# GBIF Best Practice Guide for Content Needs Assessment of Stakeholder Communities

Version 1.0

November 2013

**Suggested citation:**
Ariño AH, Chavan V, Macklin JA, Ghosh-Harihar M, Mathur V, Gaiji S, Flemons P (2013). GBIF Best Practice Guide for Content Needs Assessment of Stakeholder Communities, Ver. 1.0, Copenhagen: Global Biodiversity Information Facility, Pp. 62, ISBN: 87-92020-53-4, Accessible at http://www.gbif.org/resources/3024

**ISBN:** 87-92020-53-4
**Persistent URL:** http://www.gbif.org/resources/3024

**Language: English**



**Project Partners:** The Global Biodiversity Information Facility (GBIF); University of Navarra, Spain; Wildlife Institute of India, India; Agriculture and Agri-Food, Canada and Australian Museum, Australia

***Document Control:***

| Version | Description | Date of release | Author(s) |
|---|---|---|---|
| 1.0 | Content development | 22 November 2013 | Ariño AH, Chavan V, Macklin J, Ghosh-Harihar M, Mathur V, Gaiji S and Flemons P |

# About GBIF

**GBIF: The Global Biodiversity Information Facility**

As an inter-governmental organization, GBIF was conceived as a global mega-science initiative to address one of the great challenges of the 21st century – harnessing knowledge of the Earth's biological diversity. GBIF envisions a world in which biodiversity information is freely and universally available for science, society, and a sustainable future. GBIF's mission is to facilitate free and open access of the world's biodiversity information. To achieve this mission, GBIF helps a wide variety of biodiversity data holders, generators and users across the globe to make discoverable primary biodiversity data harmonized against agreed global standards. Website: http://www.gbif.org

# Table of Contents

## *List of Figures*

# Executive Summary

Biodiversity research is a data-intensive and collaborative science. In recent times a large cache of biodiversity data has become accessible for a wide range of uses and user communities. However, the data being accessible is only the first step, as users must then evaluate whether they are fit-for-use based on their requirements. Determining what the data are useful for, how they could be used, and by which stakeholders, are the fundamentals of content needs assessment activities. A user content needs assessment (CNA) is a systematic approach to studying the state of knowledge, ability, interest, or attitude of a defined audience or group.

User content needs assessments may be carried out at different scales, from local to global, and by a range of actors including governmental and non-governmental organizations, research project designers, funding institutions, national biodiversity information facilities and ministries responsible for biodiversity and environment. This guide is therefore intended for any individuals or institutions engaged in mobilizing data on biodiversity and making them freely accessible, serving the needs of a given set of stakeholders.

A systematic exercise of CNA helps identify and prioritize actions intended to meet the expectations of the user communities. Content needs assessment for biodiversity information is therefore both a tool to discover what stakeholders have requested, and a product — contents or results of the assessment itself.

However, content needs assessments in the biodiversity realm seem surprisingly rare. Further, conducting a content needs assessment can be a time consuming task where a set of scientific approaches together with expertise needs to be employed. Lack of a community-recognized guide makes it difficult to undertake such content assessment exercises. As a result, biodiversity data publishing is often an opportunistic activity, which is largely being determined by the researchers themselves.

This best practice guide provides step-by-step approaches for conducting a user needs assessment. It describes an optimal workflow consisting of seven steps, viz. (1) setting up of purpose, scope and objectives of the study, (2) identifying the target audiences, (3) selecting a set of optimal and easy to implement methods, (4) identifying information needed and investigation methodology, (5) collection and analysis of the information, (6) synthesis of collected information and publishing results, and (7) develop and implement an action plan with an appropriate monitoring and evaluation mechanism.

These steps are discussed with the help of eight representative biodiversity CNAs. While two of them have global coverage, six are national/regional in nature. These studies were examined for general trends. On the basis of this understanding, we provide a set of do's and don'ts for each of the seven steps of CNA workflow.

One aim of this guide is to help design optimal content assessment exercises ranging from global to national scale, with parsimonious investment of resources and time. As a result of these content assessment exercises, the biodiversity informatics community will be in a position to have a better understanding of its stakeholder communities. This will result in target-oriented and demand-driven data publishing policies and action plans.

## Section 1: User (content) needs assessment: Why?

Biodiversity research has become a data-intensive science (Kelling et al., 2009). The more than 400 million primary biodiversity records that hundreds of data publishers are making openly and freely available through GBIF at the time of writing are a significant asset of scientific capital (Borgman, 2003, 2007), but such a body of data, though of tremendous value, is expensive to produce, maintain, and share. The cyber-infrastructure required to maintain an open access and delivery platform to support data-intensive collaborative research (Borgman et al., 2007) can only be justified if these data are fit-for-use (Hill et al., 2010) and can satisfy user needs. Determining what these data are useful for, how they will be used, and what stakeholders (scientists, conservationists, policy makers, educators) would use the data for, are the fundamentals of content needs assessment (CNA) activities.

The objective of CNA is thus to provide an assessment of biodiversity data based on user requirements at a point in time. CNA should examine the extent and adequacy of biodiversity data and information currently being generated and made accessible from the point of view of the primary target audience: scientists and decision makers. It should also identify impediments to the use of such valuable information and suggest ways to streamline pathways to increase accessibility to stakeholders (Chavan et al., 2010).

Biodiversity data are typically generated for specific purposes, but these data are often used subsequently for other unintended purposes (Faith et al., 2013). Box 1 provides several good examples of uses for biodiversity data including the potential for new scientific research and decision making related to natural resource management (see Box1). However, deficiencies in data coverage and quality have also been highlighted (Gaiji et al., 2013; Otegui et al., 2013; Yesson et al., 2007; Ariño & Otegui, 2008). Although such deficiencies could in theory be addressed on a case-by-case basis, some type of prioritizing must be put in place to address the most obvious gaps (e.g. temporal, geographical and taxonomic).

Attempts to reduce such gaps will inevitably consume resources, and these should be allocated judiciously to optimize results, i.e., using objective criteria (Ariño et al., 2011). But users' requirements also need to be evaluated in order to assess which gaps should be given priority.

However, user requirements are very diverse and they tend to evolve over time based on things like policy decisions (e.g. Aichi Targets for 2020). In 1910, Joseph Grinnell predicted that biodiversity data stored in museums as vouchered specimens and observations would one day allow us to see how the environment would change over time (Grinnell, 1910). His prediction came true one century later, when a new survey in the Sierra Nevada range in California was able to compare old and new biodiversity data to assess distribution changes (Fellers & Drost, 1996). When assessing the types, amounts, or characteristics of data needed by stakeholders, it would be insightful to foresee potential future uses for the data, as well as current ones. But this could result in diverting scarce resources to capture data not readily useable, in the hopes of having them available in the future for other purposes. CNA should become an essential tool to support optimal resource allocation by helping balance current needs with potential future needs. Assessing future needs should in turn be based on current usage trends with some room for reasonable predictions, while avoiding excessive earmarking for uncertain futures.

## Section 2: What is a User Content Needs Assessment?

A user content needs assessment is a systematic approach to studying the state of knowledge, ability, interest, or attitude of a defined audience or group. It differs from a simple needs assessment in that it does not limit itself to evaluating the gap between the present state (what users have or are using) and the desired state (what users would like to have or use). However, the purpose of a CNA is ultimately to focus on the ends (where we want to go) rather than the means (how we are to reach that end) (Witkin & Altschuld, 1995). In doing so, CNA helps setting priorities and determine criteria for solutions that can help reach that end, and leads to action that will improve the overall reliability of the assessment.

CNA for biodiversity information is therefore both a tool to discover what stakeholders need, and a product — the contents or results of the assessment itself. The stakeholders' needs should contain their current and future requirements, including data, metadata, data products, computing power, or processes to support their activities.

A successful user CNA exercise should therefore include:

- an overview of current usage and understanding of biodiversity information by users,
- a set of forecasts of future usage based on current trends, and
- recommendations on future priorities or actions to improve data usage or availability.

The overview of current usage should be informative; provide quantification where possible, and include enough data and metadata to allow statistical support.

The forecast should provide measurable indicators of success, as the accuracy of CNA should be tested against its own predictions to determine its efficiency. The validation of the recommendations coming out of the CNA will therefore rely on these measures.

Conducting a CNA can be a time-consuming task where a significant amount of expertise must be brought together. Previous authors of CNAs, e.g., Lyal (2004) and Taylor (2006) suggest setting up a multi-disciplinary steering group to provide an optimum understanding of the field.

## Section 3: Overview of the CNA Workflow

To assess biodiversity information, an optimum CNA should involve the following seven steps:

1. Set purpose, scope and objectives of the study
2. Identify the target audiences within the agreed framework
3. Select a set of method(s) for the CNA exercise (e.g., type of survey)
4. Identify the information required and the investigation methodology
5. Collect and analyse the information
6. Synthesize and disseminate the outcomes, and
7. Develop and implement an action plan with an appropriate monitoring and evaluation mechanism.

The first step of the assessment is the most critical step as all subsequent steps will depend on what the purpose and objectives are. Any CNA exercise, especially if it is based on surveys, is time-consuming and often resource-intensive. Failure to properly scope the exercise could thus be wasteful, as the obtained data would be unlikely to answer the questions the CNA intended to be answered. Therefore, careful consideration should be given to setting the purpose and objectives so as to balance the costs of running the exercise and the expected returns.

The last step in biodiversity information CNA is equally important: results from the CNA would otherwise remain a mere academic exercise. An action plan is a natural consequence of any CNA, for information gaps and unfilled requirements identified

through the exercise need to be addressed. In turn, the action plan should include mechanisms to ensure that it caters to the identified requirements.

A CNA tries to predict what types of information or data will be required by practitioners or stakeholders, in order to facilitate the availability of such information. This provides a ready mechanism to evaluate whether the CNA has been successful. If the information becomes available but it is not used as assessed by follow-up actions, the CNA exercise needs to be critically reviewed and eventually redefined and/or revised.

## Section 4: Steps in a CNA

### Step 1: Set purpose, scope and objectives of the study

In general terms, a CNA produces information about stakeholder's uses, ideas, wishes, methods, needs, and preferences. The purpose of collecting this information can be multiple:

1. To develop plans. In order to plan data gathering and data maintenance programmes, institutions sharing data (and organizations collating and curating them, such as GBIF) need to know who the potential users are, their demand for data, and the feasibility of maintaining the infrastructure required to effectively serve the data to them.
2. To help define and solve problems. The current biodiversity crisis can be decomposed into many individual components, such as species range contractions or shifts, impact of invasive species, or ecosystem changes. Often, these issues become evident only when sufficient data are available (see e.g. Otegui & Ariño, 2009), and a CNA can identify problems that require more data to solve. Some of these events can become critical environmental problems, and a CNA may help in allocating resources to make data available sooner in order to solve them.
3. To help decision makers and planners set priorities. Decision-makers and funders are often faced with having more plans than they can afford. To prioritize potential conservation programmes, they need to know the impact of providing these. Content needs assessment provides a method to learn about what has already been done and how that has helped in decision making. This allows the decision- or policy-maker to make informed decisions about needed investments.

