GLOBAL
BIODIVERSITY
INFORMATION
FACILITY

Best practice guide for

‘Data Discovery and Publishing Strategy and Action Plans’

October 2010

www.gbif.org

Suggested citation:

GBIF. 2010. Best practice guide for 'Data Discovery and Publishing Strategy and Action Plans' version 1.0. Authored by Chavan, V. S., Sood, R. K., and A. H. Arino. 2010. Copenhagen: Global Biodiversity Information Facility, 29 pp. ISBN: 87-92020-12-7. Accessible online at http://www.gbif.org.

Cover design: Ciprian Marius Vizitiu

ISBN: 87-92020-12-7

Contents

1. Data Discovery & Publishing Strategy and Action Plans: Why?

The availability of good quality data is vital to resolve certain key issues related to biodiversity conservation and use: among others, food security, invasive species, control of disease vectors, marine productivity, etc. This relies upon easy discovery and enhanced accessibility of primary biodiversity data[1] to anyone, anytime, anyplace. This calls for prioritised, determined, and persistent efforts by all relevant stakeholders in expediting the discovery, digitisation, and publishing of primary biodiversity data. This is even more necessary as Article 17 of the Convention on Biological Diversity[2] (CBD) expects contracting parties to *"facilitate the exchange of information, from all publicly available sources, relevant to the conservation and sustainable use of biological diversity, taking into account the special needs of developing countries"*. The countries who are not party to binding international treaties should also follow the same spirit as highlighted in Article 17 by making biodiversity data publicly available for all.

There is a tremendous need for data at both national and international levels to address these key issues. In order to meet such huge requirements, it is essential for countries to have a biodiversity data discovery and mobilisation strategy in alignment with their overall biodiversity strategy and action plan. Formulating such a strategy requires extensive planning, teamwork, and participation of all concerned parties at national, thematic and global scales.

Global biodiversity data and information are necessary to support well-informed decision making at the global level. However, existing data discovery and publishing efforts are often amateur and opportunistic in nature, aimed at tapping low hanging fruits. Data discovery and publishing strategy has a vital role in delivering the data required to implement the national or institutional biodiversity action plan. Without such a strategy it is hard to mobilise available data resources systematically. This strategy is also important in designing the work plan for new data collection and dissemination. Furthermore, drawing investment, socio-political support and recognition for data discovery and publishing would be made possible through comprehensive strategy and action plan development.

[1] Primary biodiversity data is defined as digital text or multimedia data record detailing facts about instance of occurrence of an organism, i.e. on the what, where, when, how and by whom of the occurrence and the recordings.

[2] http://www.cbd.int/convention/articles.shtml?a=cbd-17

2. Data Discovery & Publishing Strategy and Action Plans: Status

It is a well known fact that data and information critical to biodiversity conservation and resource management related decisions are not readily available. Part of the problem is associated with the complex nature of biodiversity data. In addition, global biodiversity data are available in diverse formats and resolutions. Biodiversity data is scattered and is held by different individuals, organisations and institutions. In many cases data are either incomplete, inaccessible, or both. Currently, it is a major challenge for users to discover, access, and use available information leading to an understanding of the biological basis for biodiversity conservation and planning.

The current progress in global biodiversity data discovery and mobilisation is linear, geographically uneven, and opportunistic. The progress is generally "within the comfort zones" of the data custodians and publishers. There are few, if any, specific demand-driven and deterministic data discovery and mobilisation strategies amongst data publishers. Such a lack of strategies leads to lack of action plans that can provide useful answers to stakeholder communities. Therefore, there is a need for demand-driven discovery and mobilisation strategies.

3. Data discovery & publishing strategy and action plans: Components

The purpose of any strategy is to direct action towards a desired outcome. For the objective of this document the purpose of the strategy will be to facilitate discovery and accessibility to optimum data (quality as well as quantity) leading to informed decision making and sustainable use of biotic resources. This means the strategy should be addressing the needs of target audiences and key stakeholder communities, as well as responsible actors. Chapman, 2008[3] has provided an exhaustive list of users and usage of biodiversity data. Major responsible actors whose work processes will be affected by such a strategy include data originators/collectors, data managers, and data publishers, science funding agencies and all players involved in various stages of the data life cycle. Such a

[3] Chapman A. 2005. Uses of Primary Species Occurrence Data, version 1.0. Copenhagen: Global Biodiversity Information Facility. 106 pp. ISBN: 87-92020-01-1 (available as part of GBIF Training Manual 1: Digitisation of Natural Histiry Collections Data, ISBN: 87-92020-07-0, accessible at http://www.gbif.org/communications/resources/print-and-online-resources/online-publications/gbif-training-manual-1-digitisation-of-natural-history-collections-data/)

strategy has to be resource management, user demand, conservation and science driven and should provide approaches that would alleviate the community's responsibilities' and functions. An important goal of the strategy is to bring in clarity about roles, responsibilities and expectations of relevant stakeholders. It should also provide insight into the impact of a strategy at various degrees of implementation.

As depicted in Figure 1, the data discovery & publishing strategy and action plan consists of six components:

1. Content Needs Assessment (CNA)
2. Data Gap Analysis
3. Data discovery & publishing strategy
4. Action plans and business proposals
5. Resources mobilisation and implementation
6. Performance evaluation and monitoring

Figure 1: Components of the Data Discovery & Publishing Strategy and Action Plan

Such a strategy needs to be developed and implemented at all levels – local to global scale, e.g. institutional, regional/thematic, national and global. However, granularity and detail of the scope will often change at every level. Strategies at local scale (Institutional) will focus on specific actions while strategies at national and global level will often provide broad guidelines/best practices. Further, a comprehensive and complete national

and global strategy will be a reality only if it is based on numerous local, regional/thematic strategies.

In the subsequent sections, the adoption and implementation of each component will be discussed in detail.

4. Content Needs Assessment: Why & How?

The objective of Content Needs Assessment (CNA) is to get a first-hand idea about the user needs of the biodiversity data. CNA should examine the extent and adequacy of biodiversity data and information currently being generated and accessible from the point of view of decision makers. It should also identify impediments to the use of such information and suggest ways to design formats for biodiversity information to increase their accessibility to decision makers.

