Support to Building the Inter-American Biodiversity Information Network

Trust Fund #TF-030388

GUIDELINES

on

Biodiversity Information Management

for

Donor-financed Projects

(Document 9)

April 2005

NIPPON KOEI

UNEP WCMC

Support to Building IABIN (Inter-American Biodiversity Information Network) Project

Document 9 – Guidelines on Biodiversity Information Management for Donor-financed Projects

Background

The World Bank has financed the current support work under the Japanese Consultants Trust Fund. The objective is to assist the World Bank in the completion of project preparation for the project Building IABIN (Inter-American Biodiversity Information Network), and for assistance in supervision of the project. The work undertaken covers three areas: background studies on key aspects of biodiversity informatics; direct assistance to the World Bank in project preparation; and assistance to the World Bank in project supervision. The present document is one of the background studies, although applicable beyond the context of IABIN.

The work has been carried out by Nippon Koei UK Co Ltd, in association with the UNEP World Conservation Monitoring Centre.

Table of Contents

CHAPTER 1 INTRODUCTION

1.1 Purpose of the Document

This document is aimed at World Bank Task Managers and implementing project managers and contractors dealing with projects in which biodiversity information management is a significant component. It is intended to provide guidance on good practices and applicable standards in the field of "bio-informatics" in order to assist in making such projects efficient, effective, and sustainable. These Guidelines are intended to complement rather than replace such guidance as the World Bank Task Managers' ICT Toolkit, 2003, (ICT Toolkit) by presenting the practices and standards specific to the content domain of "bio-informatics" (such as taxonomy reference archives, and biological nomenclature and classification systems), and emphasizing the areas of information technology (such as GIS) that are most relevant to biodiversity.

Although deriving from the project Building the Inter-American Biodiversity Information Network, the guidelines are meant to be applicable to any region for projects in which biodiversity information management forms part of the requirement.

1.2 Scope of Biodiversity Information

The term "biodiversity information" is difficult to define in a global context, for there is no consistent and accepted meaning. Various views as to the scope and meaning have evolved from different sectors of the environmental science community, and three differing major views have developed, as follows:

First view - Biodiversity information means taxonomy: The taxonomic community has interpreted the Convention on Biological Diversity (CBD) as support and justification for increased scientific research in their specific field. Hence the apparent view that "biodiversity information" **equals** taxonomy, even though this scientific endeavour provides only a partial picture, and is only one of many classes of information important to the conservation of biodiversity. The Global Biodiversity Information Facility (GBIF), for instance, concentrates on scientific issues in taxonomy (naming and relationships) and on specimen collections in museums and herbaria, even though its name might imply a broader scope.

Second view - Biodiversity information means species-related information: This view of the scope extends from taxonomy and museum specimens to species observational data – e.g. distribution and populations of species. This implies information on the occurrence and movement of species, their protection status, and natural habitat requirements.

Third view - Biodiversity information has broad ecological scope: Biodiversity information as implied by the Convention on Biological Diversity extends beyond species-

centric data, to include biodiversity management and ecosystems information – that would include protected areas, habitats, ecosystem condition and monitoring, conservation strategies and methodologies, population dynamics, actions towards conservation (conventions, regulations, action plans), and so on. The Convention also encompasses information related to socio-economic considerations and concepts such as "equitable sharing of benefits" and "sustainable development".

The full breadth (the "broad ecological scope") encompasses a number of major categories as follows:

Taxonomic Information

- Taxonomic reference systems and registries
- Species nomenclature and synonymy
- Species identification
- Museum, herbarium and botanic garden specimens

Species Information

- Species distribution
- Species population and dynamics
- Conservation status
- Threats
- Behaviour and habitats
- Species conservation activities (in-situ and ex-situ)
- Species "hot-spots"

Protected areas

- Location and distribution
- Purpose
- Protection status, international and national
- Management
- Relationship to species
- Ecosystem protection

Ecosystems

- Characteristics

- Distribution and dynamics
- Threats
- Status and condition
- Long term monitoring
- Relationship to species

Responses

- Conventions and treaties
- Legislation and regulation
- Strategies and policies
- Action plans and projects

These five major categories provide the core information required for effective decision-making on the range of topics relevant to biodiversity. Any particular project will have a scope of relevant information that falls somewhere in this spectrum.