When designing a particular biodiversity CNA exercise, the scope should also be clearly delimited. Content needs may refer to:

- What is required: raw data, processed data, ranges, imagery, etc.
- The provenance, sources, origin of the data
- The geographical range or time frames relevant to the purpose
- Quantities of, and qualifiers for, required data
- The quality level, or indicators about the uncertainty or reliability of the data
- The targeted taxonomic groups

The purpose (what the CNA is intended for) and the scope (what types of information it is concerned with) are linked together to attain objectives. For example, a CNA aimed at policy makers will likely have to address questions such as, what target groups are important (e.g., CITES-listed species) or what types of data will likely be relevant for designing protective regulations (e.g., whether presence-only or abundance data are required). On the other hand, scientists might be more concerned with indicators of accuracy for presence data, uncertainty, or exchange formats that may be relevant to ensure availability of high quality data that are fit-for-purpose.

Ultimately, assessing content requirements leads to actions by stakeholders toward fulfilling them. As actions will generally require resources, priorities can be set among the list of identified needs through objective criteria. However, while such prioritization can be determined during data collection as one of the aspects, actual prioritization should best be left to the stakeholders themselves, according to their agendas.

## Step 2: Identify the target audience

Determination of the target audience who will be the subject of the CNA should be an integral part of the objectives. Selection of the target audience depends upon the questions that are identified and what geographical and thematic scale data publishers, and/or biodiversity information networks wish to address the issues. For instance, if the objective of the CNA exercise is to determine user needs for better management of biotic resources in a given protected area, then the target audience of such an exercise would be the protected area's managers, policy makers, local, state and federal administrators, biodiversity research institutions, non-governmental organizations, citizens from fringe areas, etc. (Chavan et al., 2010). This list can be expanded as needed.

Target audiences will generally belong to one or more of a number of categories. The CNA exercise should take into account the types of requirements that are more likely to be

associated with the categories and be tailored to suit specific characteristics and considerations. Below we list some examples. **Error! Reference source not found.**

| Category | Members | Characteristics | Tailoring |
|---|---|---|---|
| Research | - Scientists<br>- Programme officers<br>- Evaluators | - Fine-grained detail<br>- High accuracy<br>- Fringe data (often unknown)<br>- New data types | - Depth/specificity of questions<br>- Many options in multiple-choice responses<br>- Space for suggestions<br>- Terminology (intensive literature pre-survey) |
| Conservation, management | - Resource managers<br>- Technicians<br>- Consultants | - Immediacy<br>- Completeness, fitness-for-use assessment<br>- Sourcing<br>- Sensitiveness | - Field-specific terminology, ontologies<br>- Taxon-group lists<br>- Indexing/listing categories<br>- Language localization |
| Industry and applied science | - Executives<br>- Engineers<br>- Researchers<br>- Technicians | - Accuracy<br>- Immediacy<br>- Completeness, fitness-for-use assessment<br>- Sourcing | - Closed lists<br>- Trust measures, ground truthing<br>- Cost/effort considerations<br>- Effectiveness perception |
| Policy | - Policy makers<br>- Decision makers | - Digested, processed data<br>- Referral | - Localization<br>- Language |
| Education and awareness | - NGO<br>- Schools<br>- Educators<br>- Academics | - Approximate data<br>- Derived data<br>- Generalizations | - Flexible response models<br>- Language localization |

## Step 3: Selection of a set of methods

Two broad categories of methods are available for a CNA, information mining and surveys. Both types of methods can be retrospective or prospective. Information mining is largely retrospective, collating documented uses in response to specific needs. However, it can also be used to construct models and forecast possible developments based on discovered

trends on the use of data. Surveys can be both retrospective, providing information about past needs, and prospective, describing what users need to further their research.

**Information mining** is done by the CNA practitioner on existing evidence. It will require research to find such evidence. This approach may include:

- Literature review and analysis:
    - Scholarly papers
    - Books
    - Data papers
    - Grey literature
    - Media
- Case studies:
    - Reports
    - Projects
- Information mining:
    - Assessments of assets (what types of databases have been developed and/or published)
    - Searches in databases
    - Database analysis (comparing what fields are frequently used)
    - Citizen-science sourcing

**Surveys** are prepared and executed by the CNA practitioner but will depend on other experts to provide the data to be collated and analyzed. They can take several forms:

- Surveys:
    - Online surveys, generally using a web tool
    - Offline surveys, usually through e-mail or written forms
- Workshops and brainstorming sessions
    - On-site meetings
    - Remote:
        - Webinars, online discussions (asynchronous)
        - Video or teleconferences (synchronous)
    - Public hearings
- Interviews
    - In person
    - Remote

- Longitudinal studies (commissioned cases)

### 3.1. Things to consider in choosing methods

Regardless of the chosen method there is a trade-off between various factors. Some of these factors are intrinsic to the type of method, while others may depend on how these methods are applied. Some of these factors and considerations include:

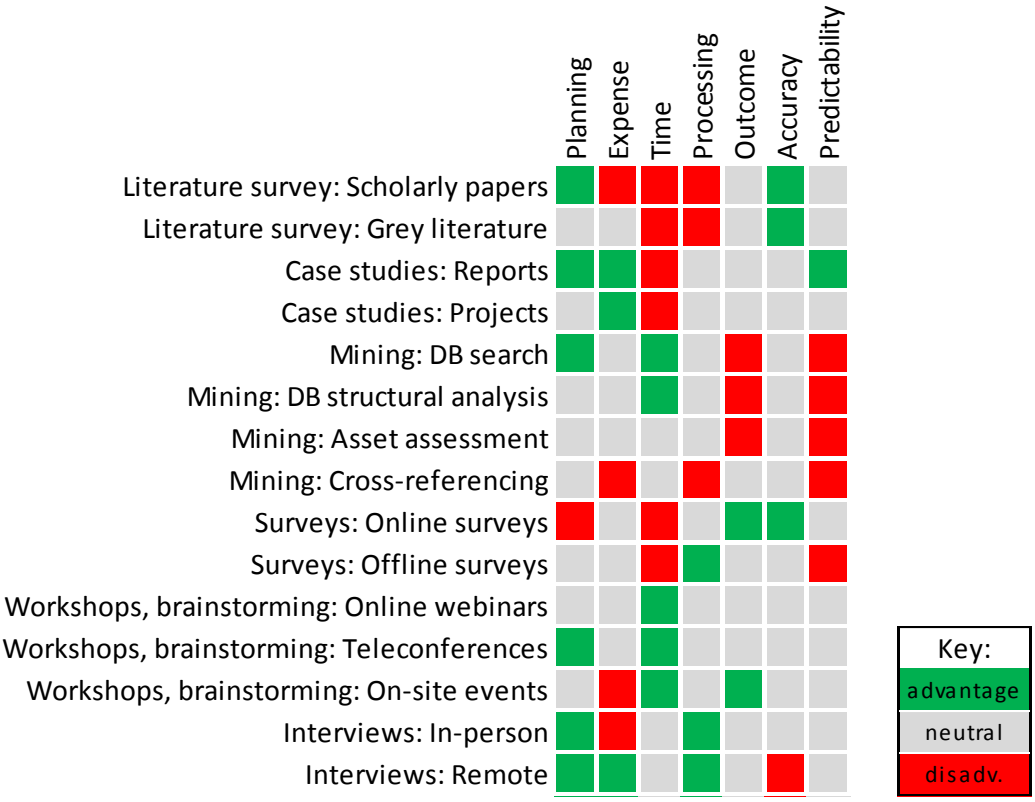
- Planning requirements, level of detail
    - **High**: Detailed planning. For example, a web-based, multiple-choice survey.
    - **Low**: Planning reduced to outlines; details to emerge from the exercise. For example, brainstorming.
- Expense, resource use
    - **Costly**: Requires considerable resources. For example, in-person interviews or research projects such as literature reviews.
    - **Inexpensive**: Expenses are reduced by including methods that fit within general running costs, or can be diluted among participants. For example, e-mail surveys.
- Time costs, immediacy
    - **Lengthy process**: Results require a long time to be gathered and consumed, often due to work delegated to a few individuals who must gather and analyse the data, or do research. For example, analysing and normalizing data from case studies.
    - **Short-term results**: Results can be produced quickly because they have been pre-processed, or just need to be collated. For example, textual reports from limited seminars and consultations that are collated together.
- Post-processing requirements
    - **High**: Results cannot be used unless significantly reprocessed, due to disparate formats, complex data extraction from textual information, the high volume of data, and/or the need for recoding data for statistical analysis. For example, indicators found in literature reviews, surveys needing recoding, or deep analysis of extensive surveys.
    - **Low**: Post-processing not generally required, as the data are already amenable for presentation without requiring further manipulation. For example, simple e-mail surveys.

- Outcomes
    - **Many**: The CNA addresses multiple questions and subjects, with fine detail. For example, extensive surveys or large seminars with parallel sessions.
    - **Concentrated**: The CNA is designed to assess one or few particular questions. For example, an e-mail questionnaire or consultation.
- Accuracy and reliability
    - **High**: Results represent the state of the art and come from a wide user base. For example, a precisely targeted survey that reaches a significant fraction of the potential users.
    - **Low**: Results come from a small representative user base, or a user base that is biased. For example, interviews in a small circle of collaborators that may not represent the breadth of the user base.
- Prospective capacity
    - **High**: The CNA produces an assessment including trends that may eventually prove accurate. For example, retrospective analyses on former CNA exercises compared to the current situation.
    - **Low**: The CNA is not concerned with predicting what users will need in the future in terms of biodiversity data, but analyzes what they are using now. For example, a survey on current uses and needs.

In general, the higher the investment in planning, resources, and time and user base, the better, more accurate and more reliable the results. Predictive capacity of the CNA (i.e., the ability of the assessment to correctly determine what users will need from the assessment onwards) will in turn depend on correctly converting reliable, material results into predictive trends.

Each method has advantages and disadvantages according to how it scores against these factors. **Table 1** summarizes the advantages and disadvantages of various methods. No method scores only advantages, and therefore careful consideration should be given to whether the drawbacks may be overcome.

