–812.

Kelling, S., Fink, D., La Sorte, F. A., Johnston, A., Bruns, N. E. & Hochachka, W. M. (2015). Taking a 'Big Data' approach to data quality in a citizen science project. *Ambio* **44**, 601–611.

Kelling, S., Yu, J., Gerbracht, J. & Wong, W. K. (2011). Emergent filters: automated data verification in a large-scale citizen science project. In *2011 IEEE Seventh International Conference on e-Science Workshops (eScienceW)*, pp. 20–27. Stockholm, Sweden.

Kéry, M. & Schaub, M. (2012). *Bayesian Population Analysis using WinBUGS. A Hierarchical Perspective*. Academic Press, Burlington.

Kissling, W. D. (2015). Animal telemetry: follow the insects. *Science* **349**, 597.

Kissling, W. D., Hardisty, A., García, E. A., Santamaria, M., De Leo, F., Pesole, G., Freyhof, J., Manset, D., Wissel, S., Konijn, J. & Los, W. (2015). Towards global interoperability for supporting biodiversity research on essential biodiversity variables (EBVs). *Biodiversity* **16**, 99–107.

van Kleunen, M., Dawson, W., Essl, F., Pergl, J., Winter, M., Weber, E., Kreft, H., Weigelt, P., Kartesz, J., Nishino, M., Antonova, L. A., Barcelona, J. F., Cabezas, F. J., Cardenas, D., Cardenas-Toro, J., et al. (2015). Global exchange and accumulation of non-native plants. *Nature* **525**, 100–103.

Kõljalg, U., Tedersoo, L., Nilsson, R. H. & Abarenkov, K. (2016). Digital identifiers for fungal species. *Science* **352**, 1182–1183.

La Salle, J., Williams, K. J. & Moritz, C. (2016). Biodiversity analysis in the digital era. *Philosophical Transactions of the Royal Society B: Biological Sciences* **371**, 20150337.

Lausch, A., Bannehr, L., Beckmann, M., Boehm, C., Feilhauer, H., Hacker, J. M., Heurich, M., Jung, A., Klenke, R., Neumann, C., Pause, M., Rocchini, D., Schaepman, M. E., Schmidtlein, S., Schulz, K., Selsam, P., Settele, J., Skidmore, A. K. & Cord, A. F. (2016). Linking Earth Observation and taxonomic, structural and functional biodiversity: local to ecosystem perspectives. *Ecological Indicators* **70**, 317–339.

Lepage, D., Vaidya, G. & Guralnick, R. (2014). Avibase – a database system for managing and organizing taxonomic concepts. *ZooKeys* **420**, 117–135.

Leray, M. & Knowlton, N. (2016). Censusing marine eukaryotic diversity in the twenty-first century. *Philosophical Transactions of the Royal Society B: Biological Sciences* **371**, 20150331.

Liu, J., Pacitti, E., Valduriez, P. & Mattoso, M. (2015). A survey of data-intensive scientific workflow management. *Journal of Grid Computing* **13**, 457–493.

Loh, J., Green, R. E., Ricketts, T., Lamoreux, J., Jenkins, M., Kapos, V. & Randers, J. (2005). The Living Planet Index: using species population time series to track trends in biodiversity. *Philosophical Transactions of the Royal Society B: Biological Sciences* **360**, 289–295.

Lomolino, M. V., Riddle, B. R., Whittaker, R. J. & Brown, J. H. (2010). *Biogeography*. Sinauer Associates, Sunderland.

MacKenzie, D. I., Nichols, J. D., Royle, J. A., Pollock, K. H., Bailey, L. L. & Hines, J. E. (2006). *Occupancy Estimation and Modeling*. Elsevier Academic Press, Amsterdam.

Mathew, C., Güntsch, A., Obst, M., Vicario, S., Haines, R., Williams, A., de Jong, Y. & Goble, C. (2014). A semi-automated workflow for biodiversity data retrieval, cleaning, and quality control. *Biodiversity Data Journal* **2**, e4221.