It is important to conduct CNA of all representative users of biodiversity information, including policy-makers, land managers, planners, business and industry representatives, scientific and international organisations. It is crucial to evaluate the quality, quantity, and type of biodiversity data that is accessible to decision makers, natural resources managers, and conservation agencies. CNA should also investigate the ways and means in which biodiversity data and information is generated and made accessible to cross sectional users as it will influence usability and applicability of such data in decision making. Decision makers from local governments, private industry and conservation organisations must be included in the CNA exercise.

4.1 Content Needs Assessment: How?

The CNA needs to be done at all levels within Participant networks (local, national, thematic, regional and global). As depicted in Figure 2, the ideal CNA will involve the following six steps.

a) Determining Purpose and Objectives
b) Identification of Target Audience
c) Methods of CNA Exercise
d) Design the survey/questionnaire
e) Collection, Analysis & interpretation
f) Dissemination and follow-up actions

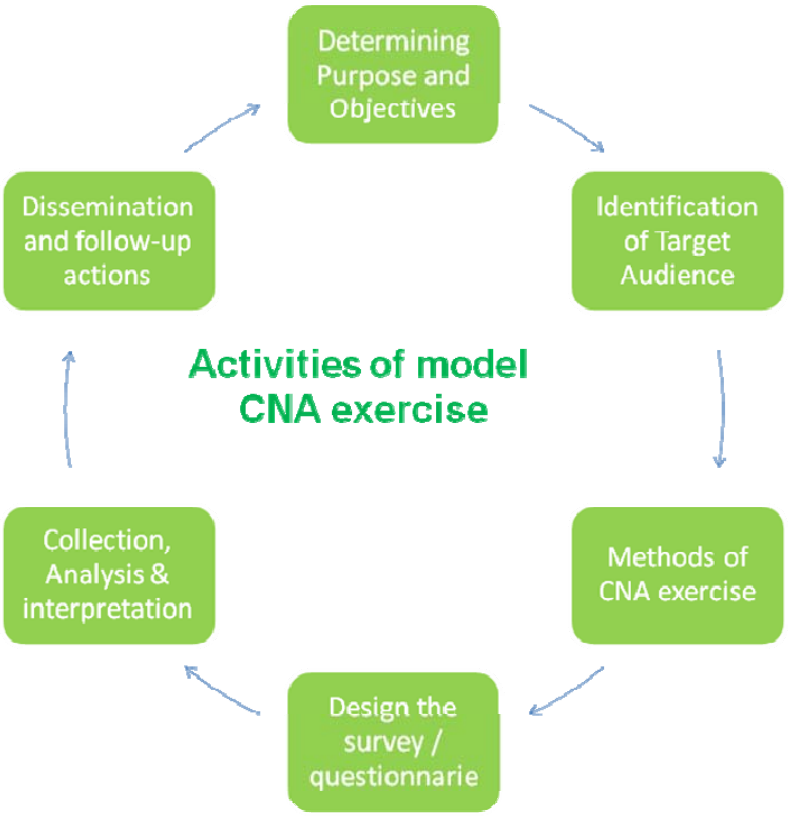

Figure 2: Activities of model CNA exercise.

a) **Determining purpose and objective**: This important step of determining the purpose and objective of the CNA exercise will help in having a clear understanding of what sort of information data publishers, and/or biodiversity information networks, want to collect (through such CNA exercise) and why. This is critical for achieving anticipated results or outcome from the CNA exercise. For instance, understanding user needs of primary biodiversity data was the objective of a GBIF-conducted CNA survey[4]. Needless to say that purpose and objectives of each CNA exercise will vary depending on thematic, geographic scale of the exercise and its target audience.

b) **Identification of target audience**: Determination of the target audience that will participate in the study is very important. The selection of the target audience depends upon which questions and at what geographical and thematic scale data publishers, and/or biodiversity information networks wish to address. For instance, if the objective of the CNA exercise is to determine user needs for better management

[4] http://www.surveymonkey.com/s.aspx?sm=XWXP0hKQSbFwb4fgn5uZ%2bQ%3d%3d

of biotic resources in a given protected area, then the target audience of such an exercise would be protected area managers, policy makers, local, state and federal administrators, biodiversity research institutions, non-governmental organisations, citizens from fringe areas, etc. This list can be expanded as necessary.

c) **Methods of CNA exercise**: For the success of the CNA exercise it is critical to decide which method(s) or approaches will be employed. The choice of these methods depends on purpose and objective, coverage of target audience, and granularity of answers sought through the exercise. Some of the methods used to conduct such a study are: (a) Surveys (online & offline), (b) Interviews (in-person or remote), (c) literature survey and analysis (scholarly, gray and popular media), and (d) Brainstorming sessions. The last one includes workshops of representatives of stakeholder communities, and/or public hearings with the community itself.

   In the recent past, GBIF Secretariat and its various Task Groups have employed a combination of approaches including online surveys[5], interviews, literature survey, and focused brainstorming sessions[6] for its CNA exercise. The choice of language is very critical for seeking adequate feedback to a CNA exercise. Given that the stakeholders community in biodiversity is globally spread, and includes a significant non-English speaking population, CNA exercises at national, regional and global scales should be conducted in more than one language. A recent survey conducted by the GBIF Content Needs Assessment Task Group was conducted in five languages[7],[8], viz. English, Spanish, French, Russian and Chinese. It is interesting to note that the return from this exercise from non-English speakers was vastly higher than a similar exercise conducted in English only[9].

Appendix -I provides some sample questions to conduct content needs assessment surveys. Users are encouraged to modify these questions or add additional questions to meet their requirements.

[5] http://www.gbif.org/communications/news-and-events/showsingle/article/gbif-content-needs-assessment-survey-2009/
[6] http://www.gbif.org/communications/news-and-events/showsingle/article/gbrds-stakeholders-planning-workshop/
[7] http://www.gbif.org/communications/news-and-events/showsingle/article/gbif-content-needs-assessment-survey-2009/
[8] http://www.gbif.org/informatics/primary-data/task-groups/cna-tg/
[9] GBIF (2010) Report of the GBIF Task Group on Global Strategy and Action Plan for Mobilisation of the Natural History Collections Data (GSAP-NHC TG): Annex 2: Selected highlights of the GSAP-NHC TG Survey, Global Biodiversity Information Facility Secretariat, Copenhagen, Denmark, pp. 104.

d) Designing of the survey/questionnaire: Design of the survey or questionnaire is the most critical aspect to extract accurate information or facts from the stakeholder communities, irrespective of which method or approach is employed for the CNA exercise. However, from our experience of both conducting such surveys, and overseeing the CNA exercises, this is the most neglected or often rushed activity of the entire exercise. All too often survey questions are 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 guide books[10,11] available which can help in designing a productive survey.