**Table 1.** Advantages (green) and disadvantages (red) of several CNA methods as compared to others in a set of categories. Grey: Neutral (not particularly advantageous or disadvantageous).




## 3.2. Criteria for choosing appropriate method(s)

Some of the factors described in previous sections contribute to the costs of the exercise in terms of manpower, expenses, or time; other factors describe the advantages afforded by the exercise. These factors can then become selection criteria by weighing the relative advantages and disadvantages against the constraints and allowances of the exercise. We may include within "costs" the effort (time, brainpower, processing requirements) and expenses (resource allocation, expenditure), and as "benefits" the returns, accuracy and predictive capacity. Pitting costs against benefits may help in the selection of a particular type of CNA according to the desired level of results, budgets or other criteria.

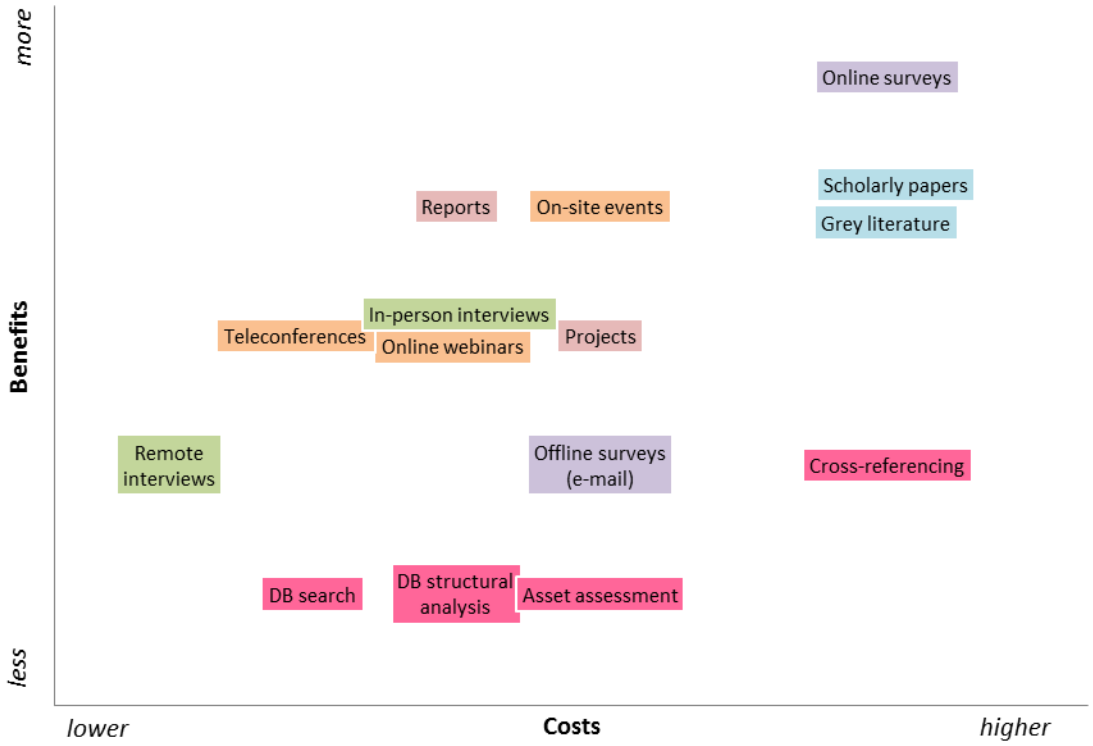



**Figure 1.** Analysis of costs-benefits for various CNA methods. Units are relative. Lavender: surveys; orange: workshops and brainstorming; pink: information mining; blue: literature review; green: interviews; brown: case studies.

In most cases, the CNA types belonging to one category behave rather similarly in the costs/benefits plots, which could, in principle, allow for some simplification. However, there may be cases where certain kinds of costs can be more readily assumed than others; separating costs would add a number of dimensions to the analysis. We may collapse the CNA types into main categories and separate the kind of costs as in **Table 2**. Here each family of methods falls into a particular combination of level of expense/resource use and effort, to yield a benefit level relative to other families. Surveys and case studies seem to represent the best compromise between costs (although the effort is significant) and high returns, but literature review, a generally lengthy and detailed process, may also be rewarded with excellent results.

A review of a decade of biodiversity CNA exercises (see section 5) has revealed that surveys are the most often selected method, perhaps representing a good perceived trade-off between accuracy and costs.

**Table 2.** CNA categories classified according to two cost components (expenses and effort) and benefits (darker gray: higher benefit; lighter gray: lower benefit).

| | | EFFORT | | |
|---|---|---|---|---|
| | | LOW | MEDIUM | HIGH |
| EXPENSE | HIGH | Workshops | Data mining | Literature review |
| | MEDIUM | Interviews | | Surveys |
| | LOW | | Case studies | |

### 3.3. Criteria for going with multiple methods

Multiple methods can be considered when the relative drawbacks and advantages can represent different subsets. The main reason for using more than one method is to obtain complementary results. Three main guidelines can be considered:

- Select methods belonging to different categories. Often, methods in one category share features. For example, all reviews are costly in terms of time, but can produce highly accurate results. These can be complemented by a faster method such as a workshop to consider the findings.
- Use the weakness of one method as a strength in another. For example, mining databases for metadata such as database structure yields tentative information on what designers thought was important to record, but may also represent past needs. On the other hand, interviews may help assess why the designer thought that structure was important and whether it could still be valid.
- Leverage available resources. Costs of methods may impose constraints, but if a particular resource is available (for example, volunteer time within an NGO) it may pay off to select more than one method whose costs are based on time, instead of resources, to maximize returns.

In our review of eight biodiversity CNA exercises (section 10) the majority combined several methods. More generally, certain multi-mode surveys seem to elicit higher response rates (Greenlaw & Brown-Welty, 2009).

## Step 4: Identify the information required and the investigation method

Design of the survey or questionnaire is critical for obtaining relevant and accurate feedback in the form of facts and opinions from the stakeholder communities, irrespective of which method or approach is employed for the CNA exercise. Unfortunately, this is

often neglected or rushed. For example, survey questions can be ambiguous or confusing. One way to avoid this problem is to pilot test a survey with several people before administering it to a large group. There are several best practice guidebooks available (e.g. Rea & Parker, 2005 or Flower, 2001) that can help in designing a productive survey.

For the purpose of a biodiversity CNA exercise, survey questions should aim at understanding the (i) profile of data users, (ii) current trends in usage of biodiversity data, (iii) gaps in accessible data, (iv) areas where more biodiversity data are required by the major stakeholder communities, (v) qualitative and quantitative requirements of biodiversity data, (vi) requirements of ancillary data resources, etc., among other aspects. Annex 1 lists a set of questions included in the GBIF CNA exercise conducted in May-June of 2009.

We will base some examples in this guide on a number of CNA exercises that have sought to profile the uses of data by biodiversity stakeholders. Questions regarding the kinds of biodiversity data used in research, biodiversity collections, or informatics infrastructure are common. For the sake of simplicity, let's assume we are specifically interested in knowing what users need to manage data in natural science collections. Questions can then be centred on the collections themselves, their contents, the processes undertaken in them, how digitization is accomplished, what scientific production is associated, statistics about the collections such as number of accessions, and so on.

As surveys seem to constitute the most commonly used method for CNA in recent times (**Figure 6**), the remainder of this step will deal primarily with surveys.

## 4.1. Collection design

### *4.1.1. Target/core data*

A CNA exercise can produce a huge amount of data, but not all of it may be useful in the end. On the other hand, it is often difficult to decide beforehand what data may be needed according to the study's objectives. If the team designing the CNA exercise is small they may not be able to adequately represent all aspects of expertise in the field being covered. Thus, the CNA exercise may make allowances for some data eventually not being used, as long as the required data are collected. Failure to collect required data, to identify useful categories of data, or to provide enough opportunities to answer (e.g. answer options) in closed questions[1] can only be overcome by conducting a second round

[1] A *closed question* provides a set of possible answers to choose from. An *open question* allows any text or data to be put in.

of the CNA, or by indirect analysis (e.g., tabulating textual responses), both cases adding to the overall time and effort of the exercise.

Generally atomized data can always be merged into more general data, but general data cannot easily be parsed into more precise components. For example, if we asked whether an author had published less than 10, 10-30, or more than 30 papers using biodiversity data resulting from a particular, databased natural history collection, we would not know what percentage of authors had never published such a paper. On the other hand, providing a more detailed answer level (such as 0, 1-4, 5-10, 11-20, and so on) may allow such insight.

The level of detail may also be disadvantageous if it is too high. In most cases, questions for a CNA could be codified as fractions, percentages or other quantities. For example, the question above could be specified as a blank field for the respondent to provide an actual number. Such a number might be accurate (for example, 42), or an estimate (for example, 40, which could perhaps mean anything between 35 and 45, or even more). However it would be difficult to ascertain the degree of precision of the estimate. For a respondent it is generally easier to choose from a menu of options than to try to produce an actual figure, which might entail some data searching.

Therefore, a design of core data collection should ensure:

- Adequate coverage of the field. A pilot review of literature may help selecting current questions being addressed.
- Ample representation of practitioners or stakeholders. The user base may not be restricted to one particular field.
- Questions for different categories of stakeholders (e.g., policy makers, researchers, technicians, volunteers) may require different depth levels.
- A limited number of questions, avoiding redundancies. Data that can be derived from other data should not be asked (e.g., whether a user uses a database to manage a NHC is a redundant question if another question asks which database s/he uses, unless the first question is a gate to a set of questions regarding use of databases).
- Adequate codification of potential answers, avoiding:
    - Imprecise answer options, and
    - Excessive (and tiresome) options that may make answering difficult.
- Provisions for textual responses that would cover unforeseen answers.

### *4.1.2 Auxiliary data for stratification*

Many answers may need to be rescaled against some other factor to become meaningful. For example, a question about the current holdings in a respondent's collections may be interesting in itself, but to derive the rate of accrual or the individual cost of one accession we would need to know how long the collection has taken to reach its current size, what the budget across all collections is, and the total number of collections. While some of these data may be core, others may come from the profile of the respondent. Stratifying respondents by a number of parameters such as age, gender, size of institution, position, budget, etc., may help such questions become more meaningful.

## 4.2. Information survey

In this approach we review documents (scientific papers, government reports, workshop reports, databases, etc.) for relevant identified problems and previous assessments, and mine and collate such information. This gives us the present "state of the art", and in multiple-method CNAs provides a first approach to the contents to be assessed. According to Lyal, 2004, among other things this information is important for:

- Questioning how users employ currently-available information and facilities; and
- Assist in asking relevant institutions how they respond to needs from the area being assessed.

