Best Practice Guide for Data Gap Analysis for Biodiversity Stakeholders

Arturo H. Ariño, Vishwas Chavan & Javier Otegui

September 2016

Best Practice Guide for Data Gap Analysis for Biodiversity Stakeholders

Arturo H. Ariño[1], Vishwas Chavan[2], Javier Otegui[3]

Table of Contents

[1] University of Navarra, Pamplona, Spain
[2] World Institute of Nature, Pune, India
[3] Formerly at Map of Life, http://mol.org

Executive Summary

The object of this guide is to elaborate on processes for conducting 'data gap analyses' (DGA) of biodiversity data in light of needs expressed by key stakeholder communities. DGAs will help in prioritizing data mobilization activities to meet these needs.

Biodiversity DGAs vary greatly in scope, depth and extent, and have particularities as compared to the more general DGA common in the business field, but share the common approach of detecting what data may be missing for whatever causes. Traditionally, biodiversity data are spatially explicit but as biodiversity is a multidimensional entity extending on concepts, space and time, DGA methods and techniques, while relying heavily on database and geospatial information analysis and representation, can be highly specific and extend into other areas as well such as time-series analysis or probability theory.

Whatever the methods, DGA's most conspicuous use is to inform policy- and decision-makers on what pitfalls need to be avoided because of lack of adequate data. A DGA will be successful if it can reliably point out such pitfalls, but even more if it can self-heal: it should show where plugs should be applied to gaps, and hopefully how. In doing so, DGA must itself avoid its own shortcomings, for a failed DGA may be of no use or, worse, of ill use.

General techniques for DGAs are often case-specific, so a general approach on how to conduct one will have to wait. However, lessons can be learned from existing exercises and a set of high-level, general principles can be derived that have been found to work. This is the purpose of this guide, which is not a manual to conduct a biodiversity DGA but an overview of such general principles to be borne in mind when designing a DGA.

To illustrate the general practices, sixteen actual DGA exercises have been summarized, and six, highly concerned with GBIF-facilitated data, have been discussed in more detail. This guide is intended for biodiversity data stakeholders: biodiversity information systems/networks, biodiversity data publishers, biodiversity organizations, research groups, information managers, national/regional/thematic biodiversity information facilities (BIFs), national/regional funding agencies, etc.

Section 1: Data Gap Analysis for Biodiversity Knowledge

Overview

'Data Gap Analysis' (DGA) is an essential step towards coordinated stewardship to ensure accessibility to appropriate, adequate, and fit-for-use primary biodiversity data to its stakeholder communities. Thus, the purpose of the 'Data Gap Analysis' is to identify discrepancies between current and ideal states (Research Data Strategy Working Group 2008) of the entire enterprise of biodiversity data management leading up to its publishing and usage (Chavan *et al.* 2010). "Ideal states" are either defined by stakeholders expressing their data needs, or deduced from the data needs and uses observed in the biodiversity-related literature.

Gap analysis for biodiversity knowledge includes, but is not limited to, a subset of the more general gap analysis prevalent in conservation planning, where the extent to which protection goals have been met in protected areas is evaluated through various methods. A "gap" in this conservation context is, in brief, the lack of representation or inadequate representation of a biotic element (plant community or animal species) in a map, in areas managed primarily for natural values (Crist & Csuti 2000). In itself, gap analysis for conservation is a customization of the even more general gap analysis applied to processes where a set of goals for an enterprise is compared to current, documented achievements through a series of metrics. This type of analysis is often conducted for the purpose of not only reviewing gaps, but also removing them through improving data collection.

Although protected areas may be a meaningful way to preserve biodiversity, their establishment does not guarantee that actual biodiversity is fully known even in these areas. It has been demonstrated that even if knowledge about protected areas actually exist, it may be spread over various, often disconnected sources, and tapping from all available sources dramatically increases the overall picture of the area. On the contrary, neglecting certain sources of data has often led to incomplete knowledge, resulting in inadequate management plans (Pino-Del-Carpio *et al.* 2011).

Extending the gap analysis for biodiversity knowledge into non-protected areas means dealing with fragmentary data and somewhat utopian goals. The ultimate knowledge about biodiversity would encompass anything related to distributions of species through space and time, knowledge of their genetic diversity, changes through time, the network of species interactions, even numbers or biomass, and all that at as fine a grain as possible and with complete accuracy and taxonomical reliability.

A data gap analysis (DGA) exercise, therefore, should always start with setting up both feasible and reasonable knowledge goals. Only then the extent of knowledge voids can be evaluated against what knowledge actually exists. As the main "customer" of biodiversity data for which gap analysis is required is the network of protected areas, it is natural to build data gap analysis from there, using the requirements and expectations of such networks as a starting point.

As technology improves and data accrue, however, goals can also be moved forward. Accurately mapping land cover at metric scale was a technically insurmountable problem until the advent of advanced remote sensing put it within reach. Therefore, a gap analysis will always be dynamic: once a gap is filled because of technology, effort, or knowledge have been put in place, then goals can be moved a step further, or improved, so a new set of gaps will appear between the state of the art and the newly desirable state of knowledge. That notwithstanding, it is hoped that the general principles guiding gap analysis may remain somewhat constant, allowing for repeated or cyclic application.

It should be noted that a DGA in biodiversity is *not* restricted to spatial gaps whatsoever. Gaps exist in data along many dimensions: space, time, taxonomy, subject, environment, etc.

(Peterson 2013). While the details of the practice of DGA will be shaped on the dimensions being analyzed, assembling a large number of techniques whose details fall out of the scope of this guide, a few general principles should be applicable to almost all DGAs that deal with biodiversity data. We will try here to address these general, overarching principles, and how to deal with them in DGA.

Data Gap Analysis: Why?

Scott (2000) summarizes the need for gap analysis around four key questions: Where do we stand today in the area of concern? Where are we headed? Where do we want to go? How will we get there?

The area of concern being biodiversity *data*, the rationale behind DGA is to assess what biodiversity data exist and can be used, how such data grow, and what data availability should be desirable. Then, the exercise gets into how to produce the desired data.

Thus, there are two main drivers behind a DGA: First, to know what data are available, and second, which additional data should be sought.

Biodiversity data are required in many endeavors—not only for conservation purposes. Understanding biodiversity leads to additional benefits, ranging from scientific knowledge to practical uses of biodiversity. Predictions, for instance, can help correct destructive trends or prevent loses (e.g. predicting the expansion of aggressive wasps may help protect pollinators). Such predictions increasingly rely on e.g. niche modelling (Peterson *et al.* 2011), and modelling typically requires training data. Although some algorithms have proven remarkably insensitive to sample size for certain types of errors (Peterson & Kluza 2003), in general models can be as accurate as the amount of training data available, and below certain thresholds it is difficult to rely on some models as their robustness is compromised (Rocchini *et al.* 2011).

A DGA comes to facilitate obtaining enough quality data to understand and model biodiversity. “Quality” here means that data required for this understanding or suitable for a specific modelling target will have to meet certain conditions, including but not limited to:

- Be enough to make valid (e.g., eventually statistically significant) inferences;
- Be appropriate and material to the target being sought;
- Become as free as possible of errors or misrepresentations;
- Be current or contemporaneous to the problem.

The problem of data trust and data reliability

Users of biodiversity data will require some idea about the reliability of the data they are using. Biodiversity data can be used as long as they are reliable. However, users actually place trust on the data: They assume that reliable data can be trusted—but lack of trust does not necessarily imply that data are unreliable. Trust may be subjective: for example, in a recent content need analysis most respondents placed trust in the researcher’s own data above trust on other’s objective data (Ariño *et al.* 2013), irrespective of whether the data were intrinsically reliable. Thus, reliable data can be trusted or not—but unreliable data should not be trusted. Providing some measure of the reliability of available data should therefore facilitate trust, and consequently, data use.

One important component of reliability is the degree of completeness of available data, although it has been shown that perception and measurable reality may not always coincide in this regard (Sousa-Baena *et al.* 2013). Trust can be built if the DGA can provide some estimate of the degree of completeness of the data.

Reliability should be applied both to the data themselves, and to the products derived from these data. DGA is one of the tools that can help deliver a reliability measure. For example,

DGA may help by pointing to data voids or issues that may affect how biodiversity data are used as a source of knowledge, especially after the gap analysis' products, e.g. a map of poor digitally accessible knowledge (Sousa-Baena *et al.* 2013) become ground truth (Scott 2000).

Data focus

This guide is concerned with **data** gap analysis within the biodiversity knowledge realm. Although the guide contains references to other concepts, the focus of the guide is specifically on the *data* component. It is not a general guide for gap analysis, which may include other concepts aside data availability. Gap analyses can be purpose-focused and, in fact, they should be conducted with a specific purpose for efficiency. But a data-oriented gap analysis may also be reused for different purposes and is thus a more basic tool onto which more comprehensive exercises may be built.

Although this guide makes frequent use of the concept of metadata, it is not a full guide about metadata. Biodiversity data knowledge will hardly ever be complete and using data in addition to models will always be required (e.g. distribution models). Modern models can address some of the classical uncertainties and biases in records. But they can best do so, if the modeler has additional information about the datasets, i.e. metadata that informs about possible uncertainties or sampling biases. Metadata treatment runs at a more specific level, and guidance on metadata gaps could be further referred to individual DGAs.

Role of Data Gap Analysis in a comprehensive strategy and action plan

Data gap analysis has three distinct roles in a strategy and action plan pursuant data mobilization:

1. Identifying categories of data that are not (yet) available but necessary for the formulated set of requirements of the stakeholders;
2. Identifying areas of data volume deprivation that prevent drawing fundamental or practical knowledge; and
3. Identifying quality components of available data that are not up to standards for meaningful discovery.

Once a data discovery strategy has been set up and the corresponding plans have been formulated, DGA becomes an exercise aimed primarily at bridging available data to required data. Therefore, the formulation of the strategy shall also require defining both ends: defining what data (and in what form and at what quality level) are available, and what data should be made available.

However, it should be noted that DGA should not be limited to such bridging, a popular idea in the more general gap analysis in other enterprises. Indeed, data stewardship indicators such as policies, funding, roles & responsibilities, trusted digital data repositories, standards, skills & training, rewards & recognition system, accessibility and preservation must be considered as part of 'Data Gap Analysis' exercise if they are not to be included in post exercise introspection (Chavan *et al.* 2010).

The Data Gap Analysis workflow

The six major steps of model DGA include:

1. Scoping the analysis and setting the expectations
2. Assessing the universe of accessible data
3. Identification of Data Gaps
4. Synthesis and dissemination of the outcomes

5. Prioritization of gap-closing, demand-driven data discovery and publishing activities
6. Evaluation of the DGA exercise

Section 2: Steps and activities in a DGA

Scoping the analysis and setting the expectations

It is essential to determine the scope of the Data Gap Analysis which is aimed at developing a comprehensive strategy and action plan towards demand-driven and deterministic data discovery and publishing for issue(s) under consideration. With regard to accessibility to fit-for-use data to address/resolve specific pre-determined issues from local-to-global significance, two questions that would determine the scope of 'Data Gap Analysis' include 'where are we now?', and 'where we want to be?'. For example, the scope of the 'Data Gap Analysis' can relate in a fairly simple way to an area planned for protection, or may be specified for the conservation of the specific targeted species or ecosystems and be descriptive of the desired number and distribution of occurrences of populations (Parrish & Dudley 2006). It is essential that institutions conducting Data Gap Analysis should clearly state 'what data gap analysis is for?' and 'what it will not answer?' to better manage expectations of stakeholder communities (Chavan *et al.* 2010).

Data expectations will need to be realistic, as data availability is subject to many hurdles. In the context of data required for environmental assessments, Fry *et al.* (2002) have identified several factors that prevent data availability from being perfect, such as outdated datasets, the bias in what data is available induced by the practitioners' skills or selective interests, or the inconsistencies arising from geographically overlapping datasets.

When scoping the DGA, a preliminary plan is developed that should also consider:

1. The resources: what is needed to conduct the DGA? Can it be done with in-house resources, or should it be outsourced? Can sections of the DGA be tasked independently? What is the cost in manpower, externalities, and expenses?
2. The deadlines and milestones: Are there identifiable points marking progress, e.g. completion of a geographical survey? What is the time frame for milestones? Are there time constraints that may suggest limiting the DGA?

Realistic scoping of the DGA is crucial to its success, and may be a point of failure if not done properly. The expected outcomes of the DGA must thus be within reach, and therefore should be stated as clearly as possible so as to prevent having to redefine them.