For the purpose of the 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 is 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 2009.

e) Collection, Analysis & Interpretation: Collecting survey data and organising it based on key categories is important. Use data summary sheets to help determine patterns in the data collected through survey. Most online survey tools can help organise information based on each question, making analysis of survey data easier. However, an online survey also comes with certain constraints as explained in the footnote below[12]. The proper analysis of survey results is important as it helps determine the

[10] Rea and Parker (2005). Designing and conducting survey research: a comprehensive guide (Jossey Boss Public Administration Series) , ISBN: 078797546X, pp. 304.
[11] Flower, F. J. (2001), Survey Research Methods. Sage Publications Inc., ISBN: 0761921915, pp. 192.
[12] Using summary data sheets is indeed very useful to determine patterns. However, two of our exercises have shown that online survey tools may not be offering the right set of organizing tools. Rather, they seem intent in making the data available in some sort of compromise between a compact and an exhaustive form, or to offer just the most basic (and often not easily reworkable) summary results. But getting the summary sheets as really needed to detect patterns implies (or at least has required in our exercises) a complete reorganization of the data, in our case in the form of databases that could be queried. . Interestingly, a particularly painful limit in our exercises was that the output from the online tools was arranged as a set of horizontally-split case-by-variable-option tables largely exceeding the spreadsheet column limits. This thoroughly impeded many of the analytics we used until all data were rearranged in a more standardized, databased format, holding one record for each respondent-variable-option. Databased formats are inherently more complex to arrange, but also more flexible and more uniform for querying; are not limited, and can be managed easier.

strengths and weaknesses of the survey, and can be greatly facilitated if the results format follows a record-format that can be queried rather than a fixed, tabular format that is often limited to the most basic queries. It can also help in deciding if there is a need to contact some of the respondents again. Data gathered during the survey should be analysed against the purpose to determine the effectiveness of the survey. Use of best practice guidelines[13],[14],[15],[16] in analysis and interpretations is highly recommended to achieve the expected outcome through the CNA exercise.

f) Dissemination and Follow-up Actions: Being at the end of the spectrum of CNA exercise, these activities (especially, dissemination) often do not receive deserved attention by the organisers. While follow-up to some degree is always treated as a natural fall-out of the exercise, the dissemination of results needs to be an integral part of the communication strategy of the organisation conducting the survey. We rather feel that a broader dissemination of results and follow-up actions helps in generating good will and increasing support to underlining the cause which resulted in the CNA exercise. The dissemination of survey results to all stakeholders helps in preparing the proper action plan. It also helps in evaluating if the study was successful or not. Based on the results, the development of a clear work plan (or follow-up activities) highlighting (i) short and long term vision, (ii) work plan with timelines, and roles & responsibilities, (iii) plans for allocation of additional resources required, and (iv) performance evaluation and auditing mechanisms to be put in place, should be mandatory.

5. Data Gap analysis

Having understood the content needs of the stakeholder communities through CNA exercises, Data Gap analysis is the next logical and crucial step forward in ensuring that data needs of the key stakeholder and user communities are satisfied in a timely manner. In fact, 'Data Gap Analysis' 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

[13] Chambers and Skiners (2005). Analysis of survey data: Wiley series in survey methodology. John Wiley & Sons Ltd., ISBN: 0471899879, pp. 376.
[14] Fink A. (1995). How to analyse survey data. Sage Publications Inc., pp. 101.
[15] Lee and Forthofer (2006). Analysing complex survey data. Sage Publications Inc., pp. 91.
[16] Fink A. (2003). How to manage, analyze and interpret survey data (2nd edition). Sage Publications Inc., pp. 141.

discrepancies between current and ideal states[17] of the entire enterprise of biodiversity data management leading up to its publishing and usage.

As expected, a 'Data Gap Analysis' exercise will help to identify both the need and accessibility to data essential for making the best possible decisions about sustainable resources management, and sound and informed decisions about conservation strategies. It further identifies data needs and focuses programmes that aim to strengthen capacity to address the gaps in representation, ecology and management[18].

As depicted in Figure 3, the major steps of model 'Data Gap Analysis' include (1) scoping the analysis and expectation setting, (2) assessing the universe of accessible data, (3) data vs needs analysis, (4) identification of data gaps, (5) prioritisation of demand-driven data discovery and publishing activities, and (6) evaluation of the 'Data Gap Analysis' exercise and strategies for future analysis.

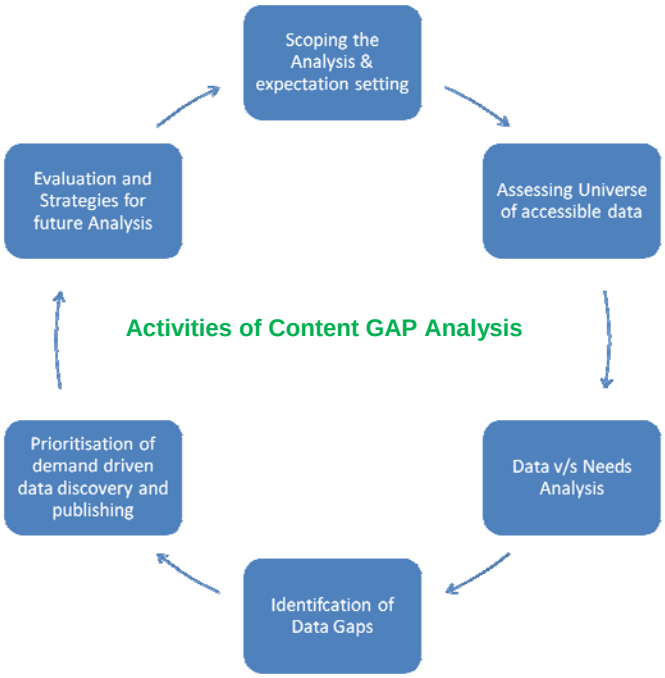

Figure-3: Workflow to identify gaps in data

Before we begin to elaborate further on each of the these steps involved in the model 'Data Gap Analysis' exercise, it is essential to dispel the prevailing misconception among many involved in the business of biodiversity data management that the scope of 'Data Gap Analysis' is limited to mapping between (a) accessible data and (b) users needs of

[17] Research Data Strategy Working Group (2008). Stewardship of research data in Canada: A gap analysis, accessible at http://data-donnees.gc.ca/docs/GapAnalysis.pdf.
[18] http://www.protectedareas.info/upload/document/gapanalysis-introduction.pdf

data. On the contrary, 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 the 'Data Gap Analysis' exercise if they are not to be included in post exercise introspection.