McGeoch, M. A., Butchart, S. H. M., Spear, D., Marais, E., Kleynhans, E. J., Symes, A., Chanson, J. & Hoffmann, M. (2010). Global indicators of biological invasion: species numbers, biodiversity impact and policy responses. *Diversity and Distributions* **16**, 95–108.

McRae, L., Deinet, S. & Freeman, R. (2017). The diversity-weighted living planet index: controlling for taxonomic bias in a global biodiversity indicator. *PLoS One* **12**, e0169156.

Meng, Q., Liu, Z. & Borders, B. E. (2013). Assessment of regression kriging for spatial interpolation – comparisons of seven GIS interpolation methods. *Cartography and Geographic Information Science* **40**, 28–39.

Meyer, C., Kreft, H., Guralnick, R. & Jetz, W. (2015). Global priorities for an effective information basis of biodiversity distributions. *Nature Communications* **6**, 8221.

Meyer, C., Weigelt, P. & Kreft, H. (2016). Multidimensional biases, gaps and uncertainties in global plant occurrence information. *Ecology Letters* **19**, 992–1006.

Michener, W. K., Brunt, J. W., Helly, J. J., Kirchner, T. B. & Stafford, S. G. (1997). Nongeospatial metadata for the ecological sciences. *Ecological Applications* **7**, 330–342.

Michener, W. K. & Jones, M. B. (2012). Ecoinformatics: supporting ecology as a data-intensive science. *Trends in Ecology & Evolution* **27**, 85–93.

Millennium Ecosystem Assessment (2005). *Ecosystems and Human Well-Being: Biodiversity Synthesis*. World Resources Institute, Washington.

Missier, P., Belhajjame, K. & Cheney, J. (2013). The W3C PROV family of specifications for modelling provenance metadata. In *Proceedings of the 16th International Conference on Extending Database Technology*, pp. 773–776. ACM, Genoa.

Newbold, T., Hudson, L. N., Arnell, A. P., Contu, S., De Palma, A., Ferrier, S., Hill, S. L. L., Hoskins, A. J., Lysenko, I., Phillips, H. R. P., Burton, V. J., Chng, C. W. T., Emerson, S., Gao, D., Pask-Hale, G., et al. (2016). Has land use pushed terrestrial biodiversity beyond the planetary boundary? A global assessment. *Science* **353**, 288–291.

Nichols, J. D. (1992). Capture-recapture models. *BioScience* **42**, 94–102.

Nichols, J. D., Bailey, L. L., O'Connell, A. F. Jr., Talancy, N. W., Campbell Grant, E. H., Gilbert, A. T., Annand, E. M., Husband, T. P. & Hines, J. E. (2008). Multi-scale occupancy estimation and modelling using multiple detection methods. *Journal of Applied Ecology* **45**, 1321–1329.

O'Brien, T. G., Baillie, J. E. M., Krueger, L. & Cuke, M. (2010). The wildlife picture index: monitoring top trophic levels. *Animal Conservation* **13**, 335–343.

Otegui, J. & Guralnick, R. P. (2016). The geospatial data quality REST API for primary biodiversity data. *Bioinformatics* **32**, 1755–1757.

Pagel, J., Anderson, B. J., O'Hara, R. B., Cramer, W., Fox, R., Jeltsch, F., Roy, D. B., Thomas, C. D. & Schurr, F. M. (2014). Quantifying range-wide variation in population trends from local abundance surveys and widespread opportunistic occurrence records. *Methods in Ecology and Evolution* **5**, 751–760.

Pereira, H. M., Belnap, J., Böhm, M., Brummitt, N., Garcia-Moreno, J., Gregory, R., Martin, L., Peng, C., Proença, V., Schmeller, D. & van Swaay, C. (2017). Monitoring essential biodiversity variables at the species level. In *The GEO Handbook on Biodiversity Observation Networks* (eds M. Walters and R. J. Scholes), pp. 79–105. Springer International Publishing, Cham.

Pereira, H. M., Ferrier, S., Walters, M., Geller, G. N., Jongman, R. H. G., Scholes, R. J., Bruford, M. W., Brummitt, N., Butchart, S. H. M., Cardoso, A. C., Coops, N. C., Dulloo, E., Faith, D. P., Freyhof, J., Gregory, R. D., et al. (2013). Essential biodiversity variables. *Science* **339**, 277–278.