BOX 1 - Spectrum of main DGA potential customers

(blue: mainly primary data producers; red: mainly data consumers. Non-exclusive).

- Naturalists
- Field biologists
- Nature explorers
- Research institutions
- Citizen scientists
- Curators
- Ecologists, biogeographers
- Analysts, modellers
- Conservation planners
- Nature managers
- Policy makers
- Funding agencies

Expected outcomes

Expectations are closely related to the *operational* scope of the DGA: will it cover data quality, data availability, or both? At local, regional, or global level? Are one or more aspects of biodiversity data (i.e. taxonomical, spatial, temporal, environmental, ecological) to be analyzed? What is the main focus?

As the DGA is aimed at helping "know what we do or don't know", priorities should be set according to the main targets requiring the DGA. Among

biodiversity data stakeholders, prime customers for DGA range from the scientists producing the primary data used to assess biodiversity or fill the gaps, e.g. ecologists or field biologists, to conservation planners and managers using processed data and needing to know what is available at the far end of the spectrum.

Langhammer *et al.* (2007) list two main principles for setting conservation priorities:

- irreplaceability (or uniqueness) of sites, and
- vulnerability.

In an extreme example, a site is irreplaceable if a species can only live there. It is vulnerable if its biodiversity value is likely to be lost in the future. DGA should therefore be conducted primarily according to these priorities, as well as some subordinate ones such as complementarity, iteration, accountability and repeatability (Langhammer *et al.* 2007). Expectations from the DGA should be closely tied to how the exercise would help determine whether both main priorities can be met.

The two main constraints for expectations, apart from conceptual feasibility, are:

- resource availability, and
- time frame.

For example, let's suppose that the DGA is aimed at providing complete coverage for the distribution of indicator species, e.g. soil microarthropods, in a network of protected areas (PAs). Conceptually, sampling could be done to assess what species are present in the areas. But this can be an extremely time-consuming task for large, heterogeneous areas. A literature or database survey may eventually yield presence data (signifying that the gap for soil indicators may be closed) or no data, in which case the gap might still exist—essentially, the same outcome of the sampling but with certainty for positive results and uncertainty for negatives, whereas the sampling method would provide probability (not uncertainty) for negatives. Therefore, the expectations of the DGA should be adjusted to its feasibility.

Ultimately, a DGA should reveal what data are accessible, whatever its means. However, an increasing demand is for digital accessibility (termed *ready access*). Sousa-Baena *et al.* (2013) distinguish digitally accessible knowledge (DAK) to include sources of data that can effectively be accessed through digital means.

BOX 2 - Expectations in a DGA according to priority and feasibility

	Priority	Feasibility
Assessing raw data quality	Medium	Hard
Assessing fitness-for-use	Medium	Easy
Identifying under-represented areas	High	Medium
Setting baselines	High	Easy
Establishing timelines	Low	Hard
Determining completeness	Medium	Hard
Measuring DAK	Medium	Easy
Estimating gap-filling field costs	Low	Hard
Digitization (DAK) likelihood	Low	Medium
Estimating DAK costs	Low	Hard
Assessing irreplaceability of sites	High	Medium
Assessing vulnerability of populations	High	Hard
Inventorize resources	High	Easy
Ensuring repeatability of Inventories	Medium	Hard

Determining baselines and indicators

A key outcome of DGA should also be its ability to assess whether change has occurred or may occur. Therefore, it is critical that the DGA exercise can proceed from previous knowledge (a *baseline*) upon which change can be measured. However, it will often be the case that baselines do not exist yet, and in this case DGA is precisely the tool to set such baselines for subsequent DGAs. In fact, discovering the absence of a baseline should be an important factor to consider starting a DGA.

Thus, when setting up a DGA one should consider, in sequence:

1. Are there already available data about a site, in whatever form?
2. If not, can expert judgement be called in?
3. If not, can the DGA provide the baseline?

Procedurally, the baseline setting should proceed as in Figure 1.

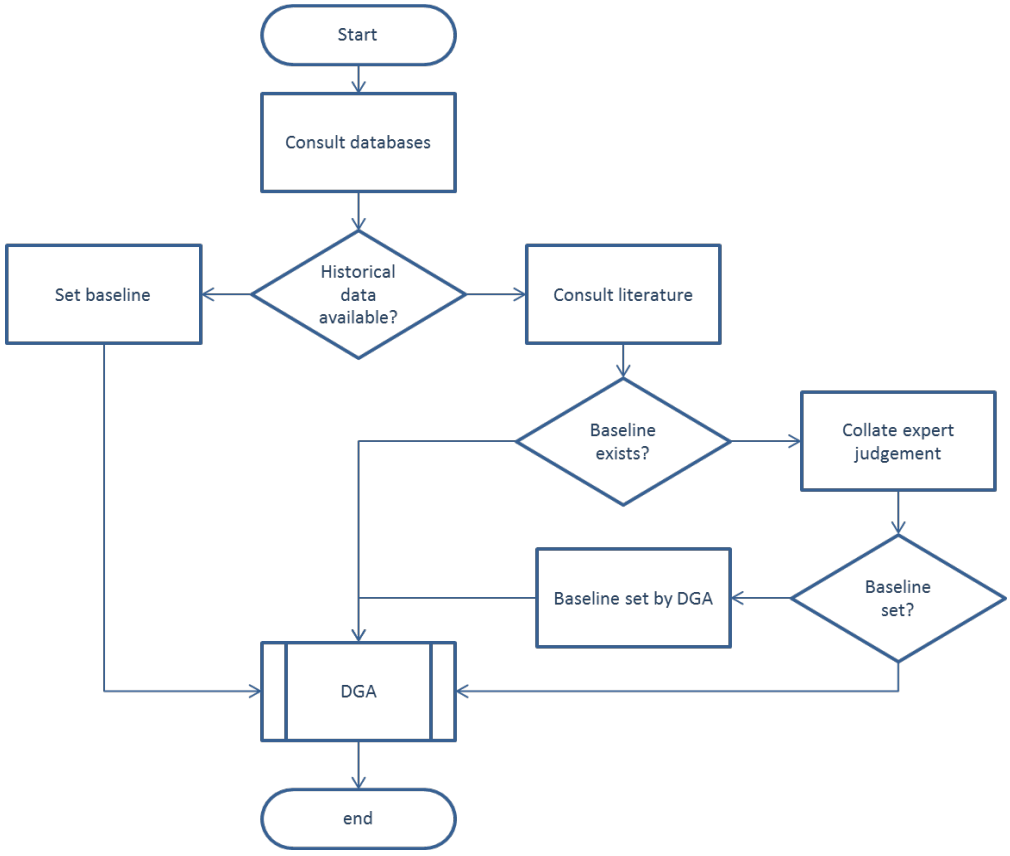

Figure 1. Procedure to set data availability baselines within DGA.

Assessing the universe of accessible data

To achieve optimal and realistic state-of-the-art understanding of 'what is accessible' and 'what is needed', it is a must to have a list of all accessible and/or available data resources. This we call 'estimating the universe of data', a comprehensive exercise that inform about five prime concepts: (a) who has what type of data, (b) in what form or format, (c) in what state of digitization, (d) if digitized, whether accessible or not accessible, and (e) its state of fitness-for-use to derive solutions for pre-determined issues. It is advisable that data resources which are currently not in digital form (e.g. natural history collections) also be inventoried. Given that such an inventory of data custodians & publishers is useful in data mobilization and publishing activities as well, we recommend that the best mechanism to

assess the 'estimate of universe' is to be built using existing metadata catalogue(s) if available, or else consider seriously to develop a metadata catalogue that can document the descriptions of data resources and help set priorities for such mobilization (Berendsohn & Seltmann 2010). Completeness of metadata documents that constitute such a catalogue will determine the degree of success of remainder activities of 'Data Gap Analysis' exercises (Chavan *et al.* 2010). Ancillary elements may complement, or be part of, the prime concepts above and contribute to their specific completeness: for example, "type of data" may in turn include qualifiers such as taxonomical validation, or the availability of associated digital publications may qualify the "state of digitization" of the resource.

Digitally accessible data

Available data can exist in digital or analog form. The metadata catalogue will refer to existing data, be it in analog or digital form, and should state what form the data are in. Digital does not necessarily mean easily accessible: for instance, digital data can be behind a security or pay wall, or be kept private, or be merely unknown. Digitally accessible knowledge (DAK) (see previous section) can be readily accessible (e.g. a shared database with clear metadata and standard format) or require some facilitation (e.g. indexing, linking, or authorization), although this extra step makes them less of a DAK. For DGA purposes, however, we will maintain a difference between knowledge that can be accessed digitally (either directly or after some facilitation, including digitization), from existing knowledge that cannot be accessed because a barrier prevents such access. In turn, the existence of this knowledge may be known (we shall term it *locked knowledge*, LK) or not known by other stakeholders (*buried knowledge*, BK).

BOX 5 - Some definitions

DAK – Digitally Accessible Knowledge: Primary data that are both digital and accessible in standard formats

LK – Locked knowledge: Data that are known to exist, but cannot be accessed because of some barrier (e.g. paywall, obsolete digital systems, inability to digitize)

BK – Buried knowledge: Data that exist but whose existence is not known or cannot be ascertained by users.

DGA is thus concerned with DAK to establish corresponding baselines, and will seek to estimate whether there are LK or BK as part of the closure exercise.

It should be noted that, in the context of biodiversity data, DAK can be extended to the structured analog data that can be readily converted into digital form, for instance through capture or OCR of tables already kept as digital images or files (e.g. reports in PDF format).

DGA is thus most efficient if restricted to DAK, but subsequent DGA iterations can progress to other types of data as well.

Accessing data needed for exercise

DGA covers gaps in data. Therefore, the choice of methods to access data may determine the outcome of DGA if selective (some specific data may be discovered by specific methods). Thus, it is important to be aggressive at data mining, and often more than one method should be tried, covering possible sources. On the other hand, some sources may be costly in terms of time and resources (e.g. having to conduct surveys or purchase RS), and a correct scoping of the DGA should help finding the balance between desired data and obtainable data.

Potential sources of hard data (e.g. collected primary biodiversity data) should include:

- Databases and data aggregators
- Literature (digitized and legacy), including data papers
- Reports, management programs, grey literature

- Remote sensing products
- Citizen science output
- Surveys

In addition, data could be derived from secondary products such as distribution models and secondary classification of remote sensing. For example, the Arizona Gap Analysis team used an airborne video camera in conjunction with field sampling to develop a large number of training sites, inexpensively, in order to derive the corresponding discriminant function for vegetation cover classes based on the properties of those classes (Mulder 1988). It should be noted, though, that data derived from secondary products may be subject to interpretative issues (e.g. methodological constraints and their resulting uncertainties) that might limit their integration with "hard" data.

Methods for assessing universe of data

Among other possible criteria, existing or potential data can be categorized according to how difficult it would be to get at them, and what form they are in (Figure 2).

		Access/process difficulty		
		Easy	Medium	Hard
Capture or existence form	Digital	Databases Indexes Digital inventories CS output	Unstructured files Maps, digital RS Survey results	Locked files Unknown files
	Analogical	Imaged reports, tables Imaged museum data labels, structured ledgers	Unscanned papers Old imagery Unmarked literature Field notes	Locked, private collections Forgotten or unknown collections
	Future	Automated surveys and monitoring New CS endeavors	New field surveys Planned RS	Unsampled remote or inaccessible sites Unknown organisms

Figure 2: Access to, and forms of, biodiversity data. Black: DAK. Blue: LK. Red: BK. Green: Mineable data. Grey: Future data.

DGA needs to assess what data are already available, and what data could be made available (at the corresponding costs) before the gaps can be found. A successful DGA will measure the relative difficulty (and/or costs) of getting to the upper left square of Figure 2.

Determining the universe of available data, whether accessible or not, entails making assumptions about the proportion of known data within the total data. Techniques that look at data existing in overlapping formats or systems, and that have been collected differently, can be used at advantage to statistically deduct what data are missing.

Pino-Del-Carpio *et al.* (2013) demonstrated that similar data retrieved from different sources were complementary, and that no data source fully represented the available data. For instance, biodiversity data of biosphere reserves can be found in at least three independent data sources: scientific literature, management plans, and databases facilitated through GBIF. On average, a protected area will yield only about half- to two-thirds of biodiversity

data appearing in all three sources, and the rest will appear in any two, or any one exclusively.