5.1 Scoping the 'Data Gap Analysis' & Expectation Setting: 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 do 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[19]. 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 the expectation of stakeholder community.

5.2 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 informs about (a) who has what type of data, (b) in what form or format, (c) in what state of digitisation, (d) if digitised, 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 inventorised. Given that such an inventorisation of data custodians & publishers is useful in data mobilisation and publishing activities as well, we recommend 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. Completeness of metadata documents that constitute such a catalogue will determine the degree of success of the remainder of activities of the 'Data Gap Analysis' exercise.

[19] http://www.protectedareas.info/upload/document/gapanalysis-introduction.pdf

5.3 Data accessibility v/s Data Needs Mapping: Now that we have (a) needs of the user community as identified through the 'Content Needs Assessment' (as described in Section 4) and (b) understanding of 'estimate of the universe', the next logical step is to map them against each other in an attempt to answer the question whether needs expressed by the user community(ies) are being met through accessible data or not. Answers to this question will determine the further step towards 'demand-driven' and 'deterministic data discovery and publishing strategies and action plan development.

We feel that such mapping is a continuous process, and that it needs to be conducted at regular intervals. One of the major reasons for this is the incomplete and in-progress state of biodiversity studies in different regions of the world, which continually generate new data. As new data become available, such a mapping exercise can provide realistic insights of the effort required to bridge the data gaps to address pre-determined issues or data needs of a particular stakeholder community.

5.4 Identification of Data Gaps: 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 prioritised activities ranging from collections of data to its publishing, resulting in free and open access to data.

5.5 Prioritisation of Demand Driven Data Discovery and Publishing: It is natural that the 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 lists of user communities cannot be met at the same time. This may result from a 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.

Therefore, the prioritisation of demands for data by the stakeholder community is essential, as to which demands need to be addressed first and which can be taken up at a later date. The criteria for such prioritisation differ depending upon the gravity of user demands, types of data requirements (quantity and quality), geographic, ecosystem, and thematic scope of the demands for the data, etc.

5.6 Evaluation and Strategies for future 'Data Gap Analysis': 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 bench mark for future 'Data Gap Analysis' studies. This calls for a closer evaluation of every 'Data Gap Analysis' to understand what were the positive or negative aspects, what was missing, and how it can be rectified in the next exercise. This will help to build strategies and approaches, making 'Data Gap Analysis' a productive exercise, leading to demand-driven and deterministic data discovery and publishing initiatives.

6. Data Resources Discovery System (DRDS): Why and How?

The discovery of biodiversity data resources is essential for the accessibility of data. From the users' perspective, a data resources discovery mechanism is a useful first step to ensure access to adequate, appropriate, authentic, current or up-to-date, fit-for-use data. However, from the data custodian or publishers point of view, it is a mechanism or platform to publicise its expertise, credentials, and establish reputation in its area of enterprise by attracting increased investment, collaborations, usage of its data products, due credits, recognitions as well as social, political and financial support. Furthermore, discovering resources is a great tool to exemplify the tax payer's investment in the organisation's mission. Data discovery should act as a decision support mechanism for data users to make decisions as to which data resources are useful for their analysis and interpretation activities.

The foregoing discussion is compelling enough to emphasise that a data resources discovery mechanism or platform at the individual researcher, institution, but most importantly at the national and global scale, is essential. However, such a mechanism or platform at those levels is currently not available. Even if a form of mechanism is present in some thematic areas (e.g. Biodiversity Collections Index[20] for discovery of natural history collections) or for isolated regions, it is inadequate to have real-time, seamless and hassle-free discovery of registered data resources. On the contrary, existing efforts

[20] http://www.biodiversitycollectionsindex.org

towards 'biodiversity information infrastructure' development are more focused or aimed at 'data accessibility'. This often results in making only low hanging fruits accessible (Berents et.al, 2010)[21]. This paradigm or trend in biodiversity informatics needs to be changed from the local to global scale to have data discovery as the first priority of any enterprise dealing with a data life cycle. This calls for a data resources discovery strategy and action plan at all levels. Such a strategy and action plan should lead to the establishment of 'Data Resources Discovery System (DRDS)'.

Therefore, as indicated earlier, DRDS should be aimed at facilitating the registration of data resources by their custodians and publishers, whereas for users it should be a hassle-free, seamless, easy-to-use discovery mechanism. We further believe that driving forces such as obligations towards funding agency, citizens' right to information, and quest for recognition and credit, as well as the increasing demand for transparency, authenticity, and reliability of scientific outcomes are reasons that are compelling enough for every institution to act towards fulfilling these obligations/demands.

6.1 DRDS: Unique Characteristics

Data resources discovery system (DRDS) is hardly a new or innovative approach. In the past there were discovery systems in the form of catalogues or directories. However, today's DRDS will have to be dynamic, always work in progress, and real-time data resources registration and discovery system. Further, it can be interactive, facilitating two-way exchanges between data owners/custodians/publishers and prospective users.

Unlike systems of the past, technology allows us to build a web of discovery systems (systems of systems) whereby each institution, nation, or thematic network establishes its own DRDS, which can communicate with each other through a community-agreed protocol facilitating data resources discovery to potential users at anytime, anyplace, irrespective of where and by whom a data resource is registered.

Since its inception nine years ago, the GBIF network has realised that one of the major challenges of existing biodiversity informatics infrastructure is to provide an innovative means to the discovery and access of all relevant information and data resources. In fact,

[21] Berents, P., Hamer, M., and V. Chavan. 2010. Towards demand-driven publishing: Approaches to the prioritization of digitization of natural history collections data. Biodiversity Informatics, 7: 113-119.

as indicated earlier, our current ability to discover distributed, isolated and unknown data and information resources is limited. GBIF is currently addressing this challenge through the development of a Global Biodiversity Resources Discovery System (GBRDS) for the registration and discovery of biodiversity information and data resources and services.