Pereira, H. M., Navarro, L. M. & Martins, I. S. (2012). Global biodiversity change: the bad, the good, and the unknown. *Annual Review of Environment and Resources* **37**, 25–50.

Peterson, A. T., Soberón, J., Pearson, R. G., Anderson, R. P., Martínez-Meyer, E., Nakamura, M. & Araújo, M. B. (2011). *Ecological Niches and Geographic Distributions*. Princeton University Press, Princeton.

Pettorelli, N., Wegmann, M., Skidmore, A., Mücher, S., Dawson, T. P., Fernandez, M., Lucas, R., Schaepman, M. E., Wang, T., O'Connor, B., Jongman, R. H. G., Kempeneers, P., Sonnenschein, R., Leidner, A. K., Böhm, M., et al. (2016). Framing the concept of satellite remote sensing essential biodiversity variables: challenges and future directions. *Remote Sensing in Ecology and Conservation* **2**, 122–131.

Pollard, E. & Yates, T. J. (1993). *Monitoring Butterflies for Ecology and Conservation: The British Butterfly Monitoring Scheme*. Chapman & Hall, London.

Porter, J., Arzberger, P., Braun, H.-W., Bryant, P., Gage, S., Hansen, T., Hanson, P., Lin, C.-C., Lin, F.-P., Kratz, T., Michener, W., Shapiro, S. & Williams, T. (2005). Wireless sensor networks for ecology. *BioScience* **55**, 561–572.

Potts, J. M. & Elith, J. (2006). Comparing species abundance models. *Ecological Modelling* **199**, 153–163.

Proença, V., Martin, L. J., Pereira, H. M., Fernandez, M., McRae, L., Belnap, J., Böhm, M., Brummitt, N., García-Moreno, J., Gregory, R. D., Honrado, J. P., Jürgens, N., Opige, M., Schmeller, D. S., Tiago, P. & van Swaay, C. A. M. (in press). Global biodiversity monitoring: from data sources to essential biodiversity variables. *Biological Conservation* https://doi.org/10.1016/j.biocon.2016.07.014.

RDA-CODATA Legal Interoperability Interest Group (2016). Legal interoperability of research data: principles and implementation guidelines (eds P. Uhlir and G. Clement), p. 41. Research Data Alliance, New York.