The methods for gathering this information will depend on whether it is structured or not. Scientific literature and reports need to be parsed, and separate lists compiled. For example, GBIF has compiled in Mendeley[2] a list of papers using data gathered through GBIF. Tagging the papers according to content allow simple statistics to be derived describing which fields have apparently benefited most from the availability of data.

On the other hand, databases can be consulted directly, although there are two types of information that can be retrieved:

- structure of the databases (metadata), and
- content in the databases.

These data also need to be compiled, although in the case of databases the compilation can be done automatically by structured queries, for example looking at the frequency of selected keywords or subjects in the datasets. To accomplish this, a basic knowledge of database management may be necessary.

[2] http://www.mendeley.com/groups/1068301/gbif-public-library/

## 4.3. Questionnaire

In most cases, surveys will be conducted (Crawford, 1997) through questionnaires sent through electronic means. These should cover both core data and additional data that can allow for stratification of core data.

It is difficult to achieve a suitable outcome without a well-designed questionnaire, but there is no theoretical basis to design a flawless one (Crawford, 1997), i.e. a "perfect" questionnaire that will provide all required answers without error, ambiguity, or uncertainty. CNA surveys could be developed to be exploratory, allowing for a measure of freedom in responses, care being taken not to ask too many or too complex questions. Specific manuals for survey design are available (e.g., Crawford, 1997 or Walonick, 2003), but the CNA design must be closely tailored to the selected target groups.

### *4.3.1 Model template of the questionnaire*

The questionnaire should contain sections aimed at different objectives of the CNA. Wherever possible, questions should not be mixed. Questions can be closed (i.e., having fixed answers) or open. Closed questions afford easier analysis, but open-ended questions may help uncover issues that the designers had overlooked. A good approach is to always add an open-ended "escape box" to any closed question.

The CNA conducted by GBIF in 2009 (Faith et al., 2013) contained six sections. Based on that survey and other similar CNA exercises (e.g., Environmental Law Institute, 2001; GEOCONNECTIONS, 2007; InBIF, 2011; Lyal, 2004; Meerman & Clabaugh, 2004; Tann et al., 2008; Taylor, 2006) an approximate model could contain the following sections:

| Section | Purpose |
|---|---|
| Respondent profile | Stratification of the questionnaire and contextualization where appropriate |
| Current use | Set of questions to assess how and for what the users are using biodiversity data now. Equivalent to the information survey. |
| Current access | Technical section aimed at discovering how or through what means (paper, databases, cloud, web, etc.) the users access or fulfill their biodiversity data requirements. |
| Nature of required data | An extensive set of questions to find out what kinds of data users require or would potentially require. |
| Data quality/quantity requirements | Level of quality and amount of data deemed useful for the user's purposes |
| Free comments | Any contributions the user could add to the survey. |

### *4.3.2 Customizing the template for national, thematic, regional or institutional surveys*

Different stakeholders may have different priorities for data types, thus benefiting from some customization (see **Error! Reference source not found.**).

- Localization is perhaps the most distinctive, and efficient, customization. Delivering a global survey in different languages elicits a much higher turnout, as evidenced by comparing the GSAP-NHC (Berendsohn et al., 2010) and CNA-TG exercises (Ariño et al., 2013), for there may be many respondents not well served by English during the survey. On the other hand, localization during the design-phase of a country-specific survey is easier to accomplish, for it can be prepared in the country's main language(s) directly.
- Thematic surveys may allow for a more specific, deeper set of questions. A reasonable limit for a questionnaire is about 20 questions, and a thematic survey may allow a narrower scope and thus a more complete coverage.
- Customization for institutional surveys, in turn, may allow for a wider questionnaire. Institutional responses may be undertaken as an institutional task, perhaps charged to a team, and therefore the time exacted from a single respondent may still be within limits while the full response set may be larger. Also, questionnaires aimed at institutions may ideally gather archival (and thus highly reliable) data, if the respondents can command human resources within the institution's infrastructure.

## 4.4. Data base

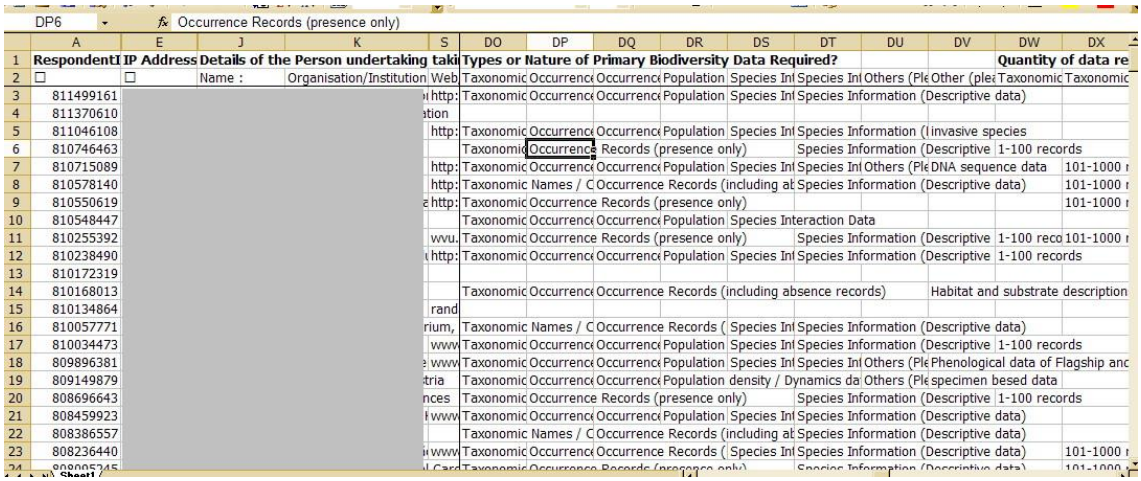
Responses to the questionnaire need to be tabulated prior to analysis unless all analytical needs are directly met by the survey software, which will seldom be the case. This will often entail constructing a database, unless the chosen method for administering the questionnaire already provides such a database. Most available on-line questionnaires, however, use a tabular/spreadsheet paradigm that has little flexibility as compared to a relational database and, for instance, prevents cross-tabulation, cross-check, or import into a statistical package.

### *4.4.1 File / Item model*

A sensible data structure will facilitate analysis, either directly or through export to some statistical software package. The database should receive the raw data from the survey software. For instance, the commonly used tool SurveyMonkey[3] produces output as a set of spreadsheet tables, recording individual respondents in rows and single options for each question as a column. Cells are filled with the selected, verbatim options (see **Figure 2**).

[3] https://www.surveymonkey.com/

As the number of options can exceed some spreadsheet's maximum column capacity, additional sheets are produced by the software holding the remaining columns.



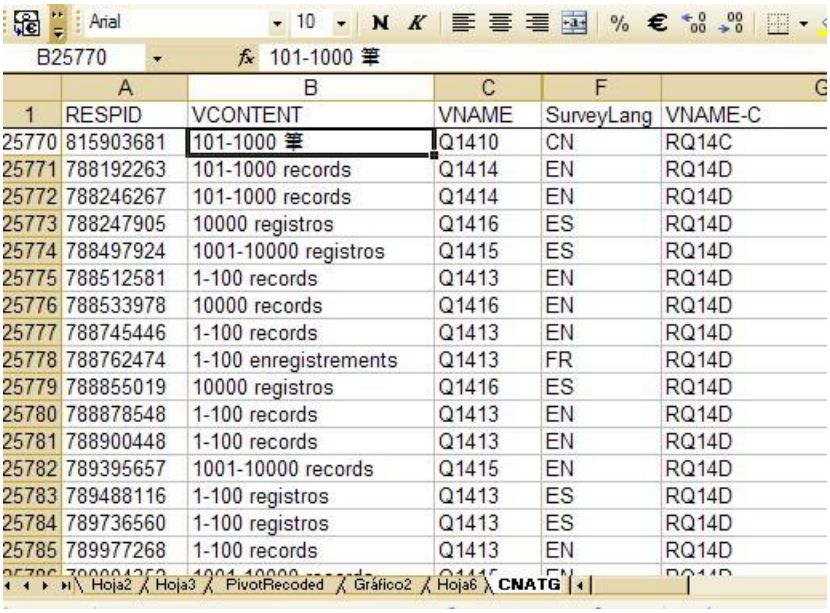
DP6 | Occurrence Records (presence only)

| | A | E | J | K | S | DO | DP | DQ | DR | DS | DT | DU | DV | DW | DX |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | RespondentI | IP Address | Details of the Person undertaking taki | | | Types or Nature of Primary Biodiversity Data Required? | | | | | | | | Quantity of data re | |
| 2 | ☐ | ☐ | Name : | Organisation/Institution | Web | Taxonomic | Occurrenc | Occurrenc | Population | Species Int | Species Int | Others (Ple | Other (plea | Taxonomic | Taxonomic |
| 3 | 811499161 | | | | http: | Taxonomic | Occurrenc | Occurrenc | Population | Species Int | Species Information (Descriptive data) | | | | |
| 4 | 811370610 | | | | ation | | | | | | | | | | |
| 5 | 811046108 | | | | http: | Taxonomic | Occurrenc | Occurrenc | Population | Species Int | Species Information (I | invasive species | | | |
| 6 | 810746463 | | | | | Taxonomic | Occurrenc | Records (presence only) | | | Species Information (Descriptive | | | 1-100 records | |
| 7 | 810715089 | | | | http: | Taxonomic | Occurrenc | Occurrenc | Population | Species Int | Species Int | Others (Ple | DNA sequence data | | 101-1000 r |
| 8 | 810578140 | | | | http: | Taxonomic | Names / C | Occurrence Records (including ab | | | Species Information (Descriptive data) | | | | 101-1000 r |
| 9 | 810550619 | | | | e http: | Taxonomic | Occurrence Records (presence only) | | | | | | | | 101-1000 r |
| 10 | 810548447 | | | | | Taxonomic | Occurrenc | Occurrenc | Population | Species Interaction Data | | | | | |
| 11 | 810255392 | | | | wvu. | Taxonomic | Occurrence Records (presence only) | | | | Species Information (Descriptive | | | 1-100 reco | 101-1000 r |
| 12 | 810238490 | | | | t http: | Taxonomic | Occurrenc | Occurrenc | Population | Species Int | Species Information (Descriptive | | | 1-100 records | |
| 13 | 810172319 | | | | | | | | | | | | | | |
| 14 | 810168013 | | | | | Taxonomic | Occurrenc | Occurrence Records (including absence records) | | | | | Habitat and substrate description | | |
| 15 | 810134864 | | | | rand | | | | | | | | | | |
| 16 | 810057771 | | | | rium, | Taxonomic | Names / C | Occurrence Records ( | | Species Int | Species Information (Descriptive data) | | | | |
| 17 | 810034473 | | | | www | Taxonomic | Occurrenc | Occurrenc | Population | Species Int | Species Information (Descriptive | | | 1-100 records | |
| 18 | 809896381 | | | | e www | Taxonomic | Occurrenc | Occurrenc | Population | Species Int | Species Int | Others (Ple | Phenological data of Flagship and | | |
| 19 | 809149879 | | | | tria | Taxonomic | Occurrenc | Occurrenc | Population density / Dynamics da | | | Others (Ple | specimen besed data | | |
| 20 | 808696643 | | | | nces | Taxonomic | Occurrence Records (presence only) | | | | Species Information (Descriptive | | | 1-100 records | |
| 21 | 808459923 | | | | f www | Taxonomic | Occurrenc | Occurrenc | Population | Species Int | Species Information (Descriptive data) | | | | |
| 22 | 808386557 | | | | | Taxonomic | Names / C | Occurrence Records (including ab | | | Species Information (Descriptive data) | | | | |
| 23 | 808236440 | | | | u www | Taxonomic | Occurrenc | Occurrence Records ( | | Species Int | Species Information (Descriptive data) | | | | 101-1000 r |
| 24 | 808005245 | | | | | Taxonomic | Occurrence Records (presence only) | | | | Species Information (Descriptive data) | | | | 101-1000 r |