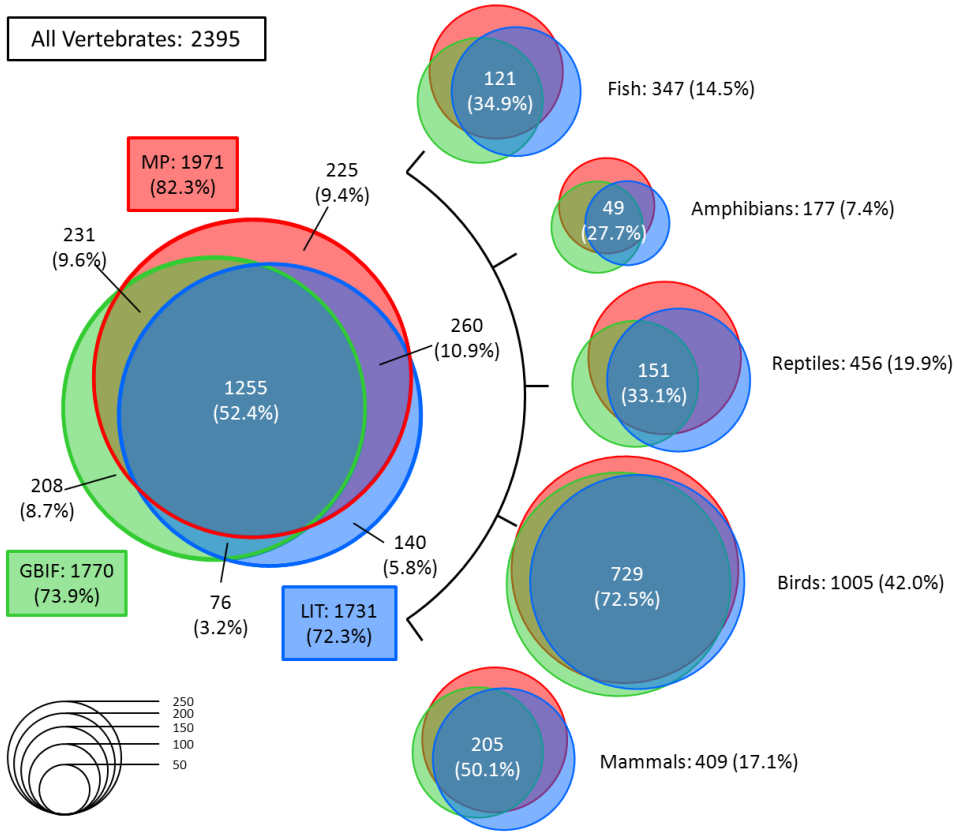

Figure 3. Number of vertebrate species listed in separate sources (GBIF, management plans, literature) for the network of protected areas in Mexico. Each source listed several species unknown by other sources. The figures can be used to estimate the number of unknown species potentially noticed by unlisted sources (Pino-Del-Carpio et al. *2013).*

In addition to data known from at least one source, there might also be data that have been escaped detection, e.g. they are in a different source. Ariño (2010) proposed using probability theory to find the size of the unknown portion of the universe of data based on data sharing sources. The proposed generalized Seber method, for instance, could be used to determine the size of existing universe of data.

Collections in the universe may have been unlisted (R_0), listed in repository A (R_A), in repository B (R_B), in C (R_C), etc., of L repositories, and therefore may have been also simultaneously listed in more than one repository (for example, R_{AB} or R_{ABC}). Assuming that listings are independent from each other; that there are $m(x: 0, A, B, C, ..., AB, AC, ..., ABC, ...)$ possible outcomes; and that a collection has a probability Θ_x of belonging to the outcome X, all possible outcomes follow a multinomial distribution

$$f(R_0, R_x, ..., R_m \mid R) = \frac{R!}{R_0! R_x! ... R_m!} \theta_0^{R_0} \theta_x^{R_x} ... \theta_m^{R_m}.$$

which is the generalized version of the particular case for four outcomes proposed by Seber and Felton (1981, eq. 11). Therefore, similar substitutions can be made as in that case, resulting in a generalized analogue where the correction factor is the product of the probabilities for a collection of not belonging to one repository when belonging to any other. Thus, the case with L=3 has m=8 outcomes and is

$$k = \frac{R_B + R_C + R_{BC}}{R_T - R_A} \frac{R_A + R_C + R_{AC}}{R_T - R_B} \frac{R_A + R_B + R_{AB}}{R_T - R_C}$$

where R_T is the number of collections belonging to at least one repository. The general case is

$$k = \prod_{S\{A,B,...L\}} \left[\frac{R_T - \left(R_{Si\{A..L,-S\}} + R_{Si\{B..L,-S\}j\{C..L,-S\}} + ... + R_{ABC..L} \right)}{R_T - R_S} \right].$$

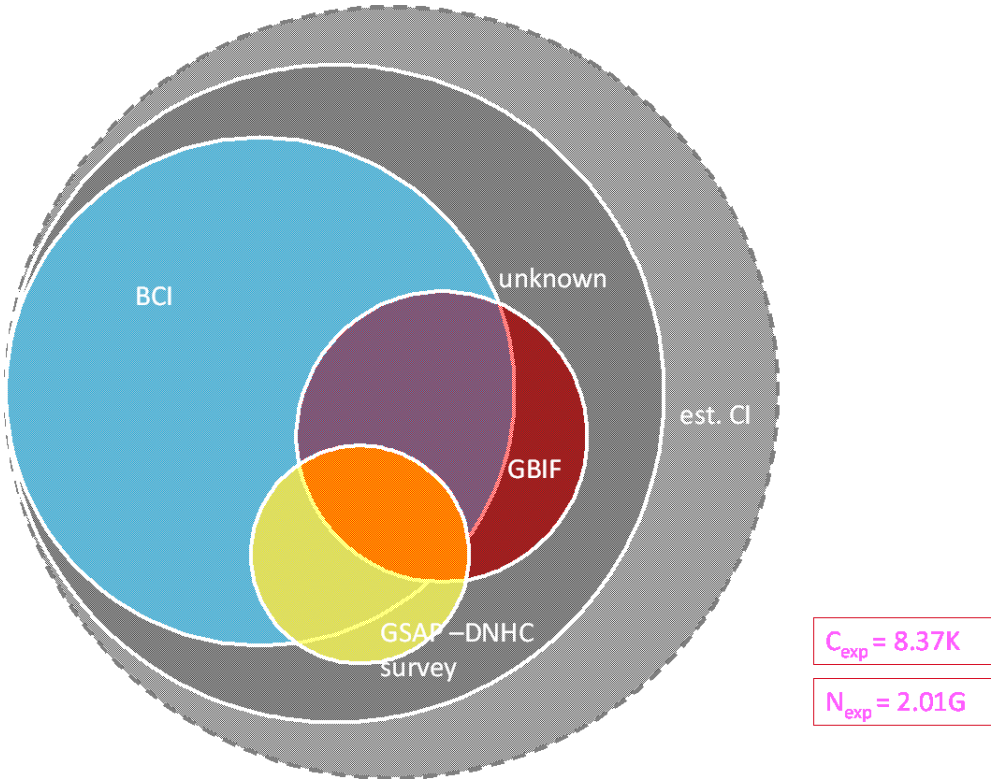

Figure 4. The generalized Seber-Felton method for estimating the universe of undiscovered data from listed repositories (Ariño 2010) and its application to the estimation of undiscovered natural history collections.

A second method for assessing the universe of potential data relies on surveys conducted through the community. The success of such surveys will depend on their correct implementation, which will seek to maximize reach and return for optimal representativeness. Reaching the community to be surveyed will require elaborating lists of potential interviewees and wide dissemination of the survey effort. Often, the survey will be part of a more general inquest, to which it can piggy-tail, or data could be gleaned from another survey's body of responses.

Content Needs Assessments (CNA) are candidate surveys that may help assessevidence the universe of data. Best practices for CNA--see Faith *et al.* (2013) and Ariño *et al.* (2013)--may help design a survey that could yield data for this exercise.

Data Resources Discovery System

One of the biggest challenges in effective use of biodiversity data is pinpointing the location of useful data. A Data Resources Discovery System (DRDS) seeks to provide users with ways to substantially increase their ability to discover and access relevant biodiversity information and data resources. The DRDS enacts a registry of links to data, enabling their location in a distributed knowledge system. Without a discovery system, data that may exist may also remain unreached not because they cannot be reached, but because knowledge about its existence (and access to the data themselves) is impaired. A core component of a DRDS is the metadata catalogue (see below) where resources are listed.

GBIF's Data Resources Discovery System

One of the largest available DRDS is GBIF's Global Biodiversity Resources Discovery System (GBDRS). It currently serves as: 1) a registry of resources and services and 2) a set of discovery services interacting with existing infrastructure such as GBIF to facilitate the discovery of biodiversity information. The most important component, the Registry, would facilitate the inventory of information resources by creating a single annotated index of publishers, institutions, networks, collections (datasets), schema repository and services. The GBRDS was not conceived to be simply a collection of centralized indexes but an integrated 'Yellow Pages' reference of all biodiversity information resources, reconciling all distributed resources and providing a meaningful way to discover them in a distributed manner (Chavan, Sood, *et al.* 2010).

Role of Metadata catalogue as Data Resources Discovery System

Simply put, a metadata catalogue is a list of available resources, with information describing them. In this context, metadata – meaning "data about the data" – refers not to information contained in the resource but information about the resource, data that describe the dataset, putting it in a particular context and improving its discoverability. The most common pieces of metadata are:

- Resource identification: title, rationale, identifiers such as a Digital Object Identifier (DOI).
- Information on authorship: who collected the data, who managed the information, who described the resource.
- A set of keywords or search and discovery optimization tools, such as geospatial, temporal and/or taxonomic coverages of the dataset.
- Access and usage licenses applied to the dataset.
- A means for accessing the dataset: a URL, a citation, contact information, and
- A means for accessing further information about the dataset, e.g. links to publications describing or using it.

Since one of the main goals of a metadata document is to improve discoverability of a dataset, it makes sense to develop unified, common ways of describing datasets. Currently, there are some standard ways of building these metadata documents. The Ecological Society of America created the "Ecological Metadata Language" (EML) standard as a way of unifying the way ecological data sets are described, and it is widely used. GBIF's Integrated Publishing Toolkit (IPT) uses EML documents to store metadata about the provenance of a dataset (in DarwinCore Archive format), and a different custom type of document (called meta.xml) to describe the contents of the dataset (the different types of information it contains, relationship between measurements and observations). Please refer to the DarwinCore archive documentation (http://rs.tdwg.org/dwc/terms/guides/text/index.htm) for more information.

Identification of data gaps

Mapping data expectations to data availability

With the needs of the user community identified through the 'Content Needs Assessment', and the extent of the 'universe of data' estimated and understood, we now need to superimpose one on the other to assess whether or not the available data are enough to fulfil the needs expressed by the user community/communities. This is commonly known as 'mapping expectations to availability', and the answer to this question will lead to a tailored set of strategies to tackle this issue and develop a solid action plan for identifying data gaps.

Such mapping, due to the ever-changing nature of data availability, has to be a continual process that needs to be conducted at regular intervals. The incomplete and in-progress state of biodiversity studies in different regions of the world continually generate new data, well-established biodiversity institutions (such as Natural History Museums) will keep gathering new biological information, and existing collections will perform cleaning tasks on their records that will lead to better, more complete and more comprehensive information. Thus, this mapping effort has to be done periodically to ensure a proper understanding of the information landscape. Even when the available information is not enough for the current needs, a continuous mapping exercise as new or better data become available can provide realistic insights into the efforts required to bridge that gap, to address pre-determined issues or data needs of a particular stakeholder community (Chavan *et al.* 2010).

It could be argued that data expectations could grow to encompass the whole inventory of biodiversity in an area. The mapping exercise could yield a two-level gap, e.g. one between mobilized data and existing data for an area, and another between covered species and presumed true richness in that area. Assessing only the difference between mobilized and existing data would underestimate the most severe data gaps in regions where even existing, un-mobilized data is extremely scarce. While the first level can be covered through mobilization efforts, the second level can only be solved through field work, producing a "sample" data that effectively represents reality and thus closing the gap. Mapping data expectations vs. data availability may thus help set priorities to mobilization or data collection, according to the likelihood of reducing the gap.

The need for a Data Gap Analysis

According to a recent world-wide survey on the biodiversity community users' needs (Ariño *et al.* 2013), taxonomic data and occurrence records are the two most demanded types of data, but people put more trust on their own data rather than records accessed through an aggregator. GBIF is currently the largest aggregator of these kinds of data, but the fact that these records come from a variety of sources makes it almost impossible to ensure a proper degree of network-wide reliability at the source. This fact can easily make the use of

aggregators a double-edged sword, feeding the now classic dilemma of few-but-good vs. lots-but-doubtful information.