Reichman, O. J., Jones, M. B. & Schildhauer, M. P. (2011). Challenges and opportunities of open data in ecology. *Science* **331**, 703–705.
Robertson, T., Döring, M., Guralnick, R., Bloom, D., Wieczorek, J., Braak, K., Otegui, J., Russell, L. & Desmet, P. (2014). The GBIF Integrated Publishing Toolkit: facilitating the efficient publishing of biodiversity data on the internet. *PLoS One* **9**, e102623.
Roche, D. G., Kruuk, L. E. B., Lanfear, R. & Binning, S. A. (2015). Public data archiving in ecology and evolution: how well are we doing? *PLoS Biology* **13**, e1002295.
Royle, J. A. & Dorazio, R. M. (2009). *Hierarchical Modeling and Inference in Ecology: The Analysis of Data from Populations, Metapopulations and Communities*. Academic Press, London.
Sauer, J. R., Link, W. A., Fallon, J. E., Pardieck, K. L., David, J. & Ziolkowski, J. (2013). The North American Breeding Bird Survey 1966–2011: summary analysis and species accounts. *North American Fauna* **79**, 1–32.
Schimel, D. S., Asner, G. P. & Moorcroft, P. (2013). Observing changing ecological diversity in the Anthropocene. *Frontiers in Ecology and the Environment* **11**, 129–137.
Schmeller, D. S., Mihoub, J.-B., Bowser, A., Arvanitidis, C., Costello, M. J., Fernandez, M., Geller, G. N., Hobern, D., Kissling, W. D., Regan, E., Saarenmaa, H., Turak, E. & Isaac, N. J. B. (in press). An operational definition of essential biodiversity variables. *Biodiversity and Conservation*. https://doi.org/10.1007/s10531-017-1386-9.
Schmeller, D. S., Weatherdon, L. V., Loyau, A., Bondeau, A., Brotons, L., Brummitt, N., Geijzendorffer, I. R., Haase, P., Kuemmerlen, M., Martin, C. S., Mihoub, J.-B., Rocchini, D., Saarenmaa, H., Stoll, S. & Regan, E. C. (2017). A suite of essential biodiversity variables for detecting critical biodiversity change. *Biological Reviews* https://doi.org/10.1111/brv.12332.
Schurr, F. M., Pagel, J., Cabral, J. S., Groeneveld, J., Bykova, O., O'Hara, R. B., Hartig, F., Kissling, W. D., Linder, H. P., Midgley, G. F., Schröder, B., Singer, A. & Zimmermann, N. E. (2012). How to understand species' niches and range dynamics: a demographic research agenda for biogeography. *Journal of Biogeography* **39**, 2146–2162.
Segata, N., Boernigen, D., Tickle, T. L., Morgan, X. C., Garrett, W. S. & Huttenhower, C. (2013). Computational meta'omics for microbial community studies. *Molecular Systems Biology* **9**, 666.
Skidmore, A. K., Pettorelli, N., Coops, N. C., Geller, G. N., Hansen, M., Lucas, R., Mücher, C. A., O'Connor, B., Paganini, M., Pereira, H. M., Schaepman, M. E., Turner, W., Wang, T. & Wegmann, M. (2015). Agree on biodiversity metrics to track from space. *Nature* **523**, 403–405.
Stephens, P. A., Mason, L. R., Green, R. E., Gregory, R. D., Sauer, J. R., Alison, J., Aunins, A., Brotons, L., Butchart, S. H. M., Campedelli, T., Chodkiewicz, T., Chylarecki, P., Crowe, O., Elts, J., Escandell, V., et al. (2016). Consistent response of bird populations to climate change on two continents. *Science* **352**, 84–87.
Stephenson, P. J., Brooks, T. M., Butchart, S. H. M., Fegraus, E., Geller, G. N., Hoft, R., Hutton, J., Kingston, N., Long, B. & McRae, L. (2017). Priorities for big biodiversity data. *Frontiers in Ecology and the Environment* **15**, 124–125.
Storch, D., Marquet, P. A. & Brown, J. H. (2007). *Scaling Biodiversity*. Cambridge University Press, Cambridge.
Sullivan, B. L., Aycrigg, J. L., Barry, J. H., Bonney, R. E., Bruns, N., Cooper, C. B., Damoulas, T., Dhondt, A. A., Dietterich, T., Farnsworth, A., Fink, D., Fitzpatrick, J. W., Fredericks, T., Gerbracht, J., Gomes, C., et al. (2014). The eBird enterprise: an integrated approach to development and application of citizen science. *Biological Conservation* **169**, 31–40.
*Sullivan, B. L., Wood, C. L., Iliff, M. J., Bonney, R. E., Fink, D. & Kelling, S. (2009). eBird: a citizen-based bird observation network in the biological sciences. *Biological Conservation* **142**, 2282–2292.
Thessen, A. E. & Patterson, D. J. (2011). Data issues in the life sciences. *ZooKeys* **150**, 15–51.