**Figure 2.** A small section of a raw file as produced by the SurveyMonkey survey software, delivered as an Excel spreadsheet. Each row corresponds to one respondent (personal data obscured).

As this layout is not amenable to direct analysis, data are transferred to a data model where each record is an individual option or response supplied by each respondent to each question (see **Figure 3** for an example in a spreadsheet). If the survey is a multiple-language one, in order to nullify language differences between surveys free-text answers coming from fixed options are recoded homogeneously across all localized surveys, and merged together into a single file. The original language must however be retained as a field, allowing for grouping when the language factor will be required later in the analysis. Also, verbatim responses (in their original language, before any recoding) need to be retained for reference as fields.

This rearrangement also allows for any number of variables and variable options in the survey output (one for each possible answer in multiple-choice questions) to be represented without limits, while also greatly reducing the dimensionality of the table to one variable for each question. In the case of multiple-choice range questions, variables can be created where a weighted index substitutes several individual options within a range by the centroid of the chosen options within that range.



| | A | B | C | F | G |
|---|---|---|---|---|---|
| 1 | RESPID | VCONTENT | VNAME | SurveyLang | VNAME-C |
| 25770 | 815903681 | 101-1000 筆 | Q1410 | CN | RQ14C |
| 25771 | 788192263 | 101-1000 records | Q1414 | EN | RQ14D |
| 25772 | 788246267 | 101-1000 records | Q1414 | EN | RQ14D |
| 25773 | 788247905 | 10000 registros | Q1416 | ES | RQ14D |
| 25774 | 788497924 | 1001-10000 registros | Q1415 | ES | RQ14D |
| 25775 | 788512581 | 1-100 records | Q1413 | EN | RQ14D |
| 25776 | 788533978 | 10000 records | Q1416 | EN | RQ14D |
| 25777 | 788745446 | 1-100 records | Q1413 | EN | RQ14D |
| 25778 | 788762474 | 1-100 enregistrements | Q1413 | FR | RQ14D |
| 25779 | 788855019 | 10000 registros | Q1416 | ES | RQ14D |
| 25780 | 788878548 | 1-100 records | Q1413 | EN | RQ14D |
| 25781 | 788900448 | 1-100 records | Q1413 | EN | RQ14D |
| 25782 | 789395657 | 1001-10000 records | Q1415 | EN | RQ14D |
| 25783 | 789488116 | 1-100 registros | Q1413 | ES | RQ14D |
| 25784 | 789736560 | 1-100 registros | Q1413 | ES | RQ14D |
| 25785 | 789977268 | 1-100 records | Q1413 | EN | RQ14D |

**Figure 3**. Example of data rearranged as an item-oriented database in an Excel datasheet for frequency analysis. Each row is an individually selected option in the survey, with fields for record ID (RESPID), question code (VNAME-C), original language (SurveyLang), delocalized answer option code (VNAME), and verbatim answer (VCONTENT).

### *4.4.2 Workflow from raw data to statistics*

The unified datasheet should always be carefully checked for duplicates, errors and misalignments before summary statistics and frequency data are compiled (**Figure 4**. Flow chart of the analytical design for a multi-language Content Needs Assessment (CNA) Survey. CN-S: Simplified Chinese; CN-T: Traditional Chinese; DB: database; EN: English; ES: Spanish; FR: French; RU: Russian; QC: quality control). Some additional data can be collected from other sources for further analysis, e.g., the respondent's city coordinates can be obtained from georeferencing facilities if there is a need to geolocate responses for stratification purposes.

Some data will need to be cross-tabulated, recoded or summarized. This is easily done using pivot tables or issuing queries directly to the database. Often, the responses are subject to frequency analyses, either directly on the data variables, or on the cross-tabulations among variables.

Frequencies can then be plotted or mapped as appropriate in order to address trends from questions either originally designed in the survey's goals, or emerging from the analytical process. Also, hypotheses can be tested as needed.

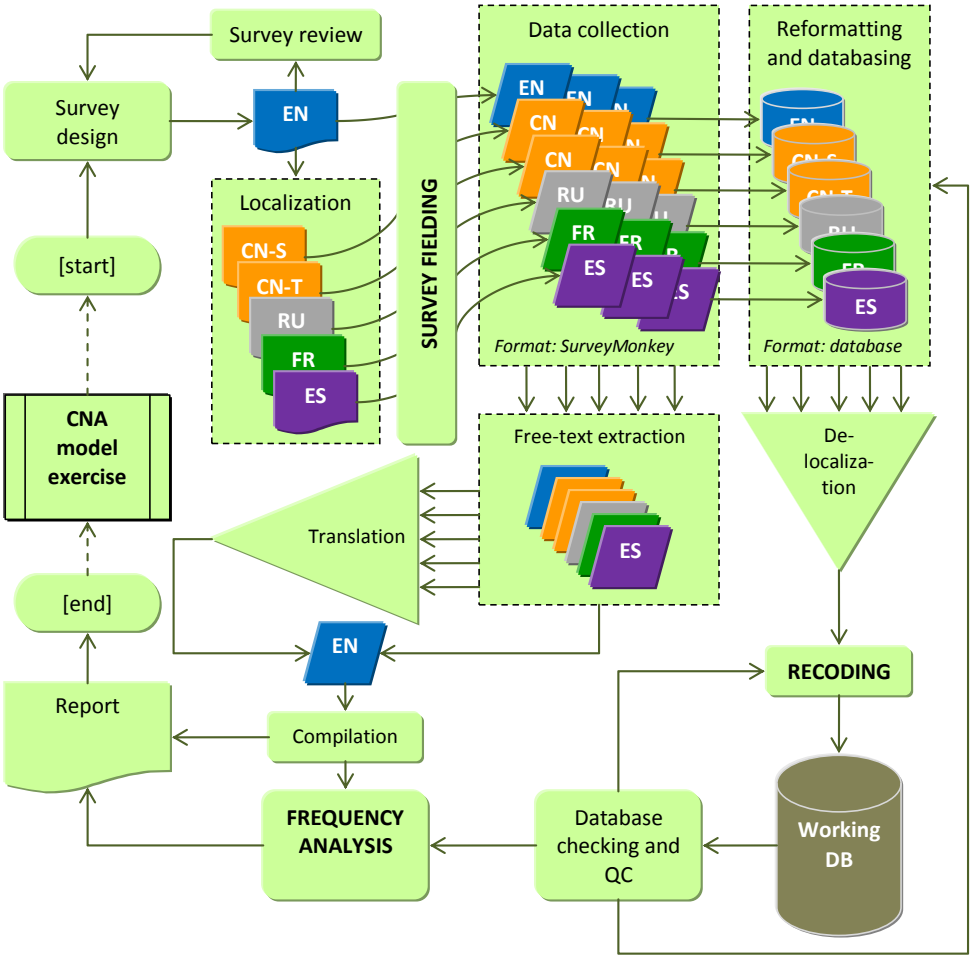



**Figure 4.** Flow chart of the analytical design for a multi-language Content Needs Assessment (CNA) Survey. CN-S: Simplified Chinese; CN-T: Traditional Chinese; DB: database; EN: English; ES: Spanish; FR: French; RU: Russian; QC: quality control

In summary, a good practice for designing the data gathering process should include:

- Relevant coverage of the field and the community
- Make sure the information collected is as atomized as possible
- Do not go into unnecessary detail in the requested information
- Avoid redundant questions that can be derived during post-processing
- Gather auxiliary data and metadata for stratification
- Customize the questionnaire according to your potential respondents' depth of knowledge
- Consider modelling your questionnaire in six sections: Respondent profile, Current use, Current access, Nature of required data, Data quality/quantity requirements, Free comments
- Allow free text input as a complement to each fixed or multiple choice question where relevant
- Get the resulting data into an item-oriented data structure to simplify recoding

# Step 5: Collect and analyse the information

## 5.1. Determining the sample size

Since surveys and interviews are the most preferred method for CNA exercises, the question of how many answers are needed to achieve significance arises. The number of answers will ultimately be limited by the size of the audience to be polled in on-line surveys, or by the resources available in interviews. The actual turnout, however, will be determined by the size of the audience that has actually been reached and the response rate of those reached. These realities can be affected by how the questionnaire has been designed, e.g., the perceived difficulty of filling out the survey, or the level of interest it can attain, e.g., whether the audience sampled has been correctly targeted.

From a statistical point of view, the traditional frequency analysis will require different sample sizes according to the complexity of the desired answers and to whether a known statistical distribution is required for inference. For example, if we were interested in estimating the percentage of practitioners using a certain data management application, we would require just 4 responses for a 95% confidence interval if we were happy to go with a $B$ = ±50% potential error of that estimate ($n = 1/B^2$ for a simple proportion), an allowance widely used for initial assessments (Krebs, 1999). However, for management implications (e.g., deciding on a software purchase) a ±25% error estimate for that fraction would be desirable, resulting in 16 answers (or 100 if we were actually researching it, thus not settling for an error beyond 10%). Any good statistical primer (e.g., Sokal & Rohlf, 2012) will provide formulas for estimating sample sizes according to desired confidence intervals, precision, accuracy, and expected distribution of responses.