Users relying solely on their own collections, trusted and well-known, will benefit from an easier understanding of the benefits and drawbacks of their records, but their scope of analysis can be somewhat limited. On the other hand, users who make use of data source aggregators will find a more extensive landscape of available information, at the expense of a tougher data processing workflow. Many times, however, investing in procedures to ensure a proper degree of quality from a larger set of externally accessed data pays off. Nevertheless, DGA procedures are needed even in the case of working with well-known data sets. Until the expectations are explicitly stated, no one can know precisely whether or not the available data will be enough, or if they have an adequate degree of precision.

Properly mapping expectations to availability can give clues on how to focus the DGA process and where to put efforts in order to maximize benefits and the lower cost.

Methods of Data Gap Analysis

Gap analysis, in its most common form related to biodiversity, was introduced by Scott at the end of the 20th century (Scott *et al.* 1993). Its primary use is to identify gaps related to conservation, and its field of action is generally the network of protected areas: they were designed as an assessment of the extent to which a protected area system meets protection goals set by a nation or region to represent its biological diversity (Dudley & Parrish 2006).

Methods for this type of gap analysis are well established, and have even been made official in certain administrations (e.g. U.S.) as a standardized tool to calibrate the level of achievement of conservation goals in protected or natural areas. Usually, these methods entail geographical analysis:

- Representing a parameter of interest over space,
- Inserting the physical limits of the regions of interest to which the gap analysis is to happen,
- Extracting the regions of interest.
- Representing the desired state,
- Finding the differences,
- Representing the differences over space as a map.

On the other hand, DGA assesses data, and therefore is largely an exercise about accountability. In most cases, access to databases holding the data (from simple spreadsheets to large database managers) is mandatory. The general workflow is:

1. Identify the available data
2. Collate the relevant sections of the data
3. Arrange data in an way amenable to analysis:
    a. Flat model
    b. Relational model
    c. Open-ended model
4. Attack the data with statistical or analytical tools (Excel, statistical packages)
5. Get results in representable form
6. Produce representations of data (plots, maps, tables, charts).

A list of methods for conducting DGA currently in use include:

- Published data reviews:
    - Scholarly literature review
    - Report discovery
    - Data mining
- Species distribution assessments:

    - Inventory collation
    - Field sampling
    - Database compilation and analysis
    - Remote sensing
    - Distribution modelling
- Community surveys:
    - Content needs assessments (CNA)
    - Ad-hoc surveys

Three methods stand out in determining data gaps: database analysis, user surveys, and literature review. Other methods can generally be thought of as accessory, often being part of, or complementary to, the main three. The availability of a large volume of primary biodiversity data records (PBRs) enables database analytics, and PBRs can thus be related to each other through visualizations and statistics designed to reveal continuities and voids in space, time and taxonomy. Literature review, although very time-consuming, can be highly profitable in terms of assessing with high certainty what is known (and, therefore, what is not). Content needs assessments done through surveys can be biased, but adequate preparation can reduce the possibility of error (Ariño *et al.*, 2013)

Advantages and disadvantages of various methods

Methods differ on how data are collected and treated, what data they are primarily concerned with, and the relative difficulty of execution or of reaching the expected outcome (see Scoping section for a list of feasibility of DGA outcome expectations).

Data used can be raw or processed, primary or secondary; and hard or easy to come by. A prime consideration in choosing a method is the likelihood of getting the right data timely.

On the other hand, methods differ on their advantages and disadvantages but can also differ on the extent of such advantages, as several aspects contribute to the advantages. Even a single aspect may play down a relative advantage if the method chosen is not carried out carefully. A judicious selection of methods will ultimately weigh their relative advantages according to the feasibility (a combination of resource availability, data requirements, and execution difficulty) and the needs disclosed by the stakeholders.

Main method	Literature review	Database analysis	Stakeholders survey
Resource use	Medium	Intensive	Not intensive
Costs	High*	Medium	Low
Difficulty of execution	Medium—requires deep understanding of the field	High	Medium—requires a carefully prepared survey
Time required	Long or very long	Medium	Short
Requirements for data	Low	Very high	Low
Reliability	High	High	Medium
Persistence	High	Medium	Low—changes with shifts in interests
Uncertainty	Low	Low	High – very dependent on the survey characteristics and size of the sample

Error likelihood	Low	High—dependent on the quality and standardization of data	Low
Bias	Medium	Low	High

**Unless subsumed into the institution's running costs.*

Figure 5. Advantages and disadvantages of various methods used in DGA.

Target-specific limitations

Individual gap analysis exercises may also be limited by the nature of the target. One example is provided by (Scott *et al.* 1993), who list a number of caveats about vegetation maps compiled for gap analysis. Among them:

- Microhabitat elements that may be smaller than the polygons chosen to represent vegetation;
- Generalized lack of indication about the age or successional stage of the units;
- Sharp boundaries between vegetation plots, with no provision to represent ecotones or transitions.

Differences between Needs Analysis and Data Gap Analysis

Although CNA results can be used in a gap analysis exercise, DGA more properly focuses on what is missing towards an end, whereas CNA helps set that end by asking the right questions. CNA is the tool that may best reveal what is the desired state of knowledge (as opposed to the current state, which may be just a reflection of what practitioners can do), for its results come from an analysis of what the community knows they need in order to pursue e.g. conservation objectives. Once the goals are clearly stated, DGA can measure the corresponding gaps.

Tools and resources used for Data Gap Analysis

Data gap analysis has a specific set of tools, and software exist to assist in this endeavor (Scott, 2007). However, data gap analysis is largely an analytical exercise relying on a set of technologies and tools used in an ad-hoc manner. Among them are:

- Database engines
- GIS and mapping tools
- Spreadsheets
- Statistical packages or languages (e.g. *R*)
- Web services
- Visualization tools (e.g. BIDDSAT[4], VESPER[5], GBIF dashboard[6])

The tools to be used in the DGA exercise will be largely by the practitioner's choice, likely according to personal preferences. However, a number of criteria should be universal enough to assist in the choice of tools:

[4] http://www.unav.es/unzyec/mzna/biddsat

[5] http://www.soc.napier.ac.uk/~cs22/vesperDemo/vesper/demoNew.html (Graham & Kennedy 2014)

[6] http://www.gbif.org/analytics/global

- **Efficiency (time required to get meaningful results**). A DGA may be useless if it misses its target, e.g. a bill with a set date.
- **Error control and error reduction**. A DGA that contains errors is misleading, potentially doing more harm than good if the error level overturns the DGA conclusions in absence of error. Tools should not be error-prone. One classical example is excessive manual input/copy/paste (as opposed to referencing or linking) between data sheets during the analytical flow.
- **Level of precision.** Low precision levels (e.g. tools operating over broad aggregations of data) may make the DGA miss the signal being sought.
- **Accuracy.** Tools used to represent the data must do so accurately in order to facilitate, e.g. pattern detection.
- **Information content.** DGA methods and tools should focus more on useful information contained in the records (e.g. richness) than in simple numbers (e.g. repeated records).
- **Enhancement of the level of trust, confidence, or reliability.** A high level of reliability may lower requirements for contrasting or supporting data, thus enhancing efficiency.
- **Repeatability**. DGA should lead to similar conclusions if repeated for similar datasets.
- **Reusability**. Once a DGA method has been established, it should allow for different sets of users to reuse the method, or for adaptation to different sets of purposes.
- **Expandability.** The DGA method should be scalable and cope with larger or wider datasets without fundamental changes.
- **Customization options.** The DGA should be flexible enough to allow adaptation to particularities in the datasets or the targets.

How to assess the trends/patterns of data digitization and publishing

The DGA exercise should provide a baseline of available data to measure the gap against desired data. However it should also estimate how the gap could evolve according to the evolution of the availability of data. Thus, assessing patterns and trends in the data availability (e.g. how has it proceeded through time and space, and at what rates and from what sources) should provide insight into what can be expected from data capture action plans.

DAK in highly structured form, such as the content of the GBIF index, can be analyzed for evolution through time. shows the increase in the total size of the database. A regressive model can be fit to the data, whereupon the main consideration would be to decide which one is the best model. Linear models would predict constant-rate increase of data, whereas exponential models would predict increased-rate accrual. Any model should include an estimate of its reliability, in terms of confidence interval for the model's parameter(s).

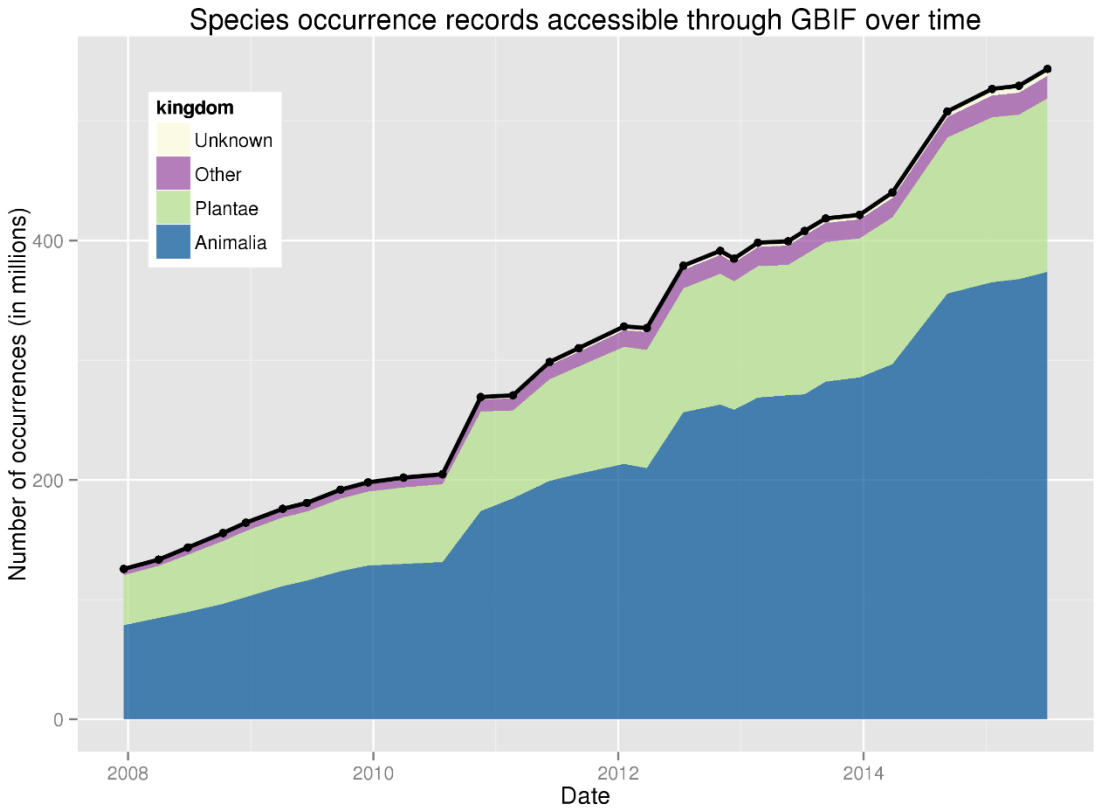

Figure 6. GBIF-mediated records from 2008 to July 2015.

However, in a DGA the patterns should be restricted to the area of interest. Therefore, data should first be filtered according to what is being assessed in the DGA. For example, regional DGAs should filter data based on georeferencing, or by locality name. It is often the case that general trends do not hold true for localized trends.

How to identify inconsistencies between demand and supply

The mapping exercise as described earlier will lead to identifying the gaps in accessible data, as well as their limitations in addressing issues that the stakeholder user community wishes to address. This revelation of inconsistencies between demand and supply of primary biodiversity data will lead to prioritization of activities ranging from collections of data to their publishing, resulting in free and open access to data (Chavan *et al.* 2010).

A reasonably simple yet efficient method to represent the mapping can take the form of a XY plot, where available data (Y) are pitched against expected data (X) for a set of instances. When properly identified and ordered, gaps should appear as clusters of data instances deviating from a regular pattern.

By way of example, let's assume that research on a given taxonomical group is expected to produce data, and that gaps (in this case, DAK gaps) may exist if such data have not been made available. We could compare the expected number of data records derived from the amount of available potential data sources such as papers, to the amount of actually available data records for the same time frames. Gaps would be represented by data points falling towards the lower right quarter of the plot.