6.2 DRDS: Suggested activities

With the growing realisation of the significance of discovery system(s), many national networks and institutions are willing to establish DRDS that will cater to their stakeholder communities. However, these agencies often require guidance on how to go about establishing such a DRDS. In our opinion, activities as depicted in Figure 4 will lead to the establishment of a DRDS to better serve the data resources discovery demands of stakeholder communities.

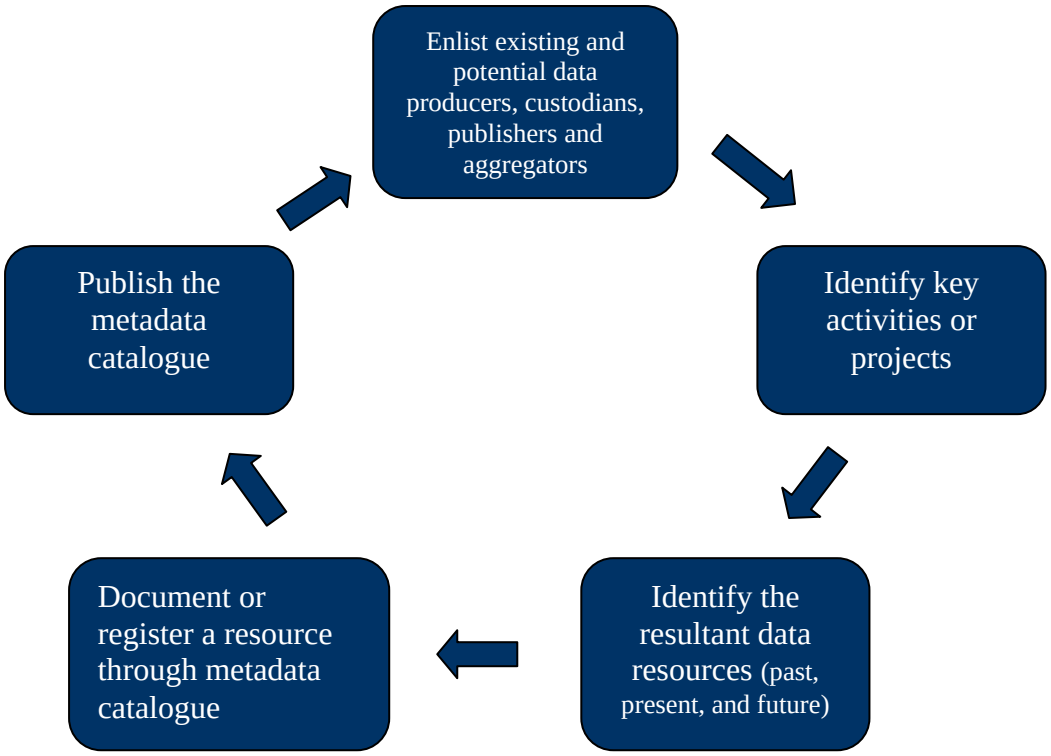

Figure 4: Activities of Data Resources Discovery System (DRDS)

These activities include (a) Listing of potential data producers, custodians, publishers and aggregators, (b) Identification of key activities and projects of these producers,

custodians, publishers and aggregators, (c) Identification of data resources, (d) Documentation, register or description of the resources through metadata catalogues, and (e) Publishing the metadata catalogue. We wish to emphasise the significance of metadata cataloguing in this entire chain of activities. The principal usage scenarios for which this metadata catalogue should be designed are data discovery, human interpretation and analytical reuse of high quality 'primary biodiversity data' for science based management of natural resources. Needless to say the data resources described through such a metadata catalogue will cover diverse scientific areas such as species distribution and abundance, measurements of characteristics of organisms, physiology, ecological processes, behaviour, experimental data, and others, and are likely to have many unforeseen uses in the future.

In this document, we will not deal with the technical nitty-gritty's of how to establish such a DRDS. There are several resources and best practice guidelines available on this topic which elaborate on issues ranging from – building of lists of agencies, their projects, and data resources to the development, enrichment, quality control and publishing of metadata catalogue, etc.

7. Data Mobilisation & Publishing Strategy and Action Plans

The last, but most critical and toughest, phase of the data discovery and publishing chain for any institution, network or nation is the development and implementation of a 'Data Discovery & Publishing Strategy and Action Plan'. In fact the success of an institution(s) or individual(s) research management venture largely depends upon ambitious yet realistic data discovery, mobilisation and publishing strategy and action plans together with their stringent implementation. However, from our experience and interactions with several agencies, institutions, and players involved in this business, we may conclude that it is conceivable that some actors might perhaps have had a limited understanding of its significance or impacts of its implementation.

In this section, we will focus our attention on strategies and action plans dealing specifically with data mobilisation and publishing. As depicted in Figure 5, there are five components to such a 'Data Mobilisation & Publishing Strategy and Actions Plan', which determine organisational success. These are (1) Purposing mobilisation and publishing, (2) Develop demand-driven and deterministic Action plans, (3) Develop Business proposal, (4) Mobilise resources, and (5) Implementation and Evaluation.

7.1 Purposing mobilisation and publishing: The reasons as to why a purposing exercise is essential have been discussed in Section 1 of this document. However, we wish to insist that data mobilisation and publishing cannot be a standalone activity all the time. Data mobilisation and publishing are a means to an end, and therefore it is essential to determine the purpose for such an end. Purposing data mobilisation and discovery will help to ensure the availability of resources leading to the completion of the task along a pre-determined timeline, as well as enhancing the scope for use of data by stakeholder communities. This also raises the profile of data mobilisation and publishing activity from being add-on and opportunistic to demand-driven and deterministic.