The best practice is to try to define what types of answers we would like to get and what allowable error is acceptable in advance, then use such estimates and allowable errors to inform the required sample size according to tables or formulae in statistical manuals (e.g., Sokal & Rohlf 2012).

In summary, a good practice for determining sample size should include:

- Deciding your desired level of accuracy for the answers
- Determining the size beforehand using statistical methods based on desired accuracy
- If in doubt, erring on the more numerous side (trying to get more answers than needed)

## 5.2. Overcoming biases in sampling

A number of biases may be present in the survey that may affect the accuracy of the results if left unchecked. Among them:

- *Selective sampling*: The survey reaches some communities selectively, leaving out others. For example, a taxonomy-oriented CNA for a vertebrate network is sent to eBird[4] users but not to FishBase[5] users
- *Stray data*: The survey is undertaken by non-targets that may provide spurious answers ("overcover")
- *Low turnout*: Potential respondents ignore, or pull out from the questionnaire
- *Systematic bias*: Respondents meeting certain criteria are more likely to respond, introducing that bias in the answers. For example, a complex questionnaire administered in English only may elicit attrition from non-native speakers, inducing a cultural bias or the underrepresentation of the non-English world.

Overcoming these biases is best done during the planning phase. Some potential actions may include:

- Collating an ample set of potential targets and equalizing observed groups. For example, if we are to send a questionnaire to one mailing list of 400 ichthyologists and to another one with 4000 ornithologists, the views of ornithologists will have a heavier statistical weight. If our study lends the same importance to both groups, we may draw a random sample of, say, 200 ichthyologists and 200 ornithologists to de-trend the answers.
- Ensuring that all potentially interested stakeholders are represented, by outreaching pyramidally: the steering group enrolls a second level of contacts that in turn outreach to a third level, and so on.
- Designing identification methods in the questionnaire to filter out undesired answers based on adequacy, e.g., requiring a "test question" to prove that the respondent belongs to a certain group.
- If using a questionnaire, making it relatively lean so as to avoid respondent's "survey fatigue".
- If resorting to interviews, ensure that the interview is conducted, if possible, in the respondent's language.

[4] www.ebird.org
[5] www.fishbase.org

- When analysing literature, ensure that a variety of sources are represented. For example, do not ignore articles or reports that are not readily available through the institution's library access or subscriptions. If necessary, obtain access to identified but inaccessible papers through colleagues or other members of the steering group.

In summary, a good practice for overcoming bias should include:

- Avoid selective sampling based on partial lists or chosen closed groups
- Ensure wide opportunity for answers, using multiple language versions if possible for global coverage surveys
- Make questionnaires affordable in time and complexity, avoid fatigue
- Balance out the audience of respondents across target categories
- Recognize and remove bogus respondents ("trolls")

### 5.3. Questionnaires: Fielding methods

Questionnaire-based surveys can be fielded in one of three ways:

- Electronically through a web page,
- Electronically through e-mail,
- In paper form.

Currently paper questionnaires are on the wane. However, e-mailed questionnaires may require the same type of processing as paper questionnaires (being essentially identical excepting the medium): data must be transferred to the receiving database. An on-line questionnaire, on the other hand, means the least amount of processing as data are entered directly from the application. A number of services are available for building and administering on-line questionnaires. Vehovar et al. (2012) list an extensive collection of online survey software classified according to a number of criteria. As of 2012 their list includes more than 300 different systems and applications.

Be it online or through e-mail, potential respondents must be made aware of the survey. This is generally done now through announcements in web pages, forums, listservers, interest groups, or by e-mailing.

Constructing a set of mailing lists and potential outlets is therefore a highly significant determinant of the portion of the audience that will ultimately be sampled. In addition, members of the steering committee may compile lists of potential respondents, perhaps as letters to heads of organizations where significant numbers of respondents may exist.

In our experience, a motivating letter to key people that may have the questionnaire cascaded down their institutions may attract a high number of respondents. However, care should be taken to avoid sampling bias as explained before, by carefully balancing targeted groups.

In summary, a good practice for fielding methods should include:

- Use web-based questionnaires allowing direct entry when possible
- Ensure wide dissemination of the questionnaire through sites, fora and mailing lists
- Make key people aware of the questionnaire and lobby for dissemination through their institutions
- Avoid concentrating on closed groups while overlooking others

## 5.4. Methods of data collection

The method of data collection will be highly dependent on the type of assessment and data source, although there are a number of tasks common to various methods.

In most cases, the best cost-effective data entry method will be a survey tool delivering a questionnaire. However there are several modes of data entry.

### *5.4.1 Modes of data entry*

Most surveys will produce data that can be collected **directly**. A web survey will use forms or applets that will convey the answer to the question (be it a choice from a list, a range, or textual answer) directly to a database collating the answers. On the other hand, some forms may be collected **indirectly** using an e-mail based approach, where the answers will be collated and sent to the receiver in a structured e-mail. The e-mail must then be processed. If it includes some kind of markup (e.g., an xml schema), the process can be automated and the answers, in turn, be entered directly into a database.

Interviews also collect data directly, although it is the interviewer who conveys the interviewee's answers to the database. In effect, the interviewer acts as a surrogate of the interviewee's actions, although there are a number of advantages: e.g., the interviewer can clarify questions on the fly, and will likely be much more acquainted with the questionnaire (thereby reducing errors by misinterpretation). Such entry can be termed **supervised**, as opposed to the **unsupervised** data entry through an online tool.

Online surveys and interviews may allow for fixed-response questions (e.g., multiple-choice, drop-down, range-select) that can go directly as elements into a database. However, answers can also take the form of free text. Collecting these data verbatim is necessary, but also makes it difficult to categorize the data. The surveyor may be required to **interpret** the data of interest from the textual answers, and perhaps recode it into

homogeneous categories. For example, in a digitization assessment we may be interested in determining the digitizing method for a collection of plants, and have prepared a drop-down list containing the most common methods of imaging (scanner, inverted scanner, camera, scan-camera). However, in a free-text box the interviewee can be given a chance to specify a different method, perhaps a novel one (e.g., a field camera). Such new options could prompt the insertion of a new category in the database.

Manual interpretation will also often be necessary in surveys where the data source is literature or other evidence. The practitioner will parse text for data and context and then fill in the corresponding data element into the survey database. In some circumstances and for certain data, automated data extraction procedures might help such as locating taxonomic names (e.g., GoldenGate[6] or Global Names[7]; see Penev et al., 2011 for a review of schemas). This parsing exercise can in turn be combined with a survey-type data collection. For example, sources can be published to a group of volunteers who scan them for data and then use a survey form to fill in these data. Crowdsourcing initiatives designed for biodiversity data collection from digitized sources (e.g., the ALA Volunteer Portal; Flemons & Berents 2012, Flemons 2011) can thus be used for a CNA when applied to literature, reports, or other similar evidence.

### *5.4.2 Choice of data entry / management applications*

As long as a survey application meets the requirement of the survey (e.g., response limits, processing, types of reports), selecting a survey system can often be a matter of convenience. For example, the institution fielding the survey may already have a license for a particular system. If cost considerations are of concern, for small surveys, a variety of simple, free web-based applications exist. Vehovar et al. (2012) include in their list about 45 free applications, nearly half of them being open source. A further one hundred follow the "*freemium*" model, offering limited functionality for small surveys but stepping up in price as more features are requested. These commonly include the ability to generate reports and to download the response table, as do an unlimited number of collected responses. Prices can range anywhere from small monthly fees of around US$ 10 a month up to more than US$ 10,000 for some down payment corporate licenses for high-end solutions with a typical 20 per cent yearly maintenance cost, offering almost unlimited storage, data collection, analysis and reporting.

Most survey applications allow for multiple question types and many can collect and digest responses producing summary reports. For many users, that is all they would possibly

[6] http://plazi.org/?q=GoldenGATE
[7] http://gnrd.globalnames.org/

need. Reporting capabilities however vary, and so does the downloadability of data in a format amenable to processing (e.g., databases or data tables). Often these capabilities come at a premium, along with larger fielding or retrieval capabilities (e.g., several *freemium* models offer a limited number of collectable answers, such as 100 or 1000 over a fixed time period unless one pays a premium).

The examined CNA exercises had modest response rates (less than 1,000 respondents) and could thus have been made possible currently with many *freemium* models at low cost, although post-processing of downloaded response tables would be required for more in-depth analysis than the basic reporting facilities of the applications would warrant. Limited budgets should therefore not be a deterrent for reasonable CNA exercises as long as the practitioner can embark on custom analysis if needed.

In summary, a good practice for data collection should include:

- Use of survey software allowing for direct unsupervised data entry if possible
- Making sure your chosen application allows for at minimum download of data tables in addition to summaries and pre-cooked plots
- The ability to code and categorize free-text answers
- Removing duplicated categories filed under different names
- Using direct supervised entry of data from interviews by specialized personnel knowing the project
- Combining crowdsourcing with supervised entry when mining non-structured sources such as literature, reports, and databases.

## 5.5. Methods of analysis and interpretation

Once data have been collected, these must be analyzed to derive the information the CNA exercise has been set up to gather.

Survey tools will almost invariably produce basic summaries such as response and option frequencies or charts. According to the feature level of the tool (quite often linked to premium versions) other perks such as data table downloads, cross-tabulations or dynamic reports will also be available.

Basic summaries (e.g., how many answers belonged to each response category for each question) may need further refinement. This can be done through data analysis. In turn, data analysis may require collating a database of the responses in a highly structured manner (see ***4.4.1 File / Item model***

If an item-oriented database has been constructed, the next step is to clean the data, looking for duplicates (e.g., from parsing free-text answers), and homogenizing records, often recoding them (see **Figure 4**). Pivot tables can then be constructed as needed for the

variables, questions, or answers, quite often allowing for stratification. The general layout can be a case/parameter table where each row is an individual respondent, and columns will have the items responded. Filters can be set as required on the stratification or other variables, but the table must contain homogeneous codes for the responses. If the survey was localized in different languages, the coded responses allow for merging responses from different locales, while retaining the information of the original locale as a field (see **Figure 3**).

Excel or equivalent spreadsheets can be used for easy analysis. Pivot tables in Excel allow summaries and breakdown of responses by category, which can then be plotted for easier visualization.

In addition, tables constructed from the item-oriented database can be easily exported to statistical packages (if not done directly by the survey software). These packages can be used to statistically test hypotheses suggested by the visualizations.