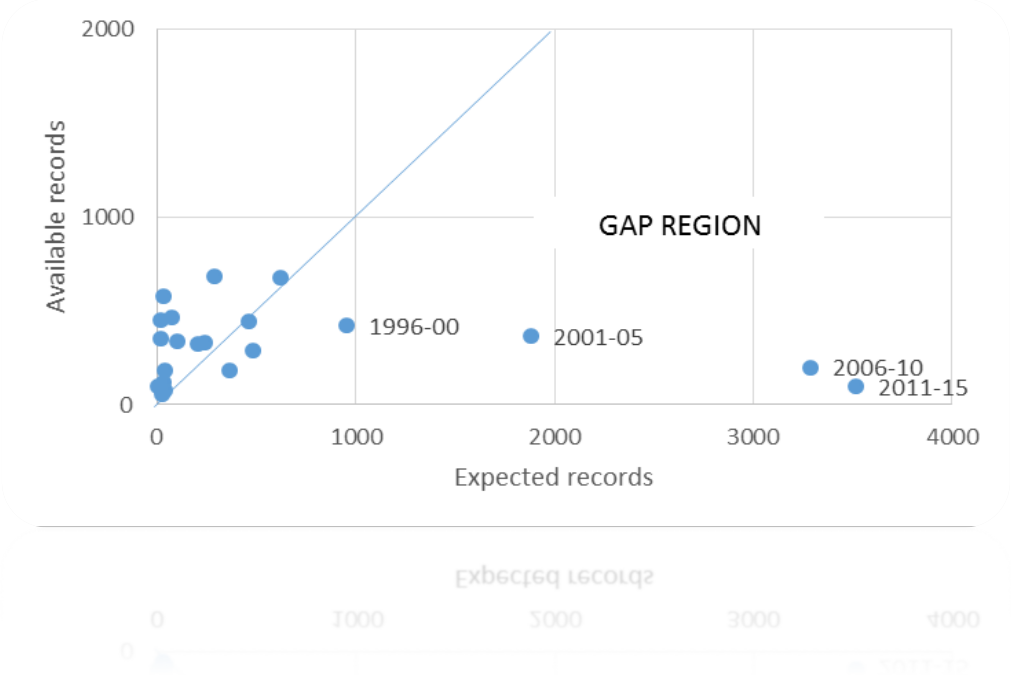

Figure 7. Expected vs available DAK for the genus Chamaeleo in the GBIF index by five-year intervals over the last 100 years. Last 20 years of research should have produced data that are not (yet) available, representing a gap in either digitization or release.

BOX 4. An example mapping of inconsistencies

Hjarding *et al.* (2014) compared digitally accessible distribution data with expert knowledge about chameleons, suggesting that expert data might be locked and missing from DAK.

We set to test this by examining how record availability compared to literature in a rather simplistic way. The underlying approach was that research on the genus should be expected to produce data that would eventually make their way as formal records, ultimately becoming DAK. Gaps would exist if such records were not produced or not made available.

For each five-year period over the past 100 years, we noted how many papers referred to *Chamaeleo* on Google Scholar (P_i), and how many records existed in GBIF (R_i). Thus, the quotient $Q_i=R_i/P_i$ could represent the average data records expected from the average paper in the period. In order to make the quotient more robust, we discarded the extreme deciles and obtained the overall Q as the average from the central deciles.

Next, we multiplied each predictor P_i by the average (over the century) Q to obtain R_e, the expected number of data records for each five-year period, and plotted R_e vs R_i. A good match should be a cloud along the diagonal line, while gaps should detach as a deficit of available records.

This example shows a distinct gap regarding recent data as opposed to older data, but other dimensions are amenable as well: for example, groups could be formed geographically or taxonomically, among others, for other types of gaps.

Interpreting the results

Published biodiversity gap analyses (see section 4) have nearly always found gaps, so it can be safely assumed that this is the expected outcome by default. Often, though, gaps might

be relative. When data instances are compared (see example above), instances below expectations will be identified as gaps. Thus, the DGA exercise may actually risk becoming a comparison exercise whereby the scarcest data sections will default to gaps, irrespective of whether other (but less scarce) data sections may also constitute a gap by right.

Care should be thus applied in interpreting the result of a gap analysis. Wherever possible, expectations about data should be assessed independently of available data. If this is not possible, gaps found should be interpreted as relative. This does not invalidate the DGA, as gap filling actions resulting from the DGA will still close the relative gaps, in turn allowing for other (relative or absolute) gaps to be found.

Two other common issues in DGA that may affect how to interpret the results are large variance and scale dependency. High variance is often associated with inadequate sampling, where e.g. quadrats contain widely different number of samples or data points; gaps found under such conditions may be due more to the irregular sampling effect than to actual lack of DAK (Peterson *et al.* 2008). On the other hand, variance is often dependent on scale. When dealing with spatially-explicit data, areas may show gaps at a given scale, but show too much variance to be meaningful at smaller scales or loose information (through data homogenization) at larger ones (Idohou *et al.* 2015). Thus, understanding the scale effects may be critical to interpret whether gaps found are significant.

Finally, DGA results should allow for interpretation of the potential usability of existing data. Existing data that cannot be used (for example, data that are known to be taxonomically wrong) should in effect be interpreted as a gap. However this should not be confused with either data that may be unfit for specific purposes, but fit for others, or data that are insufficiently described to assess its fitness-for-use (Hill *et al.* 2010), e.g. taxonomical lists not attributable to a set or known taxonomy. Thus, assessing fitness-for-use, a separate exercise, may be needed to qualify gaps in the DGA as broad or specific, and perhaps help prioritize gap-filling.

Best practice in interpreting should therefore include an assessment of whether any gaps found are affected by the scale, distribution, and usability of the recorded data.

Synthesis and dissemination of the outcomes

The DGA may be useless if it does not reach its intended targets: those that can fill the gaps, and those that need to know what those gaps are. Synthesizing and publishing the DGA is therefore important, as is ensuring that the relevant stakeholders are aware of the DGA in a manner that facilitates action.

Considerations for optimal communications

Ideally the DGA should be comprehensible without ambiguity and in whole, which means that effort should be dedicated to synthesis. A multi-level approach could be considered: All relevant sections of the DGA should be summarized, e.g. as self-contained items (bullet points), that can be expanded in another section. As most information is now conveyed through hypertext, a possible strategy would be to prepare a narrative no longer than an executive summary, where all relevant bullet points link directly to sections that expand as necessary.

Biodiversity data often lends itself to graphical representation, and this, coupled with the pattern detection ability of the human brain, calls for as much graphical synthesis as possible. Where possible, plots, maps, and graphs should be used in preference to descriptive text.

However, care should be applied to avoid mis- or over-representation of results. Even though the objective of a DGA is to detect gaps, it is more important to close them. Excessive

emphasis on dead-end gaps (those that cannot be filled, e.g. lost data in time) serve little purpose other than as a warning. Communicating the results of the DGA is largely an investment in the future. Solutions, or avenues for solutions, should be presented alongside found gaps.

How to ensure that communication reaches the target audience

As a part of the DGA, the target audience should have been identified and possibly already contacted for collaboration. The same channels can be used to communicate the completion of the DGA. However, additional audiences could be identified as a result of the DGA results, prompting a wider dissemination. Current indexing facilities will certainly capture a DGA that is posted online, but announcements by mail, distribution lists and the like should ensure ample distribution. Although indexing (and location of the DGA) will nearly always be possible, if given the choice, open access model should be highly preferable to any paywall option.

Frequency and mechanisms for communication

It is assumed that diffusion will be electronic rather than on paper, or that minimal paper copies would exist for libraries or archival purposes, but the facility of print-on-demand should be considered, as well as licensing through a CC model.

DGAs are cyclic in nature, as once a new state of knowledge has been reached, new goals can be set. Thus, a stable site where successive DGAs can be posted is desirable, as opposed to separate or ad-hoc publishing ventures.

Dynamic (online) vs. static publication

Often the DGA will be a complex exercise that may require significant processing, thus favoring static products (e.g. reports). However, actionable reports containing sections that can be automatically updated may be more accurate and timely. See for example the analytical tools available in GBIF portal. If the DGA is published online through a portal, the report may be accompanied by a web page where automatic updates could be consulted for those analyses that may allow it.

Prioritization of gap-closing, demand-driven data discovery and publishing activities

It is natural that activities described earlier will help identify several gaps in currently accessible data. However, not all of them can be bridged at the same time. Or in other words, aspirations and wish list of user communities cannot be met at the same time. This may result from lack of resources for various data life cycle activities, or simply because data do not exist and need to be freshly collected through monitoring and survey activities. This is the case, for example, in large tracts of Central Africa or tropical Asia, among others (Meyer, Weigelt, *et al.* 2015) where taxonomic uncertainty is rather high.

On the other hand, the landscape of available data is in a state of flux, as many stakeholders may be changing their policies about releasing already-existing data, which may result in rapid shifts in gaps and therefore priorities (Meyer, Kreft, *et al.* 2015). Therefore, prioritization of demands for data by the stakeholder community is essential. Criteria for such prioritization differ depending upon gravity of user demands, types of data requirements (quantity and quality), geographic, ecosystem, and thematic scope of the demands for the data, etc. (Chavan *et al.* 2010).

In setting which community demands need priority, a balance between focus on the largest gaps and a focus on areas and species of conservation concern, opportunities for closing gaps in long time series, economic considerations like return-on-investment, etc. would ultimately create a richer information basis than one that merely follows the immediate needs expressed by users, and that may not always be coincidental. However, it should be expected that in a sufficiently representative sample of stakeholders, this balance is likely to emerge as well.

Evaluation of the DGA exercise

Once the DGA has been done under best practice, it's time to evaluate how it fared. The aims of evaluation are:

- to establish the reliability of the exercise, and
- to inform subsequent DGAs based on lessons learned.

The main evaluation criterion is whether the DGA met the expectations. Two basic outcomes are possible from a DGA: either expectations were met or not. Both outcomes are relevant and may need to be scrutinized:

- Fulfilled expectations need to account for spurious results, especially obvious gaps (e.g. those coming from naturally unobtainable data spaces such as time-specific baselines from unsampled regions). Expectation fulfillment needs to be verified for:
    - Scoping
    - Purpose
    - Consistency
- When expectations are not met, a "post-mortem" analysis may be helpful. In addition to the checks above, it could also be useful to ask whether:
    - A cost-benefits analysis was properly done, to weed out unrealistic expectations,
    - Time and resources were adequately calculated or there were insurmountable overruns preventing proper DGA,
    - Other similar DGA encountered similar problems.

Every new DGA may add insights to the DGA process, and a comparative analysis of DGA exercises may be insightful and helpful for planning new DGAs. Section 4 below analyzes some exemplary DGAs done over the past two decades.

Section 3: Strategies and Planning for future DGA

As mentioned earlier, Data Gap Analysis is a continual process that needs to be carried out at regular intervals. This means processes, methods and approaches need not be freshly reinvented every time. Rather, any subsequent data gap analysis exercises should be built upon experiences gained during earlier exercises. Further, the results of the previous exercises should act as a baseline or benchmark for future data gap analysis studies. This calls for closer evaluation of every data gap analysis to understand the positive or negative aspects, what was missing, and rectifications in the next exercise. This will help building strategies and approaches making data gap analysis a productive exercise, leading to demand-driven and deterministic data discovery and publishing initiatives (Chavan *et al.* 2010).

Section 4: Lessons learned from case studies

Many instances of conservation gap analyses exist in literature (e.g. Crist & Csuti 2000; Dudley & Parrish 2006; Langhammer *et al.* 2007; NSW National Parks and Wildlife Service 2001; Brooks 2004; Peterson & Kluza 2003; Peterson 2005; Jarvis *et al.* 2011) but the

number of biodiversity data gap analyses is still relatively scant. **Error! Reference source not found.** identifies some of these, and we further discuss a selection primarily dealing with GBIF-enabled data as exemplary cases.

Of the sixteen cases we analyze in **Error! Reference source not found.**, most relied on some form of database analysis, requiring already existing datasets to be made available for the DGA. Other methods were less often used or used secondarily. Among those, literature surveys, geographical analyses and visualizations were more frequent.