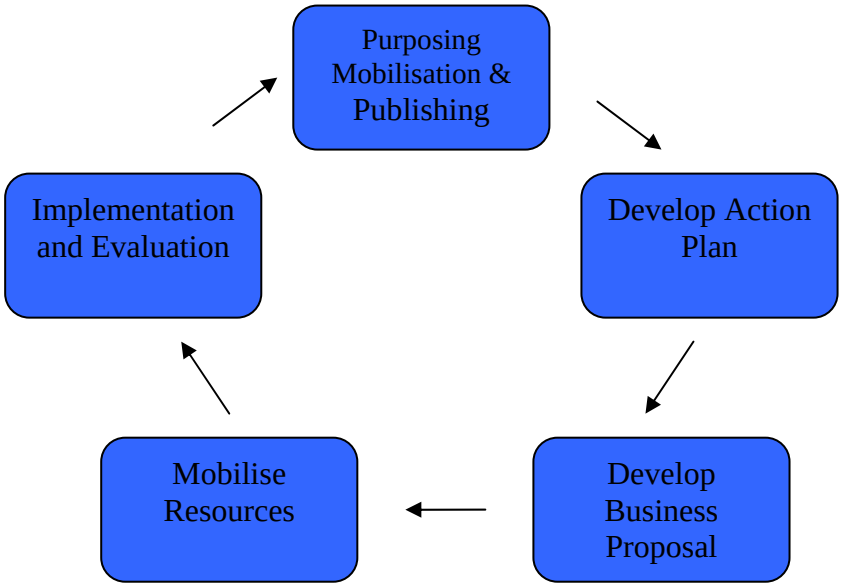

Figure 5. Components of Data Mobilisation & Publishing Strategy and Action Plan

7.2 Develop demand-driven and deterministic Action plans: Once the purpose of the mobilisation and publishing activity is zeroed in, the next logical step is to plan as to how objectives will be achieved in a timely manner. This has been adequately dealt with by Frazier et.al. (2008)[22]. Based on data digitisation requirements, an action plan must be made. Such an action plan can then be tied up with business proposals to explore the funding opportunities. Proper implementation of an action plan is essential. Follow up of an action plan is required to ensure proper implementation and evaluation of the

[22] Frazier, C. K., Wall, J., and S. Grant. 2008. Initiating a natural history collections digitisation project, version 1.0. Copenhagen: Global Biodiversity Information Facility. 75 pp. ISBN: 87-92020-05-4 (available as a standalone PDF from http://www.gbif.org).

outcomes of the data digitisation work. The Action plan should match the goals set out in the business case. It is not a bad idea to note down the types of actions required for each goal set out in the business case. As more than one method is available to achieve set goals, validating various methods and selecting the one which is likely to be most successful seems a sound tactic. One can check the selected method against the following points to ensure smooth completion of the project:

1. Does your chosen solution match goals, limitations and resources?
2. Will the solution handle future requirements and what if it does not?
3. How many staff are required?
4. How much time will it take to implement the project?
5. How much will it cost?
6. What will be the workflow?
7. What sort of action is required once the project is over?

The action plan details how the business case should actually be implemented. It contains practical information, such as what are the infrastructural requirements. It also considers the number of staff; training and how the work will proceed (commonly referred to as workflow).

Risk analysis documentation is a part of the action plan which aims to consider what to do if something goes wrong. Simple examples include what will happen if a mission-critical online database breaks or if funding is not secured for part of the project. Consideration also goes into how to minimise the risk of an event happening. Regularly backing up the data, seeking alternative funding, and having a spare server available are all simple ways of risk mitigation considered in the risk analysis documents.

It is possible that the business case outlines an overall goal that is too large to be completed in any one project and so is broken down into several smaller projects with their own business cases, action plans and risk analyses. This is a perfectly acceptable practice and the action plan for the overall business case should then outline the separate projects and how they link together to provide the overall goal. Working in this way allows large and often long term goals to be achieved in small stages without losing sight of the overall vision.

7.3 Develop Business proposal

The business case sets out what you wish to do and establishes the benefits you expect to gain from undertaking the suggested work. It also includes an assessment of the resources required to implement the project as well as identifying which resources are currently available. Any shortfall in resources should be clearly identified and the associated costs be stated. Setting these facts out in a single document allows a clear judgment to be made on the feasibility of the project.

Putting together a business case enables an institution or individual to clearly set out goals and limitations in a clear fashion. It is important to get all stakeholders involved in the project. This often means that they will be willing to help out during the implementation phase of the project. While writing proposals for such projects, it is important to go beyond the simple statement of goals and limitations to answer practical questions such as:

1. What do you/the community gain from doing this project?
2. Is the project feasible?
3. Do your goals exceed your limitations?

7.4 Mobilise resources

To ensure successful completion of the project, resources (financial, human-resource, infrastructure, political and administrative support, etc.) are essential. For instance, all projects have a cost, whether it is met by the institution or by external bodies. Finding suitable funding is something this paper cannot practically discuss in detail, as available bodies vary by country and by the exact nature of the work being undertaken. Properly building up the business case and action plan can only enhance your chances of writing a successful project proposal. Another key component is mobilising partner organisations to share relevant roles and responsibilities in the project.

7.5 Implementation and Evaluation:

The last but most crucial aspect which decides timely completion and better cost-benefit outcome is implementation mechanism and frequent evaluation of progress being made by the project.

8.0 Exemplary case study: Atlas of Living Australia

In the recent past some of the GBIF Participants carried out exercises leading to development of 'Data Discovery and Publishing Strategy & Action Plans'. Here we present the exercise carried out by Australia (through Atlas of Living Australia[23]) as an exemplary case study and lessons learnt.

8.1 Data Discovery & Publishing Strategy and Action Plan exercise undertaken by the Atlas of Living Australia

The 'Data discovery and Publishing Strategy & Action Plan' exercise carried out by the Atlas of Living Australia had two major steps; viz. user needs analysis and use case prioritisation.

A. User Needs Analysis:

The ALA User Needs Analysis is a review of the users and uses for biodiversity information in Australia. The goal of the study was to develop a clear set of use cases for building and maintaining the Atlas of Living Australia (ALA) by discovering how different users and organisations locate and use biodiversity data. The ALA has used the resulting understanding of user requirements and difficulties and of the workflows underlying key tasks to guide project priorities and to help to make data more accessible and relevant.