In any CNA exercise, care should be taken to observe trends in the data: both those revealing actual patterns and also those suggesting biases. It is a good idea to plot data (especially frequencies of categories) against each other and against respondents' metadata, for example geographical origin or language. Segmentation biases at collecting time can be detected in this way, and perhaps corrected (for example, by randomly choosing responses so as to equalize the number of respondents in different independent categories). Also, if metadata trends are suspected, data can be de-trended by normalizing to the underlying trend factors. For example, in a digitization CNA we may suspect that the amount of digitization in countries might naturally be a simple function of the size of the country. We may de-trend digitization intensity by dividing the number of digitization projects by an indicator of the size of the country, and then examine the actual commitment of institutions to digitizing data in each country. See Ariño et al. (2013) for an example of trend analysis.

Finally, analysis should include methods specifically aimed at discovering gaps in the data that need to be filled during subsequent installments of the CNA exercise. For example, if we were assessing the global need for digitization infrastructure, we should make sure we had answers evenly spread across the globe. Thus, representing the locations of the respondents should allow us to detect where the survey failed and should be retaken (**Figure 5**).

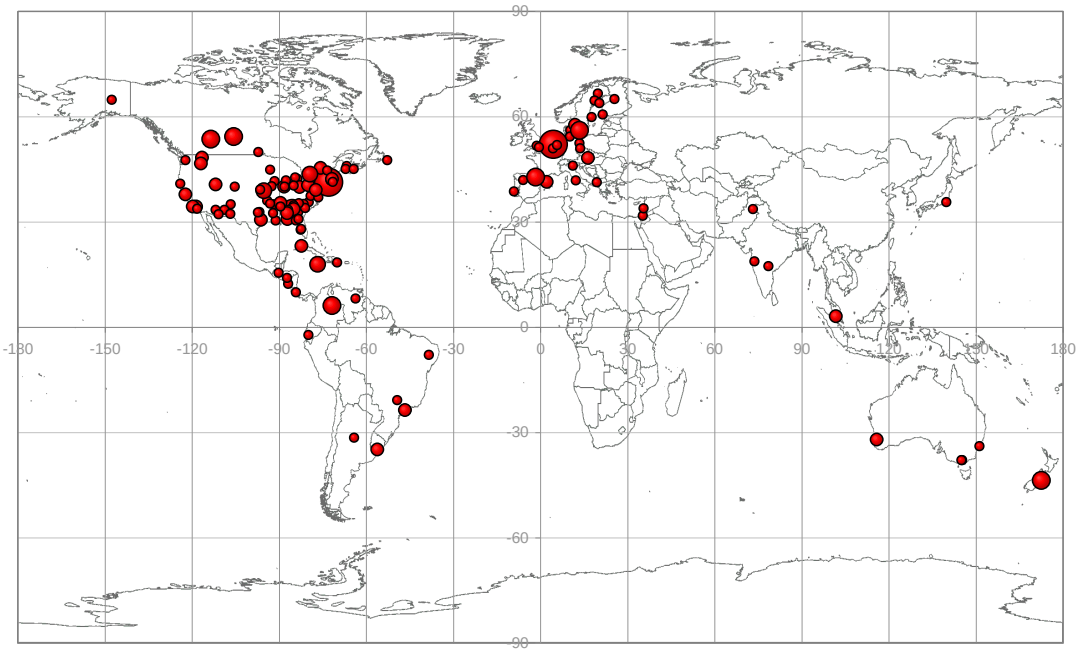



**Figure 5.** Geolocation of respondents according to institutional addresses in a survey aimed at determining needs in digitization of natural history collections (from Ariño 2010). Dot areas are proportional to the number of responses coming from the same city. Large blank areas in the map indicate a need to resample from these areas.

In summary, a good practice for analysis and interpretation should include:

- Construction of case/parameter data tables from the original database
- Use of spreadsheets and/or statistical packages
- Creation of pivot tables segmenting the data according to respondents' metadata
- Use of extensive cross-tabulation among variables
- Use of visualizations for the data whenever possible
- De-trending and relativizing data as necessary
- Performing a gap analysis on the respondent's metadata

## Step 6: Synthetize and disseminate the outcomes

### 6.1. Considerations for communicating outcomes

When content needs assessment has been completed, the final step is to present the findings to the relevant user communities as well as to existing and future data publishers. There are different ways for communicating such an outcome. Choosing an appropriate method(s) for disseminating is a topic of discussion in itself. It is best managed through

the appropriate communications arm of an organization. In the following sections we list some of the methods used for disseminating an outcome of such an exercise together with some tips on what must be done and what needs to be avoided.

### 6.2. Do's of dissemination

The first priority in any dissemination plan is returning results to study participants. Dissemination to any other stakeholder group must take place following this first step. Here are some of the do's of dissemination:

1. Adopt more than one approach to disseminate the results. Multiple approaches and combined strategy ensure wider reach to potential audiences.
2. General guidelines to ensure effective communication and usefulness include (a) being responsive, (b) being concise, (c) making it interesting, (d) highlighting key points or messages, (e) keeping it logical, (f) making sure that it is useful, (g) making it attractive, and (h) simplifying.
3. Keep it simple and easy to understand, with key messages clearly highlighted in the report or reporting formats.
4. Do explain clearly the processes employed for the exercise, and do provide an easy to understand breakdown of results, with key recommendations.
5. Facilitate an access to baseline data and algorithms if any were used.
6. Provide features for stakeholders to comment or provide feedback on generally and/or specific aspects of the report.
7. Collaborate with appropriate organizations in your field as well as general news outlets to ensure effective communication.
8. If an outcome of an exercise demands actions on part of the organization, it is a good strategy to provide a short response indicating when an appropriate action will be taken, or has been taken.

### 6.3. Optimal methods of dissemination

Key characteristics of an effective dissemination plan include: (1) orient toward the needs of the audience, using appropriate language and information levels; (2) include various dissemination methods: written including illustrations, graphs and figures; electronic and web based tools; and oral presentations at community meetings and scientific conferences; and (3) leverage existing resources, relationships, and networks fully.

Below are some of the ways that can be employed in disseminating the outcomes of a CNA exercise.

1. Media coverage
2. Press release
3. Research summary document and research articles
4. Flyers, posters, brochures and research briefs
5. Policy briefs
6. Study newsletters
7. Community publications
8. Websites and other social media
9. News and electronic media (radio, television, webcasting, etc.)
10. Local events, seminars, conferences and community meetings
11. Open letters to the community of stakeholders

This is not an exhaustive list. Communication media are always evolving. Thus, it is important to strike a balance between classical approaches and some of the trendy emerging ways to communicate.

## Step 7: Develop and implement an action plan

Once the content needs assessment is carried out, it is important to follow it up with what measures would be taken to address each outcome. Thus, the next logical step is to move forward, developing and implementing an action plan, for example as described in another guide (Chavan et al., 2010).

The action plan should at least include two generic components:

- data gap analyses, which aid the mapping of user needs against the accessible data, and
- an appropriate monitoring and evaluation mechanism.

This provides directions for further data mobilization goals, and what measures need to be adopted to achieve such goals. Eventually, the action plan may result in fulfilling the identified needs. This, in turn, may prompt users to rethink what other needs may arise as research and policy move forward — therefore, after a while a new CNA will help determining those new needs, closing the cycle.

## Section 5: Lessons learnt from the case studies

There is abundant literature on CNA, especially in the fields of social sciences, policy, and education. However, CNAs in the biodiversity realm seem surprisingly rare, although increasingly less so. In addition, other assessments seeking to understand patterns in how biodiversity is studied, used, distributed, or considered, could actually double as a CNA if they were concerned with i.e., gap analysis or the state of the art (e.g., Guralnick & Hill, 2009; Krishtalka & Humphrey, 2000; Peterson & Kluza, 2003; Peterson et al., 2010; Soberón & Peterson, 2004, 2009).

We have collected eight representative biodiversity CNAs completed in the 21st century and have examined them for general trends (Figure 6). Two of them have a global coverage while six are national/regional. Focus varied, but biodiversity informatics played a major role in all of them. They were all conducted in English, but one was also repeated in the remaining UN General Assembly official languages.

Except for a largely longitudinal study started in 2002, the length of the CNA exercises (from design and fielding to report delivery) seems to have increased steadily, from about three months in the early 2000's to almost ten towards the end of the decade. This increase is roughly matched with additional trends. The number of respondents to each survey also increased from a few dozen to several hundreds, perhaps reflecting both the expansion of the field (with more practitioners) or the need for a more in-depth assessment as time progressed.

It also seems that the preferred method(s) for the CNA exercise have evolved in this short period of time. Earlier exercises were essentially based on interviews, with some longitudinal studies in addition. However, from 2005 onwards online (or e-mail) surveys became common, at first in multiple-method procedures (with interviews) but with the last two being online surveys only. This trend may continue, as online surveys are becoming relatively affordable and easy to prepare, and once fielded they require much less of the researcher's time than interviews, and allow for a much greater turnout. However, online surveys may have two distinct features that should be accounted for:

1. They may require extensive, time-consuming preparation, as there is normally no possibility of on-the-fly adaptation as it is possible with literature or interviews. For example, a literature survey may reveal new avenues for content as it proceeds; and interviews may prompt new questions not thought of at the beginning. When crafting the online survey, the whole target field, be it narrow or wide, must be covered from the start.

2. Inexpensive or free survey tools often come limited in analytical power or depth (see 5.4.2 Choice of data entry / management applications), often requiring quite extensive post-processing and data managing to extract the answers, while in commercial packages that power is often included, trading effort for price.

As CNAs tend to be extensive and complex when conducted through surveys, the seemingly low cost of an online survey (indeed true, and appropriate for a simple set of simple questions) may in the end soar if the CNA grows in scope. Care should be exercised in determining what will be the actual cost (in money, time and/or effort) of the survey.

A change can also be perceived in how data are processed. Elaboration of outcomes and outputs is shifting toward plots and visualizations, while narrative text and tables/listing continue in reports. These may thus include digested information along with more summarized data, which is perhaps easier to understand while using less reader time—the reader can get a general idea through visualizations, referring to tables/text for deeper insights.