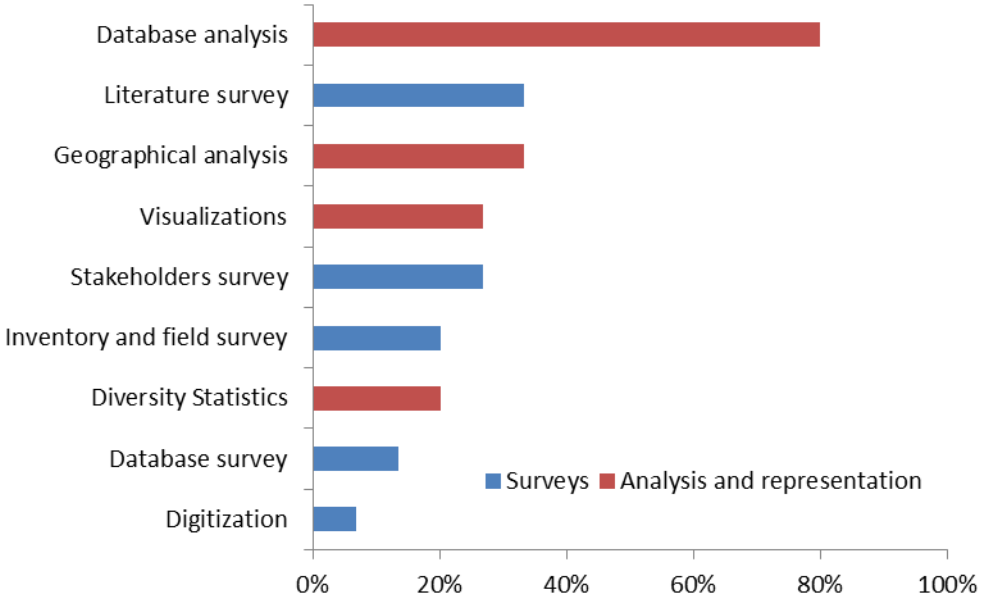

Figure 8. General techniques used in the 16 examined DGA exercises.

GBIF Secretariat – State-of-the-Network (2010)

In 2010, GBIF prepared a report to present a comprehensive overview of (a) the data discovery and publishing potential, (b) the status of data publishing, and (c) a content assessment of accessible data, in order to assess the preparedness of the GBIF network to meet its ambitious targets. Part of this report was a gap analysis targeted at taxonomic, geographic and temporal gaps in currently accessible data published through the GBIF network. Four types of assessment, namely, (a) taxonomic, (b) geographic, (c) temporal, and (d) basis of records were carried out (Chavan, Gaiji, *et al.* 2010).

Purpose (as DGA)

To assess content of data mobilized by and accessible through the GBIF network, and to determine how records were distributed in space and time, how participants contributed them and whether there were clear taxonomic biases.

Method

As part of a participating survey conducted regularly, estimates were requested from participants about their potential data holdings and mapped against contributed holdings. Reports produced by participants were tabulated and statistically analyzed to determine their data distribution. Simultaneously, the GBIF index was queried to obtain summary statistics across the targeted concepts for the DGA.

Results

The universe of available data was assessed from the survey results, and a gap was found in digitally accessible data vs. existing data. The database analysis found gaps in the spatial coverage of the data, with large bias towards the Northern hemisphere; in the taxonomical coverage, with low representation of most invertebrates and non-flowering plants; in the availability data older than a decade, including sizable data without date; and in the information about the type of data.

Unmet expectations and reasons

Data from many Participant Nodes were not contributed during the survey. Some of these nodes represented potentially large amounts of data and it is likely that their lack biased the assessment. Reasons for this failure to contribute data are multiple, but in general progress in data discovery and publishing is directly proportional to the status, mandate, capacity, vision and resources of the Participant Biodiversity Information Facilities.

Lessons learned

The DGA exercise in the 2010 State of the Network report pointed to the highly biased distribution of digitally accessible data and the need to carry out frequent monitoring on a local-to-global scale. Further, it highlighted the importance of active involvement of all network participants for a successful DGA (Chavan, Gaiji, *et al.* 2010).

University of Navarra analyses of GBIF.ES (2009, 2013)

In 2009, researchers at the University of Navarra analyzed the data hosted and served through the Spanish node of GBIF (GBIF.ES). As the node was the main hub for publishing standardized primary biodiversity data by Spanish researchers and institutions, it could be regarded as a fair representation of the digitally accessible knowledge of such data in the country (Ariño & Otegui 2009). The DGA was expanded and compared two years later using more data, also served through GBIF.ES, and more detection techniques (Otegui, *et al.* 2013b).

Purpose (as DGA)

The intent was to detect strengths and weaknesses in the availability of data, trying to find whether there were factors affecting the quality of the data. The included gap analysis examined the patterns looking specifically for coverage gaps in the temporal, spatial, and taxonomical dimensions, and proposed monitoring mechanisms for continuous evaluation of the data.

Method

The team applied visualization and organization techniques they had been developing for pattern detection and gap identification to the datasets being published by GBIF.ES through database analysis and statistical methods. For the second DGA, tools were developed and made available (Otegui & Ariño 2012) that allowed for datasets outside GBIF.ES be analyzed as well, and newer completeness measures were also produced.

Results

The DGA revealed the extent of completeness in the data being published through GBIF.ES, and the large bias towards observations published during the past few years. Artifacts in the data were detected as arising from georeferencing mistakes. Spatial distribution was bimodal with two centroids (Iberian Peninsula and Americas), and precision gap was discovered: many records had been rounded to pre-set coordinates, yielding a regular grid of records with void spaces in between. Seasonal patterns with large voids in the winter months and holiday seasons (varying according to whether the publishers were academic or administrative institutions) were also present.

Unmet expectations and reasons

The initial DGA was inherently limited to the state of the techniques developed so far. The experiences of the first DGA were used to refine the analysis during the second DGA, producing more precise data and detecting a broader set of patterns.

Lessons learned

The fact that the exercise was repeated with a two-year interval showcased the need to evaluate the DGA and refine it accordingly, and the need to do repeated gap analyses as new data come in (patterns do change). The second evaluation yielded about twice the information about gaps as the preliminary one.

Content Assessment of GBIF (2012)

After a large growth of the data available over the previous assessment of 2010, the GBIF network was complementarily reassessed in 2012/2013 (Gaiji *et al.* 2013) both by the GBIF Secretariat and the University of Navarra, followed by significant changes in the technical infrastructure (Hadoop/MapReduce platform, taxonomical backbone reworking, etc.) at the GBIF Secretariat and further development of tools at UNAV. These circumstances allowed for more efficient analytics.

Purpose (as DGA)

To re-evaluate the state of the network while simultaneously addressing rising concerns about the perceived quality of the data being published through GBIF, or the completeness or coverage of the datasets, as recommended by the task group on Content Needs Assessment (Faith *et al.* 2013), and to assess the gaps and fitness for use of the GBIF-mobilized data.

Method

Several methods were used concurrently, all within the realm of database analysis and visualization. Hive[7] tables extracted from the full GBIF index were processed and stored as data mining summary tables by the GBIF Secretariat, allowing for future repeated analyses. UNAV worked on a random sample of the full index and over several consecutive states of the indexes, enabling the study of the evolution of the gaps through time.

Results

A large series of assessments was obtained from the analyses, from data quality to evolution of gaps and how they were being closed (or growing). The updated taxonomical backbone prompted by the 2010 assessment enabled a more accurate assessment of taxonomical completeness, which was much higher and suggested that the gap was likely being closed. Geospatial checking of raw data had also improved vastly and the overall data quality had clearly improved (e.g. by tagging suspicious coordinates), and the georeferencing gap was also closing although only in recent data. However, a large gap was discovered in the temporal dimension, with many invalid or suspicious dates or no dates, affecting to more than one third of the records (Otegui, *et al.* 2013a) and up to 50% in the combination of temporal and geospatial information. Other gaps were not found to be closing. For example, the large dominance of observations of birds just increased over time. Other results patterned species richness obtainability and the rate of accrual (widening the north-south gap), or the fitness-for-use of data within pre-set, commonly used spatial ranges.

[7] http://hive.apache.org

Unmet expectations and reasons

The methodology allowed fast mining of the GBIF data index but was unable to address very specific issues such as the level of accuracy of the geospatial data, misidentification of taxa, or duplicated records between datasets. All these issues require dedicated, specifically targeted developments that may not exist yet or are still prototypes. For example, the accuracy of geospatial data could in principle be derived from workflows such as BioGeoBIF (Hill *et al.* 2009) and its Biogeomancer component (Guralnick *et al.* 2006) but need to be widely adopted; while taxonomical accuracy is highly dependent on the taxonomies followed by the publishers, and to solve duplicated records a definitively adopted standard for globally-unique identifiers is required.

Lessons learned

This study was probably the largest data gap assessment performed to date both in breadth and depth of the main index of biodiversity data. While the availability of improved infrastructures greatly facilitated the analytics, it also became evident that each result merited deeper inquiry, as gaps kept appearing as side effects whenever an analysis was performed to assess an issue. A main lesson thus was that the more comprehensive the analysis, the more likely gaps will be found that had not been suspected, often appearing only when enough data had been put to play (Ariño & Otegui 2008). One corollary was the need to develop a set of monitoring tools that should be running continuously as the index grows, enabling early warnings to prompt for correction.

GBIF Position Paper on fitness-for-use (2010)

The infrastructure program of GBIF developed an automated filter to distinguish what information appears to be correct, and what is not. With data from 2009, a component in a white paper on fitness-for-use published in 2010 (Hill *et al.* 2010) looked at what types of geospatial issues were detected by GBIF's automated filter while indexing raw data published by contributors, and how these issues would have affected data quality, while classifying the detectability and solvability of these issues.

Purpose (as DGA)

To quantify the level of error in data that could be detected or suspected by automated filters, to get an idea about what errors were most frequent, and to assess the difficulty of solving or annotating these errors and close the corresponding gaps, enhancing fitness-for-use.

Method

Statistical analysis of the annotations made by the GBIF filter that sits between the harvester and the indexer of the data contributed by publishers, and comparison of the raw and filtered datasets to analyze and categorize the information that was blocked by the filters. Results were tabulated and sorted according to various criteria.

Results

Eleven types of geospatial issues were found, of which six were recurrent: incomplete coordinates, strings in numerical fields, wrong coordinate systems, numerical sign confusion, and very particularly coordinate reversals and out-of-country coordinates. In itself this should not have constituted a gap, and some of them could have been easily solved (e.g. detected coordinate swap). However, further analyses showed that two-thirds of the issues (amounting to nearly three million records) were detected in just two specific data resources. As many data publishers tend to concentrate on specific regions or taxa, the loss of the geospatial information may not have been uniform but concentrated—thence, gaps could exist.

Unmet expectations and reasons

Comparing the two procedures led to the observation that some issues would not have been detectable had not both set of analyses taken place. In particular, as the filter works at the individual record level and the comparison of datasets works at the dataset level, some issues could only be detected when statistically comparing large amounts of data. Suspected patterns visible during the comparison, e.g. the effect of political boundaries and country's economy in the availability of data, lack detectability in a one-to-one filter.

Lessons learned

Almost all errors in the exercise arose at the time of digitization or in the distributed network, and many were resolvable. Nevertheless, the information that may pass filters, however well devised, may still contain errors that only appear when analyzing the data in the context of other data. Detected gaps may thus result from digitizing errors, and not only from actual lack of data. Unfortunately, the parsing of a record by a filter may lead to a potentially correct value in its geospatial information, but with these methods it is not possible to undoubtedly state that the processed record is correct.

EU BON Gap Analysis (2014)

The EU BON (Building the European Biodiversity Observation Network) project coordinated by the Museum für Naturkunde – Leibniz Institute for Evolution and Biodiversity Science, Germany, conducted a detailed gap analysis that was published as a project's deliverable at the end of 2014 (Wetzel *et al.* 2014). The analysis was required to set a baseline of what was available from European biodiversity databases at a high level, but extended to global databases from the European scope. Gaps were found and reviewed in a systematic manner, and produced a comprehensive set of recommendations for plugging the existing biodiversity data gaps.

Purpose (as DGA)

To assess the relevant data sources on biodiversity on a European and global scale with the aim to allow identifying and prioritizing actions to improve data availability as required by stakeholders.

Method

The requirements of different stakeholders were evaluated through a survey that included a predefined set of categories summarizing high-level questions for biodiversity proposed mainly to the EU-BON participants. Information on existing databases was compiled, and a set of potential gap categories was systematically tested on a selection of the datasets through database analysis: spatial, temporal, and taxonomical gaps, accessibility issues, trends, and data quality factors. The datasets were also evaluated against the Essential Biodiversity Variables (EBVs) covering most themes related to biodiversity.