[23] http://www.ala.org.au/

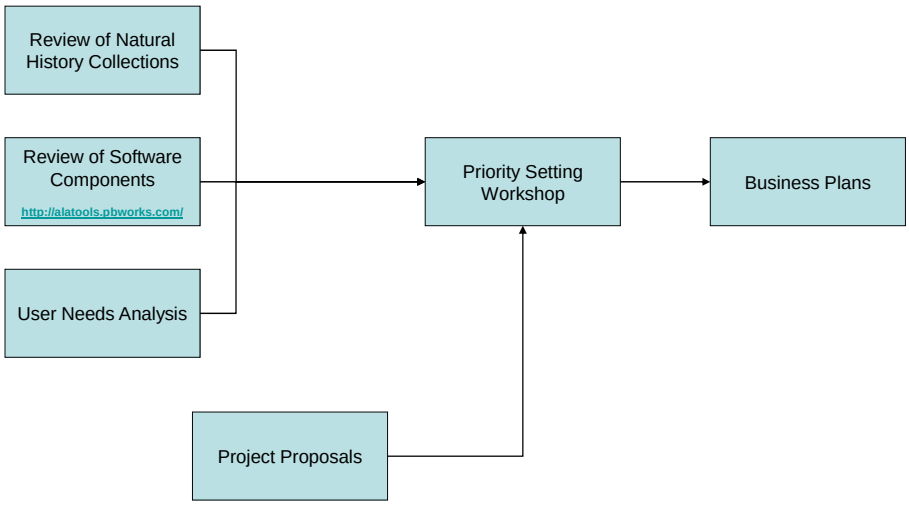

Figure 6. Atlas of Living Australia (ALA) Data Discovery & Data Publishing Strategy and Action Plan Process. URL for Review of the software component is http://alatools.pbworks.com.

Web link of detail ALA user needs analysis report is: http://www.ala.org.au/wp-content/uploads/ALAUserNeedsAnalysisReportExtract.pdf

B. Use case prioritisation:

The ALA organised a workshop to provide recommendations to the ALA Management Committee on a set of candidate projects which had been identified as possible focus areas for the ALA. The aim of this workshop was to set the agenda for the next two years of ALA activity to ensure that the ALA not only delivers core general-purpose infrastructure for managing biodiversity data but also provides solutions which address needs and even change work practices for significant user groups.

The panel comprised a spread of experts from across Australia with backgrounds in taxonomy, collections, field ecology and conservation and with knowledge of marine and terrestrial ecosystems and of vertebrate, invertebrate, plant, fungal and microbial organisms. The panel was provided with background information on the ALA, its planned core deliverables, the results of the ALA User Needs Analysis, existing activities in which the ALA or its partners are already engaged, and eleven candidate projects for consideration. The panel was tasked with ranking these eleven projects to identify those which offer the most significant benefits for users and the highest impact for ALA infrastructure development and for content shared by ALA partners.

The following potential ALA projects were considered and discussed by the panel. The projects were awarded a rank (from 1-highest priority to 8-lowest priority).

1. Extended Name Services (Rank 1)
2. ABIN biosecurity (Rank 2)
3. NRS (Rank-2)
4. NatureNet Australia (Rank 3)
5. Barcode of Life (Rank 4)
6. Image Libraries (Rank 4-5)
7. Habitat Information (Rank 5)
8. Murray-Darling Basin (Rank 7)
9. Biodiversity Conservation Strategy (Rank 7-8)
10. Images of protologues linked to APNI (Rank 7-8)
11. Fishes of Australia Online (Rank 8)

The panel ranked the candidate projects against each other to identify their relative priorities as focus areas to the ALA. There was strong agreement around the final ranking of all projects.

C. Outcome

This exercise led to identification of (a) major tasks of importance to users, (b) areas of significance for users, and (c) common subjects for importance to users.

(a) Identified major tasks of importance to users

- Distribution analysis – determining or applying the likely range for any given species
- Identification – determining the name or taxonomic group for a particular organism
- Site Assessment – reporting the list of species known, or expected to occur at a particular site
- Habitat management planning – how to best manage an area for conservation
- Managing references – maintaining a database or collection as a current information resource
- Community engagement – producing materials to educate the public
- Fact-finding – general research to find out information for any species

- Synecology / food-web analysis – exploring the interactions and dependencies between organisms
- Biosecurity – understanding introduced organisms, wildlife diseases and biological control

(b) Areas of significance for users

- Amateur observations and ad hoc data – how best to assist and encourage the capture of observational data from amateur naturalists and other independent specialists, and manage issues of quality
- Sensitive data – how to manage the many forms of sensitive and restricted data to meet the needs of users while maintaining safeguards to the satisfaction of data providers
- Names – correct and current names are highly important. How best to deal with this lack of a well-maintained and authoritative name service which addresses the needs of the many who use biodiversity data

(c)Common subjects of importance to users

- Currency – knowing that the data they are accessing is current – particularly in relation to names data
- Accuracy – an understanding of data accuracy – particularly in relation to geography and taxonomy
- Comprehensiveness – access to complete datasets – not just portions of what was potentially available
- Validation – having some measure of validation of data – to enable judgements of data suitability
- Documentation – good documentation of each data record as well as each dataset
- Ease of access – data that is easy to access and to understand its nature
- A reliable and authoritative source – trust can only come from a reliable and authoritative source of data

Appendix-I: GBIF Content Needs Assessment Survey (2009)

1. Introduction

Objective:

The objective of this survey is to assess the user needs for primary biodiversity data. The major purpose of this exercise is to identify the gaps in biodiversity data presently accessible, and make recommendations on data mobilisation strategies to bridge the gap between data needs and data access.

2. User Profile

1. Details of the Person undertaking taking this survey:

Name:	
Organisation/Institution affiliated with:	
Street/PO Box:	
City:	
State:	
Country:	
ZIP CODE:	
Phone/Mobile:	
Email:	
Web/URL:	

2. Describe your organisation (please tick one or several options)

Academic / Educational institution:	
Research Institution:	
National Agency:	
Non Governmental Organisation (NGO):	
Intergovernmental Organisation (IGO) OR Multilateral Convention:	
Private Company:	
Individual Researcher or Naturalist (e.g. citizen scientists):	
Others (please specify):	

3. Main interest/business of your organisation (please tick one or several options)

Conservation Science (including Taxonomic Research)	
Bioproductivity / Bioprospecting (Agriculture, Fisheries, Forestry, etc.)	
Biodiversity	
Biomedical and/or Public Health	
Biotechnology	
Biosecurity	
Natural Resources Management	
Industrial / Commercial Use of Natural Resources	
Exhibition / Educational / Academic	
Others (please specify)	

3. Uses of Primary Biodiversity Data

This section of the survey is designed to understand the purpose for which 'primary biodiversity data' is used by various stakeholders.