Finally, regular, peer-reviewed papers are starting to appear as dissemination media for the results, while reports were the norm in the first half of the decade. This may also reflect the increased maturity of the field.

| | | | | | | Survey type | | | Elaboration | | | Output | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Assessment (coverage: global / regional) | Focus of CNA | Year | length (months) | Language(s) | Size (respondents) | Online ■ / email | Interview | Longitudinal | Text, narrative | Tables, Lists | Plots, visualizations | Report | Journal article(s) |
| InBIF User Needs Assessment Report | Biodiversity | 2011 | 10 | EN | 170 | ■ | | | ■ | ■ | ■ | ■ | |
| GBIF Content Needs Assessment Task Group | Biodiversity use | 2009 | 8 | EN,ES,FR, RU,CN | 750 | ■ | | | ■ | ■ | ■ | | ■ |
| Atlas of Living Australia: User Needs Analysis | Users of ALA | 2008 | 6 | EN | 480 | ■ | ■ | ■ | ■ | ■ | ■ | ■ | |
| GBIF GSAP on Digitization of Natural History Collections | Digitization | 2008 | 5 | EN | 201 | ■ | ■ | | ■ | ■ | ■ | | ■ |
| United Kingdom Taxonomic Needs Assessment | Taxonomy | 2004 | 3 | EN | 99 | ■ | ■ | | ■ | ■ | | ■ | |
| Belize Biodiversity CHM - User Needs Assessment | CHM | 2004 | 4 | EN | 32 | | ■ | | ■ | ■ | ■ | ■ | |
| Federal Biosystematics Partnership (Canada) | Biosystematics capacity | 2002 | 24 | EN | 19 | | ■ | ■ | ■ | ■ | | ■ | |
| New York State Biodiversity Project: Needs Assessment | Biodiversity | 2001 | 3 | EN | 57 | | ■ | ■ | ■ | ■ | | ■ | |

**Figure 6.** Comparison of eight biodiversity CNA exercises during the last decade.

## Section 6: References


Ariño, A. H. (2010). *Selected highlights of the GSAP-NHC TG survey: Issues related to the Size of Universe question*. (pp. 1–15). Report for the GBIF GSAP-NHC Task Group.

Ariño, A. H., Chavan, V., & Faith, D. P. (2013). Assessment of user needs of primary biodiversity data: Analysis, Concerns, and Challenges. *Biodiversity and Ecology*, *8*(1), 59–93.

Ariño, A. H., Chavan, V., & King, N. (2011). The Biodiversity Informatics Potential Index. *BMC Bioinformatics*, *12*(Suppl 15), S4. doi:10.1186/1471-2105-12-S15-S4

Ariño, A. H., & Otegui, J. (2008). Sampling Biodiversity Sampling. In A. L. Weitzman & L. Belbin (Eds.), *Proceedings of TDWG* (p. 107). Biodiversity Information Standards (TDWG).

Berendsohn, W. G., Chavan, V., & Macklin, J. A. (2010). Recommendations of the GBIF Task Group on Global Strategy and Action Plan for the mobilization of natural history collections data. *Biodiversity Informatics*, *7*, 67–71.

Borgman, C. L. (2003). *From Gutenberg to the Global Information Infrastructure: Access to Information in the Networked World* (p. 324). MIT Press.

Borgman, C. L. (2007). *Scholarship in the Digital Age* (p. 336). Cambridge, Massachusetts: MIT Press.

Borgman, C. L., Wallis, J. C., & Enyedy, N. (2007). Little science confronts the data deluge: habitat ecology, embedded sensor networks, and digital libraries. *International Journal on Digital Libraries*, *7*(1-2), 17–30. doi:10.1007/s00799-007-0022-9

Chavan, V., Sood, R. K., & Ariño, A. H. (2010). *GBIF Best Practice Guide For "Data Discovery and Publishing Strategy and Action Plans". Version 1.0.* (p. 29). Copenhagen: Global Biodiversity Information Facility.

Crawford, I. M. (1997). *Marketing Research and Information Systems*. Rome, Italy: Food and Agriculture Organization of the United Nations.

Environmental Law Institute. (2001). *New York State Biodiversity Project Needs Assessment* (pp. 1–62).

Faith, D. P., Collen, B., Ariño, A. H., Koleff, P. O., Kerr, J. T., Guinotte, J. M., & Chavan, V. (2013). Bridging the data gaps: Recommendations of the GBIF Content Needs Assessment Task Group.

Fellers, C. A. &, & Drost, G. M. (1996). Collapse of a Regional Frog fauna in the Yosemite Area of the California Sierra Nevada, USA. *Conservation Biology*, *10*, 414–425.

Flemons, P. (2011). Crowd-sourcing: perpetual valuable resource or a passing shower of dubious worth? In *TDWG 2011 Annual Conference*. Biodiversity Information Standards.

Flemons, P., & Berents, P. (2012). Image based Digitisation of Entomology Collections: Leveraging volunteers to increase digitization capacity. *ZooKeys*, *209*, 203–217. doi:10.3897/zookeys.209.3146

Flower, F. J. (2001). *Survey Research Methods*. Sage Publications Inc., ISBN: 0761921915.

Gaiji, S., Chavan, V., Ariño, A. H. A. H., Otegui, J., Hobern, D., Sood, R. K., & Robles, E. (2013). Content assessment of the primary biodiversity data published through GBIF network: Status, Challenges and Potentials. *Biodiversity Informatics*, *8*(August 2012), 94–172.

GEOCONNECTIONS. (2007). *Understanding Users' Needs and User-Centered Design* (pp. 1–68).

Greenlaw, C., & Brown-Welty, S. (2009). A comparison of web-based and paper-based survey methods: testing assumptions of survey mode and response cost. *Evaluation review*, *33*(5), 464–80. doi:10.1177/0193841X09340214

Grinnell, J. (1910). The methods and uses of a research museum. *Popular Science*, 163–169.

Guralnick, R., & Hill, A. (2009). Biodiversity informatics: automated approaches for documenting global biodiversity patterns and processes. *Bioinformatics (Oxford, England)*, *25*(4), 421–8. doi:10.1093/bioinformatics/btn659

Hill, A. W., Otegui, J., Ariño, A. H., & Guralnick, R. P. (2010). *GBIF Position Paper on Future Directions and Recommendations for Enhancing Fitness-for-Use Across the GBIF Network.* *GBIF* (p. 25). Global Biodiversity Information Facility.

InBIF. (2011). *User Need Assessment Report* (pp. 1–6).

Kelling, S., Hochachka, W. M., Fink, D., Riedewald, M., Caruana, R., Ballard, G., & Hooker, G. (2009). Data-intensive science: a new paradigm for biodiversity studies. *BioScience*, *59*(7), 613–620.

Krebs, C. J. (1999). *Ecological Methdology, 2nd. Edition*. Benjamin Cummings.

Krishtalka, L., & Humphrey, P. S. (2000). Can natural history museums capture the future? *BioScience*, *50*(7), 611–617.

Lyal, C. H. C. (2004). *Taxonomic Needs Assessments Support Pack. Natural History* (pp. 1–76). London, UK.

Meerman, J., & Clabaugh, J. (2004). *User Needs Assessment* (pp. 1–235).

Otegui, J., & Ariño, A. H. (2009). Have Standards Enhanced Biodiversity Data? Global correction and acquisition patterns. In A. L. Weitzman (Ed.), *Proceedings of TDWG* (p. 92). Biodiversity Information Standards (TDWG).

Otegui, J., Ariño, A. H., Chavan, V., & Gaiji, S. (GBIF). (2013). On the dates of the GBIF-mobilised primary biodiversity data records. *Biodiversity Informatics*, *8*(1), 173–184.

Penev, L., Lyal, C. H., Weitzman, A., Morse, D. R., King, D., Sautter, G., Georgiev, T., Morris, R.A., Catapano, T.,Agosti, D. (2011). XML schemas and mark-up practices of taxonomic literature. *ZooKeys*, (150), 89–116. doi:10.3897/zookeys.150.2213

Peterson, A. T., Knapp, S., Guralnick, R., Soberón, J., & Holder, M. T. (2010). The big questions for biodiversity informatics. *Systematics and Biodiversity*, *8*(2), 159–168. doi:10.1080/14772001003739369

Rea L. M., & Parker R. A. (2005). Designing and conducting survey research: a comprehensive guide (Jossey Boss Public Administration Series), ISBN: 078797546X.

Soberón, J., & Peterson, A. T. (2004). Biodiversity informatics: managing and applying primary biodiversity data. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, *359*(1444), 689–98. doi:10.1098/rstb.2003.1439

Soberón, J., & Peterson, A. T. (2009). Monitoring biodiversity loss with primary species-occurrence data: toward national-level indicators for the 2010 target of the convention on biological diversity. *Ambio*, *38*(1), 29–34.

Sokal, R. R., & Rohlf, F. J. (2012). *Biometry: The principles and practice of statistics in biological research. 4th edition.* (p. 937). New York: W.H. Freeman & Co.

Tann, J., Kelly, L., & Flemons, P. (2008). *Atlas of Living Australia: User Needs analysis* (pp. 1–152).

Taylor, A. (2006). *United Kingdom Taxonomic Needs Assessment* (p. 35).

Vehovar, V., Manfreda, K. L., & Berzelak, J. (2012). WebSM.org - Web Survey Methodology. Retrieved November 21, 2012, from http://www.websm.org/

Walonick, D. S. (2003). *Survival Statistics* (p. 131). StatPac. Witkin, B. R., & Altschuld, J. W. (1995). *Planning and conducting Needs Assessments: A Practical Guide*.

Yesson, C., Brewer, P. W., Sutton, T., Caithness, N., Pahwa, J. S., Burgess, M., Gray, W.A., White, A.,R.J. Jones, Andrew, C., Bisby, F.A., Culham, A. (2007). How Global Is the Global Biodiversity Information Facility? *PLoS ONE*, *2*(11), e1124. doi:10.1371/journal.pone.0001124

# Appendix 1: Glossary of Terms

**Biodiversity:** "the variability amongst living organisms from all sources including, *inter alia,* terrestrial, marine and other aquatic ecosystems and the ecological complexes of which they are part; this includes diversity within species, between species and of ecosystems." (CBD definition)

**Citation (or data citation):** a process in which a data publisher can be formally acknowledged and cited as the creator of the data.

**Data publishing:** a process through which biodiversity datasets are made freely and openly available in standardized formats.

**Darwin Core**: an internationally standardized set of terms for describing the identity and occurrence of organisms.

**Darwin Core Archive:** a standardized format in which data must be presented in order to publish it through the GBIF infrastructure (also see Special Notes, below)

**Ecosystem**: a collection of living organisms, the interactions between them and with their physical environment.

**Ecosystem services**: the benefits that people obtain from ecosystems.

**Fitness for use (use when describing data):** the suitability, effectiveness or usefulness of data in delivering accurate, authenticated, replicable and scientifically valid data for analysis and forecasting in conservation and management of natural resources.

**Local government or local authority**: an administrative unit of government responsible for an area that is smaller than a state or province

**Metadata:** information (data) about a dataset

**Primary biodiversity data:** digital text or multimedia data records documenting the occurrence of organisms