Results

The results of the stakeholders survey allowed establishing a ranking of 29 biodiversity questions, separately for intrinsic importance and for effect of gaps. Invasive species and biodiversity and ecosystems and their services seemed to be the more relevant areas, although data availability was generally perceived as "fair" but not good or very good. In ecosystem functions a general lack was felt.

Indeed, after the gap analysis the EBV most frequently covered by databases was species populations, while other variables, in particular in community composition and ecosystem functions, were found to have gaps. Also, significant gaps were found in taxonomic coverage at the European level.

Spatial gaps were especially significant for Eastern Europe. Taxonomical gaps were important in particular for target species (e.g. pollinators) in specific areas, highlighting the importance of conducting a multi-dimensional analysis.

Unmet expectations and reasons

The survey found large gaps in the accessibility to the data, highlighting the importance of dark data (Heidorn 2008) needing to be released: while not necessarily buried knowledge (BK), two-thirds of the surveyed datasets had some form of access restriction potentially resulting in locked knowledge (LK). This could potentially result in both a general availability gap, but also in pseudo-gaps: for instance, most users cannot access data about a certain group and perceive it as a gap because of LK, while in fact some users can and therefore the gap does not strictly exist. As a result, not all gaps could be assessed with equal confidence, and some could in fact be pseudo-gaps.

In addition, the DGA encountered some difficulties in matching stakeholders' perception and expectations to actual data availability. While there were many reasons for this that were explored in detail, some questions about biodiversity were found wanting because of difficulties at data generation (e.g. genetic diversity or ecosystem services assessments). The lack of data may have in turn prompted low relevance (and thence low priority) from stakeholders in the survey.

Lessons learned

This extensive DGA exercise highlighted how DGA can be conducted across several scales and dimensions, and importantly how gaps can potentially be plugged by releasing LK. On the other hand, the existence of BK and difficulties or costs at obtaining certain data types may form a vicious circle, whereby stakeholders may be attaching low priority to otherwise important questions (thus reducing the possibility of plugging the gap) because of that very difficulty at getting the data.

Global Vertebrate Records in GBIF (2015)

Through the analysis of more than 157 validated point occurrences of more than 21,000 vertebrate species facilitated through GBIF, Meyer, Kreft, *et al.* (2015) were able to describe a global picture of digital accessible information (DAI) for that group. Extensive multifactor analysis provided multiple potential drivers for the observed patterns.

Purpose (as DGA)

To determine what gaps existed globally for that group and what were the main factors currently limiting biodiversity inventory completeness in global digital accessible information (DAI), as well as identify priority regions and activities to advance it.

Method

Data accessed at the GBIF index was subject to a workflow where the taxonomical backbone was extensively standardized and records cleaned of inconsistencies. Species distribution data was checked against expert sources and aggregated and mapped at four different resolutions. A number of statistical and geostatistical analyses was run on the data to derive inventory completeness for each cell at all resolutions, broken down taxonomically. Completeness results were crossed against a number of socioeconomic factors and other proxies to try to find the main drivers responsible for the observed gap patterns.

Results

Outside a few well-sampled regions, DAI on point occurrences provides very limited and spatially biased inventories of species. Many large, emerging economies are even more under-represented in global DAI than species-rich, developing countries in the tropics. Multi-model inference reveals that completeness is mainly limited by distance to researchers, locally available research funding and participation in data-sharing networks, rather than

transportation infrastructure, or size and funding of Western data contributors as often assumed.

Unmet expectations and reasons

The study was based on GBIF-facilitated data as the best representation of DAI, producing a map of gaps—but also acknowledged that other sources of DAI may be available that might change the picture significantly. It was noted that recent trends towards sharing of previously locked information (termed LK in this paper) represent a very rapid shift towards releasing large quantities of information, specifically in some areas that currently represent gaps.

Lessons learned

Results highlight potential ways for making institution-based data mobilization more effective, but also the limitations of such efforts. Also, results expose the urgent need for integrating non-Western data sources and intensifying cooperation to more effectively address societal biodiversity information needs.

Exercise	DGA Purpose	Methods	Results		
			Achieved	**Failed - *reason***	**Main lesson**
Requirements for biogeographical analysis of birds in Mexico (Peterson *et al.* 2008)	To analyze the spatial distribution of Mexican birds in collections	Digitization, database analysis, survey of literature, richness analysis	Sampling of Mexican birds is incomplete, most specimens are concentrated in a few sites, some populations are not represented	Many quadrats were inadequately sampled in the datasets to ascertain distribution *Documented sampling bias for most species, even endemics or well-represented*	Patterns of sampling impact the representativeness of the species: gap filling require continued sampling
Gap analysis for prioritization of conservation (Koleff *et al.* 2009)	To analyze patterns and gaps in the coverage of protected areas as respects to biodiversity in Mexico	Survey, database analysis, literature review, map analysis	Indexes evaluating ecoregions were generated from the aggregation of data, finding geography-linked gaps such as lowlands and certain types of ecosystems	The gap analysis did not have enough resolution to distinguish the state and viability of populations *The study was country-wide and the chosen level of resolution was adequate to the primary objective, although not for all possible objectives*	Gap analysis may yield different results depending on the chosen level of spatial resolution. However, atomized data will always be able to be grouped for broader resolution so they are desirable.
GBIF Secretariat – State-of-the-Network (Chavan, Gaiji, *et al.* 2010)	To assess content volume of data mobilized by the GBIF network	Survey, database analysis	Estimation of the universe of data; extent of geospatial, temporal and taxonomical coverage	Some results likely biased *Data from several large participant nodes were not contributed: status, mandate, capacity, vision and resources of BIFs differ widely.*	Surveys need to be unbiased and representative.
GBIF's Position Paper on Fitness-For-Use (Hill *et al.* 2010)	To assess how issues in data could be identified	Statistics of parsing filter's results, comparison of pre- and post-processed data	Geospatial issues classified and ranked, with most gaps attributed to just a few publishers	Some issues were likely undetectable *Single-technique assessments can be insufficient to detect gaps*	Errors leading to gaps may prove to be undetectable unless combinations of techniques are used

Exercise	DGA Purpose	Methods	Results		
			Achieved	**Failed - *reason***	**Main lesson**
Assessment of ecosystem threats in South America in IABIN (Jarvis *et al.* 2010)	To use primary biodiversity data for threat assessment in South America	Database analysis	Significant amounts of data lack reliable coordinates although they could likely be georeferenced through a web service	Some of the databases could not be properly assessed *Coordinates were not interpretable or the databases were online unreliably*	Algorithms for automated georeferencing can help closing gaps once a number of issues are solved.
UNAV assessments of the GBIF Spanish node (Ariño & Otegui 2009; Otegui *et al.* 2013)	To characterize data provided through GBIF.ES	Database analysis, visualization tools	Discovery of patterns and artifacts leading to gaps in space and time	Initial DGA limited *Techniques still experimental in the first phase*	DGAs need to be repeated over time as data change and detection techniques improve
Joint GBIF-S/UNAV assessment of global data (Gaiji *et al.* 2013)	Follow-up and improvement of the assessment of the state of the network	Database analysis, visualization tools	Discovery of gaps across many dimensions and scales and clues to their technical and procedural causes	Some gaps could not be assessed despite new technologies *Technologies required to assess certain data quality issues do not exist or have not been widely adopted yet, i.e. GUIds*	A set of monitoring tools needs to be in place for early warnings. Gaps will be discovered as data increase and technology improves.
Data gaps in biosphere reserves in Mexico (Pino-Del-Carpio *et al.* 2013)	To estimate the amount of data gaps between sources of information in biosphere reserves	Literature survey, Database analysis, mapping	The use of single information sources misses large amounts of critical distribution information even in highly studied areas or sensitive species	Some estimates were not as precise as desired *Georeferencing rounding and lack of uncertainty estimations at digitization prevented exact placement of some data*	Omitting any available source is very likely to jeopardize knowledge even in highly studied areas. Complementary sources can close gaps.

Exercise	DGA Purpose	Methods	Results		
			Achieved	**Failed - *reason***	**Main lesson**
Assessment of completeness of digitally accessible knowledge for plants in Brazil (Mariane Silveira Sousa-Baena *et al.* 2013)	To assess documentation of Brazilian plants looking for gaps in current knowledge	Inventory assessment, spatial analysis of databases (speciesLink)	Spatial knowledge of Brazilian angiosperms is very unevenly distributed, as well-known sites are spatially concentrated. Areas lacking detailed botanical documentation coincide with much habitat destruction.	For the calculations of gaps and completeness, only data from sites considered well sampled were retained *When few records were available, random effects jeopardized the estimates of completeness*	Biodiversity surveys and inventory efforts can be guided by existing knowledge that is digitally accessible. Spatial summaries of completeness can guide sampling efforts.
EU-BON Gap analysis (Wetzel *et al.* 2014)	To assess the relevant data sources on biodiversity on a European and global scale.	Survey, database analysis, map analysis, visualization tools	Established a ranking of biodiversity questions to be addressed through DGA. For relevant issues, gaps are narrow for IAS but wide for ecosystem functions and services.	Not all gaps could be assessed with equal confidence. Survey results and data availability need to feed in. *Some questions (e.g. genetic diversity, ecosystem services) can only be addressed using time consuming and technical methods: difficult access to data leads to low priority*	Sensible biodiversity-related priority setting at regional and global scale is achievable by combining specific data availability assessment and stakeholder surveys for perceived gaps in DGA.
Conservation of Crop Wild Relatives in Spain (Rubio Teso *et al.* 2015)	To assess protection status and requirements for CWR	Inventorying, database analysis	A prioritized list of CWR populations and their intersection with protected areas, and which remain at risk	Significance marred by high variance in results *Common species are paradoxically less studied and recorded, yielding poorer data*	New combined conservation units (e.g. CWR species occurrence and ecogeography) may be key to pinpoint gaps relevant to genetic conservation.

Exercise	DGA Purpose	Methods	Results		
			Achieved	**Failed - *reason***	**Main lesson**
Global vertebrates in GBIF dataset (Meyer, Kreft, *et al.* 2015)	To assess gaps in digital accessible information (DAI) about species worldwide	Database analysis, visualization tools, multivariate statistics	Identification of gaps and drivers for them for species occurrences. Most DAI extremely biased.	Maps of gaps cannot be considered complete *Other sources of DAI exist that may influence results*	Targeted integration of available information and assessment of gaps is vital. Cultural shift towards sharing required.

***Table 1**. A list of selected data gap analysis exercises (in chronological order) with purpose, results, and lessons learned.*

Exercise	DGA Purpose	Methods	Results		
			Achieved	**Failed - *reason***	**Main lesson**
EUNIS assessment of data sources (Condé *et al.* 1995)	To assess data sources for European biodiversity	Survey, inventory data	Although presence data existed, data on populations were rare. A gap therefore existed in the quantification of data. Invertebrate data were largely absent.	Gap analysis was done at a very general level. *At the time of collection, databases were scattered and disconnected. Exchange standards did not exist.*	Data coming through surveys would more accurately represent gaps if local focal points were put to action.
CONABIO's meta-analysis of national biological surveys (Soberón *et al.* 1996)	To assess what information can be retrieved from foreign museum data about Mexican biodiversity	Survey of databases, survey of literature	Several examples of distribution data biased by unavailability of data or sampling problems, e.g. "collector syndrome"	Lack of access to much of the data stored without common standards *At the time of the study, data exchange standards for biodiversity had not yet been developed*	Data exchange standards and access to at least metadata was a necessary step in broader-than-national gap assessment
Visualization proposals for biodiversity assessment (Guralnick *et al.* 2007)	To showcase a survey gap analysis as a validating step of performing a species richness analysis	Database crawling, visualization tools, richness estimations	Inequities in available information for birds and mammals based on GBIF records in different regions result in estimates not converging in many regions	Estimates of richness could not progress further *Some data that would enrich the biodiversity estimates (e.g. abundance, life phase, etc.) were not available*	Smart web-enabled tools may help collate data and accelerate identifying strengths and weaknesses in global biodiversity data
Tropical biodiversity data gap (Collen *et al.* 2008)	To examine coverage of biodiversity data using four global datasets and assess discrepancies	Literature survey, database analysis, map analysis	Patterns converge to indicate fewer and scattered data in the tropics and large gaps in the Southern hemisphere	Taxonomical coverage not explicitly evaluated *Taxonomic biases may influence the ability to assess changes in biodiversity, effectiveness compromised*	The difficulty in field-filling the tropical biodiversity data gap may call for other types of data (e.g. satellite data) to be evaluated as proxies for biodiversity

Acknowledgements

We appreciate constructive reviews by Franziska Schrodt and Carsten Meyer, which helped improve this version of the guide. The completion of this guide was coordinated in GBIF Secretariat by Dmitry Schigel, with contributions from Siro Masinde, Tim Hirsch, and Sampreethi Aipanjiguly.