DEFINITION: Primary biodiversity data is defined as the digital text or multimedia data record detailing the instance of an organism – or the what, where, when, how and by whom of the organisms occurrence and recording.

The uses of primary biodiversity data are wide and varied, and encompass virtually every aspect of human endeavor – food, shelter, health, recreation, art and history, society, science and politics, etc. Furthermore, such data is essential for predicting the sustainable future of our planet, and therefore of all living beings.

1. **List the ways in which you use Primary Biodiversity Data (please choose one or several options)**

Taxonomy	
Biogeographic studies	
Species diversity and populations	
Life histories and phenologies	
Endangered, migratory and invasive Species	
Impact of Climate Change	
Ecology, Evolution and Genetics	
Environmental regionalisation	
Conservation Planning	
Sustainable Use	
Natural Resources Management	
Agriculture, Fisheries, Forestry and Mining	
Nursery and Pet Industry	
Health and Public Safety	
Bioprospecting	
Forensics	
Border Control and Wildlife Trade	
Education and Public Outreach	
Ecotourism	
Art and History	
Society and Politics	
Recreation	
Human Infrastructure Planning	
Industrial Use	
Environmental Impact Management	
Others (please specify)	

4. Access to Primary Biodiversity Data

This section is to learn how users access primary biodiversity data (please choose one or several options).

The objective is to understand the mechanisms employed and the frequency for accessing primary biodiversity data.

1. How do you access primary biodiversity data?

Through your own field works/surveys:	
Through hardcopy, literature survey (non-digital form):	
Through Primary Publications (e.g. taxonomic monographs, maps of species observations):	
Through access to offline digital data sets (CDROM/DVD/Tapes etc.)	
Through the GBIF Data Portal (http://data.gbif.org)	
Through other web based data portals (please specify)	
Through FTP sites (please specify)	
Through institutional agreements	
Through Payment basis	
Through free and open datasets within and outside of your institution:	
Through Reciprocal agreements with other groups/individuals	
Through others (please specify)	

Please provide detailed examples of each option you select:

2. Frequency of access

Daily basis	
Once a month	
Once a quarter	
Bi-annual	
Cannot determine (on Need Basis)	
Others (please specify)	

5. Quality and Quantity Requirements

The intention of this section is to understand: (a) what type/nature of biodiversity data the user requires? (b) How much biodiversity data is essential for the specific purpose? And (c) how much the quality of data matters?

1. Types or Nature of Primary Biodiversity Data Required?

Taxonomic Names / Checklists	
Occurrence Records (presence only)	
Occurrence Records (including absence records)	
Population density / Dynamics data	
Species Interaction Data	
Species Information (Descriptive data)	
Others (Please specify)	

2. Quantity of data required for each data type?

	1-100 Records	**101-1000 Records**	**1001-10000 Records**	**10000+ Records**
Taxonomic names/checklists				
Occurrence records				
Population density/dynamics				
Multimedia resources				
Others (please specify)				

3. For which type of environments do you use/need more primary biodiversity data?

	Frequent Use	Less Frequent Use	Occasionally required	Not required
Marine: Coasts				
Marine: Oceans				
Marine: Deep Seas				
Marine: Islands				
Marine: Estuarine				
Inland: Wetlands				
Inland: River basin				
Inland: Lakes				
Terrestrial: Tropical Forests				
Terrestrial: Temperate Forests				
Terrestrial: Deserts				
Terrestrial: Grasslands				
Terrestrial: Agro-ecosystem				
Terrestrial: Mountains				
Others (please specify)				

4. Which data at the ecosystem level are the most required by you and at what Scale?

	Global	Regional	National	Provincial	Local
Ecoregions					
Vegetation coverage					
Protected areas					
Temperature					
Precipitation					
Soil					
Watersheds					
Basins					
Others (please specify)					

6. Species level data requirement

The objective of this section is to understand data on which taxa's is most often required.

1. Which data at the plant species level are most required by you and at what scale?

Please specify child taxa or common names in the box below

	Global	Regional	National	Provincial	Local
Plants: Monocots					
Plants: Dicots					
Plants: Bryophytes					
Plants: Pteridophytes					
Plants: Gymnosperms					
Plants: Algae					
Plants: Others (Please specify)					

2. Which data at the Animal species level are the most required by you and at what scale?

Please specify child taxa or common names in the box below

	Global	Regional	National	Provincial	Local
Phylum: Acanthocephala					
Phylum: Annelida					
Phylum: Arthropoda					
Phylum: Brachiopoda					
Phylum: Cephalorhyncha					
Phylum: Chaetognatha					
Phylum: Chordata					
Phylum: Cnidaria					
Phylum: Ctenophora					
Phylum: Cycliophora					
Phylum: Echinodermata					
Phylum: Echiura					
Phylum: Ectoprocta					
Phylum: Entoprocta					
Phylum: Gastrotricha					
Phylum: Gnathostomulida					
Phylum: Hemichordata					
Phylum: Mesozoa					
Phylum: Mollusca					
Phylum: Myxozoa					
Phylum: Nemata					
Phylum: Nemertea					
Phylum: Onychopora					
Phylum: Phoronida					
Phylum: Placozoa					
Phylum: Platyhelminthes					
Phylum: Porifera					
Phylum: Rotifera					
Phylum: Sipuncula					

Phylum: Tardigrada					
Others (please specify)					

3. Which data at the Fungi, Virus and Microbial species level are most required by you and at what scale?

Please specify child taxa or common names in the box below

	Global	Regional	National	Provincial	Local
Microbes					
Fungi					
Virus					
Others (Please specify)					

4. What are the most important characteristics that you generally want for species occurrence data?

Precise/accurate geo-referenced data	
Metadata on uncertainty about geographical/georeferenced data	
Pre-1990 data	
Post-1990 data	
Type specimens in scientific collections	
Source of information	
Images	
Synonyms of species name	
Common name of species	
Species habitat description	
Others (please specify)	