Tittensor, D. P., Walpole, M., Hill, S. L. L., Boyce, D. G., Britten, G. L., Burgess, N. D., Butchart, S. H. M., Leadley, P. W., Regan, E. C., Alkemade, R., Baumung, R., Bellard, C., Bouwman, L., Bowles-Newark, N. J., Chenery, A. M., et al. (2014). A mid-term analysis of progress toward international biodiversity targets. *Science* **346**, 241–244.
Turak, E., Brazill-Boast, J., Cooney, T., Drielsma, M., DelaCruz, J., Dunkerley, G., Fernandez, M., Ferrier, S., Gill, M., Jones, H., Koen, T., Leys, J., McGeoch, M., Mihoub, J.-B., Scanes, P., et al. (*in press a*). Using the essential biodiversity variables framework to measure biodiversity change at national scale. *Biological Conservation* https://doi.org/10.1016/j.biocon.2016.08.019.
Turak, E., Harrison, I., Dudgeon, D., Abell, R., Bush, A., Darwall, W., Finlayson, C. M., Ferrier, S., Freyhof, J., Hermoso, V., Juffe-Bignoli, D., Linke, S., Nel, J., Patricio, H. C., Pittock, J., et al. (*in press b*). Essential biodiversity variables for measuring change in global freshwater biodiversity. *Biological Conservation* https://doi.org/10.1016/j.biocon.2016.09.005.
Walls, R. L., Deck, J., Guralnick, R., Baskauf, S., Beaman, R., Blum, S., Bowers, S., Buttigieg, P. L., Davies, N., Endresen, D., Gandolfo, M. A., Hanner, R., Janning, A., Krishtalka, L., Matsunaga, A., et al. (2014). Semantics in support of biodiversity knowledge discovery: an introduction to the Biological Collections Ontology and related ontologies. *PLoS One* **9**, e89606.
Wieczorek, J., Bánki, O., Blum, S., Deck, J., Döring, M., Dröge, G., Endresen, D., Goldstein, P., Leary, P., Krishtalka, L., Tuama, É. Ó., Robbins, R. J., Robertson, T. & Yilmaz, P. (2014). Meeting Report: GBIF Hackathon-workshop on Darwin Core and sample data (22–24 May 2013). *Standards in Genomic Sciences* **9**, 585–598.
Wieczorek, J., Bloom, D., Guralnick, R., Blum, S., Döring, M., Giovanni, R., Robertson, T. & Vieglais, D. (2012). Darwin Core: an evolving community-developed biodiversity data standard. *PLoS ONE* **7**, e29715.
Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* **3**, 160018.
Witharana, C. & Lynch, H. (2016). An object-based image analysis approach for detecting penguin guano in very high spatial resolution satellite images. *Remote Sensing* **8**, 375.
*Wood, C., Sullivan, B., Iliff, M., Fink, D. & Kelling, S. (2011). eBird: engaging birders in science and conservation. *PLoS Biology* **9**, e1001220.
Yang, X., Blower, J. D., Bastin, L., Lush, V., Zabala, A., Masó, J., Cornford, D., Díaz, P. & Lumsden, J. (2013). An integrated view of data quality in earth observation. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **371**, 20120072.
Yang, Z., Wang, T., Skidmore, A. K., de Leeuw, J., Said, M. Y. & Freer, J. (2015). Spotting east African mammals in open savannah from space. *PLoS ONE* **9**, e115989.
Yilmaz, P., Kottmann, R., Field, D., Knight, R., Cole, J. R., Amaral-Zettler, L., Gilbert, J. A., Karsch-Mizrachi, I., Johnston, A., Cochrane, G., Vaughan, R., Hunter, C., Park, J., Morrison, N., Rocca-Serra, P., et al. (2011). Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. *Nature Biotechnology* **29**, 415–420.
*Zuckerberg, B., Fink, D., La Sorte, F. A., Hochachka, W. M. & Kelling, S. (2016). Novel seasonal land cover associations for eastern North American forest birds identified through dynamic species distribution modelling. *Diversity and Distributions* **22**, 717–730.

X. SUPPORTING INFORMATION

Additional supporting information may be found in the online version of this article.

Appendix S1. Information on projects.

Table S1. Principal steps of the information supply chain to build Essential Biodiversity Variable (EBV) data sets as applied to the Baltic Sea zooplankton monitoring (BALTIC) data set with indication of abundance records and taxa available for trend analysis.

Table S2. Detailed description of workflow steps used in the eBird project.

Table S3. Detailed description of workflow steps used in the Tropical Ecology Assessment and Monitoring (TEAM) project.

Table S4. Detailed description of workflow steps used in the Living Planet Index (LPI) project.

Table S5. Detailed description of workflow steps used in the Baltic Sea zooplankton monitoring (BALTIC) project.

(*Received 27 January 2017; revised 4 July 2017; accepted 5 July 2017*)