References

Ariño AH (2010) Approaches to estimating the universe of natural history collections data. *Biodiversity Informatics* 7(2): 81-92. Available at: https://journals.ku.edu/index.php/jbi/article/view/3991.

Ariño AH, Chavan V & Faith DP (2013) Assessment of user needs of primary biodiversity data: analysis, concerns, and challenges. *Biodiversity Informatics* 8(1): 59-93. Available at: https://journals.ku.edu/index.php/jbi/article/download/4094/4199.

Ariño AH, Chavan V & Macklin JA (2013) *GBIF best practice guide for content needs assessment of stakeholder communities*. Copenhagen: Global Biodiversity Information Facility. Available at: http://www.gbif.org/resources/3024.

Ariño AH & Otegui J (2009) Meta-análisis de los datos de biodiversidad suministrados a través de gbif.es, Pamplona. Available at: http://www.gbif.es/ficheros/metagbifes.pdf.

Ariño AH & Otegui J (2008) Sampling biodiversity sampling. In Weitzman AL & Belbin L, eds., *Proceedings of TDWG. Biodiversity Information Standards (TDWG)*, 107. Available at: http://www.researchgate.net/publication/258243679_Sampling_Biodiversity_Sampling.

Berendsohn WG & Seltmann P (2010) Using geographical and taxonomic metadata to set priorities in specimen digitization. *Biodiversity Informatics* 7: 120-129. Available at: https://journals.ku.edu/index.php/jbi/article/view/3988/3808.

Brooks TM et al. (2004) Coverage provided by the global protected-area system: is it enough? Bioscience, 54(12): 1081-1091. Available at: http://bioscience.oxfordjournals.org/content/54/12/1081.full.pdf.

Chavan V et al. (2010) *State-of-the-Network 2010: Discovery and Publishing of Primary Biodiversity Data through the GBIF Network*. Copenhagen: Global Biodiversity Information Facility, 36 pp. Available at: http://www.gbif.org/resource/80666.

Chavan V & Penev L (2011) The data paper: a mechanism to incentivize data publishing in biodiversity science. *BMC Bioinformatics* 12(Suppl 15): S2. Available at: http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-12-S15-S2.

Chavan V, Sood RK & Ariño AH (2010) *GBIF best practice guide for Data Discovery and Publishing Strategy and Action Plans*. Version 1.0. Copenhagen: Global Biodiversity Information Facility. Available at: http://www.gbif.org/resources/2614.

Collen B et al. (2008) The tropical biodiversity data gap: addressing disparity in global monitoring. *Tropical Conservation Science* 1(2): 75-88. Available at: http://tropicalconservationscience.mongabay.com/content/v1/08-06-09-Ben_Collen_et_al.html.

Condé S et al. (1995) Databases on species, habitats and sites: survey and analysis 1995-96. Copenhagen: European Topic Centre on Nature Conservation. Available at: http://www.eea.europa.eu/publications/92-9167-034-0/at_download/file.

Crist P & Csuti B (2000) Gap Analysis. In *A handbook for conducting gap analysis*. Moscow, ID, USA: USGS Gap Analysis Program, 151-157.

Dudley N & Parrish J (2006) *Closing the gap. Creating ecologically representative protected area systems: a guide to conducting the gap assessments of protected area systems for the Convention on Biological Diversity*. CBD Technical Series, 24. Montreal: Secretariat of the Convention on Biological Diversity. Available at: https://www.cbd.int/doc/publications/cbd-ts-24.pdf

Faith DP et al. (2013) Bridging biodiversity data gaps: Recommendations to meet users' data needs. *Biodiversity Informatics* 8(1): 41-58. Available at: https://journals.ku.edu/index.php/jbi/article/view/4126.

Fry C et al. (2002) Analysis of baseline data requirements for the SEA directive - final report. TLR Limited. Available at: http://www.southwest-ra.gov.uk/media/SWRA/Environment/Analysis_Baseline_Data_Requirements.pdf

Gaiji S et al. (2013) Content assessment of the primary biodiversity data published through GBIF network: status, challenges and potential. *Biodiversity Informatics* 8: 94-172. Available at: https://journals.ku.edu/index.php/jbi/article/view/4124/4201.

Graham M & Kennedy J (2014) Vesper: visualising species archives. *Ecological Informatics* 24: 132-147. Available at: http://linkinghub.elsevier.com/retrieve/pii/S1574954114001113.

Guralnick RP et al. (2006) BioGeomancer: automated georeferencing to map the world's biodiversity data. *PLoS Biology* 4(11): e381. Available at: http://dx.doi.org/10.1371/journal.pbio.0040381.

Guralnick RP, Hill A & Lane M (2007) Towards a collaborative, global infrastructure for biodiversity assessment. *Ecology Letters* 10(8): 663-72. Available at: https://dx.doi.org/10.1111%2Fj.1461-0248.2007.01063.x.

Heidorn PB (2008) Shedding light on the dark data in the long tail of science. *Library Trends* 57(2): 280-299. Available at: http://muse.jhu.edu/journals/library_trends/v057/57.2.heidorn.html.

Hill AW et al. (2010) GBIF position paper on future directions and recommendations for enhancing fitness-for-use across the GBIF Network. Copenhagen: Global Biodiversity Information Facility. Available at: http://www.gbif.org/resource/80623.

Hill AW et al. (2009) Location, location, location: utilizing pipelines and services to more effectively georeference the world's biodiversity data. *BMC Bioinformatics* 10(1): S3. Available at: http://www.biomedcentral.com/1471-2105/10/S14/S3.

Hjarding A, Tolley KA & Burgess ND (2014) Red List assessments of East African chameleons: A case study of why we need experts. *Oryx* 49(4): 652-658. Available at: http://www.journals.cambridge.org/abstract_S0030605313001427.

Idohou R et al. (2015) Diversity of wild palms (*Arecaceae*) in the Republic of Benin: Finding gaps in the national inventory combining field and digital accessible knowledge. *Biodiversity Informatics* 10: 45-55. Available at: https://journals.ku.edu/index.php/jbi/article/view/4914/4493

Jarvis A et al. (2011) An integrated adaptation and mitigation framework for developing agricultural research: synergies and trade-offs. *Experimental Agriculture* 47(2): 185-203. Available at: http://www.journals.cambridge.org/abstract_S0014479711000123.

Jarvis A et al. (2010) Providing means for a better understanding of biodiversity: improving primary data and using it for threat assessment and in situ conservation planning in South America. Progress report no. 1. Cali: CIAT. Available at http://www.oas.org/dsd/IABIN/Component3/CIAT/Informe%20de%20Avance.pdf

Koleff P et al. (2009) Identificación de prioridades y análisis de vacíos y omisiones en la conservación de la biodiversidad de México. In *Capital natural de México, vol. II: Estado de conservación y tendencias de cambio*. México City: CONABIO, pp. 651-718.

Langhammer PF et al. (2007) *Identification and gap analysis of Key Biodiversity Areas*. Gland, Switzerland: IUCN.

Meyer C, Kreft H et al. (2015) Global priorities for an effective information basis of biodiversity distributions. *Nature Communications* 6: 8221. Available at: http://www.nature.com/doifinder/10.1038/ncomms9221.

Meyer C, Weigelt P & Kreft H(2015) Multidimensional biases, gaps and uncertainties in global plant occurrence information. *PeerJ PrePrints* https://dx.doi.org/10.7287/peerj.preprints.1326v2h.

Mulder NJ (1988) Digital image processing, computer-aided classification and mapping. In Küchler AW & Zonneveld IS, eds. *Vegetation Mapping*. Dordrecht, Netherlands: Kluwer Academic Publishers, 269-316.

NSW National Parks and Wildlife Service, 2001. Western data audit and gap analysis: Western region, Resource and Conservation Assessment Council.

Otegui J et al. (2013a) Assessing the primary data hosted by the Spanish node of the Global Biodiversity Information Facility (GBIF). *PLoS ONE* 8(1): e55144. Available at: http://dx.plos.org/10.1371/journal.pone.0055144.

Otegui J et al. (2013b) On the dates of the GBIF-mobilised primary biodiversity data records. *Biodiversity Informatics* 8(1): 173-184. Available at: https://journals.ku.edu/index.php/jbi/article/download/4125/4202.

Otegui J & Ariño AH (2012) BIDDSAT: visualizing the content of biodiversity data publishers in the Global Biodiversity Information Facility network. *Bioinformatics* 28(16): 2207-2208. Available at: http://www.ncbi.nlm.nih.gov/pubmed/22730433.

Parrish J & Dudley N (2006) What does gap analysis mean? A simple framework for assessment. In Dudley N & Parish J, eds. *Closing the gap: Creating ecologically representative protected area systems*. Montreal: Secretariat of the Convention on Biological Diversity, 14-17.

Peterson AT et al. (2011) *Ecological niches and geographic distributions*. Princeton, NJ, USA: Princeton University Press. Available at: http://press.princeton.edu/titles/9641.html.

Peterson AT (2005) Kansas gap analysis: the importance of validating distributional models before using them. *The Southwestern Naturalist* 50(2): 230-236.

Peterson AT & Kluza DA (2003) New distributional modelling approaches for gap analysis. *Animal Conservation* 6(1): 47-54. Available at: http://doi.wiley.com/10.1017/S136794300300307X.

Peterson AT, Navarro-Sigüenza AG & Benítez-Díaz H (2008) The need for continued scientific collecting; a geographic analysis of Mexican bird specimens. *Ibis* 140(2): 288-294. Available at: http://doi.wiley.com/10.1111/j.1474-919X.1998.tb04391.x.

Peterson T (2013) Example of survey gap analysis: Kenya. In Biodiversity Informatics Training Program. Nairobi. Available at: http://biodiversity-informatics-training.org/wp-content/uploads/2013/09/D2T4_ATP_KenyaExample.pdf.

Pino-Del-Carpio A et al. (2011) Communication gaps in knowledge of freshwater fish biodiversity: implications for the management and conservation of Mexican biosphere reserves. *Journal of Fish Biology* 79(6): 1563-1591. Available at: http://doi.wiley.com/10.1111/j.1095-8649.2011.03073.x.

Pino-Del-Carpio A. et al. (2014) The biodiversity data knowledge gap: Assessing information loss in the management of Biosphere Reserves. *Biological Conservation* 173: 74-79. Available at: http://dx.doi.org/10.1016/j.biocon.2013.11.020.

Research Data Strategy Working Group (2008) Stewardship of research data in Canada : a gap analysis. Available at: http://publications.gc.ca/collections/collection_2009/cnrc-nrc/NR16-123-2008E.pdf.

Rocchini D et al. (2011) Accounting for uncertainty when mapping species distributions: The need for maps of ignorance. *Progress in Physical Geography* 35(2): 211-226. Available at: http://ppg.sagepub.com/cgi/doi/10.1177/0309133311399491.

Rubio Teso, ML et al. (2015) In situ conservation of CWR in Spain: present and future. *Crop Wild Relative* (10): 24-26. Available at: http://www.researchgate.net/publication/272504193_In_situ_conservation_of_CWR_in_Spain_present_and_future.

Scott JM (2000) A handbook for conducting gap analysis, Moscow, ID, USA: National Gap Analysis Program.

Scott JM et al. (1993) Gap Analysis - A geographic approach to protection of biological diversity. *Wildlife Monographs* 123(1): 1-41.

Soberón J, Llorente J & Benítez H (1996) An international view of national biological surveys. *Annals of the Missouri Botanical Garden* 83(4): 562-573.

Sousa-Baena MS, Couto Garcia L & Peterson AT (2013) Completeness of digital accessible knowledge of the plants of Brazil and priorities for survey and inventory. *Diversity and Distributions* 20(4): 369-381.

Sousa-Baena MS, Garcia LC & Peterson AT (2013) Knowledge behind conservation status decisions: Data basis for “Data Deficient” Brazilian plant species. *Biological Conservation*. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0006320713002255.

Wetzel F et al. (2014) EU BON Deliverable 1.1: Gap analysis and priorities for filling identified gaps in data coverage and quality, Berlin, Germany. Available at: http://www.gbif.fr/sites/default/files/documents/eu_bon_deliverable_1_1_final_v30.pdf.