In terms of circumscribing the scope of "biodiversity information", it is important to note that this rather broad definition does NOT extend as widely as "environmental information" – i.e. does not encompass information on pollution loads, renewable and non-renewable resource extraction and utilisation, and many other factors normally considered part of State-of-the-Environment reporting, although many of the same principles and issues may apply.

1.3 Characteristics of Biodiversity Information

An information systems specialist may argue that "data is data", hence all the conventional approaches to ICT development are similarly applicable, for instance as outlined in the ICT Toolkit. One also hears the counter-argument that biodiversity information is "special", requiring specialist scientific knowledge, and is not amenable to conventional systems analysis and development methodologies. There is an element of truth in both views, although it should be emphasised that **all** of the advice of the ICT Toolkit is applicable – although there are some domain-specific characteristics that need to be considered in addition.

Biodiversity information deals mainly with the observational aspects of conservation biology – a very descriptive science. The intrinsic nature of the information and the way it is customarily collected and presented make it difficult to integrate with other non-biological information due to some of the following characteristics:

- Biodiversity information is often both descriptive and subjective, rather than quantitative. Assessments of the state of ecosystems are often entirely narrative and

contain un-standardised relative terms such as “declining” “improving”, “healthy”, “fragmented”, with little or no quantitative information.

- Biodiversity information is inherently geographic (spatially-based) for example, species distributions, observational sample locations, boundaries of watersheds, eco-regions and so on, and is often presented in mapped form.
- There are few agreed standard ways to classify or typify habitats or ecosystems (or bio-geographic zones, or biomes, or vegetation cover, etc, etc). Where such classifications exist, they tend to be applicable only in a limited region.
- There is little long-term systematic monitoring of ecosystems – nor agreement on what to monitor – hence no baselines from which to measure change, or assess the impact of implemented actions.
- There is no agreed way to “value” biodiversity or to assess the “health” of an ecosystem or the state of its biodiversity, even in relative terms.

Because of these factors, biodiversity reports on ecosystems, protected areas, countries, districts, and species may contain huge amounts of information that is difficult to relate even to similar assessments of similar areas, and impossible to effectively link to non-biological information. The two related natural fields of climate change and oceanography are seemingly much more advanced in knowing what is important to measure, how it can be measured, and what are causes and effects. Arguably, this is due to the intrinsically more complex nature of biology (and/or ecology), but it also stems at least in part from the “gentleman scientist” and “natural history” roots of biodiversity.

Some significant implications for ICT projects involving biodiversity information are therefore:

- The requirement to handle non-quantitative and narrative information, and the associated need to utilise specialised vocabularies and reference sources to add semantic information and metadata.
- The frequent need to incorporate Geographic Information Systems (GIS) technology to store, interrelate and visually present the information. It is interesting to note that in the ICT Toolkit, all examples of “environment” sector projects indicate a GIS component and most in the related sector of Health, Nutrition and Population
- The nature of biodiversity also implies more than usual difficulty in determining precise system specifications and data structures and even in determining the “business case”. And yet as noted in ICT Toolkit, good specifications (especially functional specifications – what should the system do for the “business”) are the key to successful implementation, in a field where “information systems appear to have an alarmingly

high failure rate". For that reason, more than usual attention must be paid to stakeholder participation in defining needs, and in ways to integrate and harmonise data.

1.4 Principal Barriers to Effective Bio-Informatics Systems

It was recognised many years ago that scientific understanding of the Globe's environment, including its biological diversity, was essential to any efforts to achieve sustainable development and resource utilisation that protected future generations. The need for improved scientific cooperation and biodiversity information sharing has been noted with boring regularity and predates the Convention on Biological Diversity (CBD) by decades. This concern was foremost at the Stockholm Conference of 1971 that led to the formation of the United Nations Environment Programme (UNEP) in 1972 and soon after, its environmental information arm, the Global Environmental Monitoring System (GEMS), providing the first major global overviews of data and trends.

The UN Forum on Environmental Information in Montreal in 1991 confirmed that in spite of GEMS and the GIS-based UNEP-GRID Project, environmental information was:

- *Fragmented*
- *Difficult to access*
- *Of uncertain quality*
- *Inconsistent*
- *Lacking a scientific base in methods and models*
- *Not suitable for decision-making.*

The Rio Summit's Agenda 21 in 1992 set the direction for the next decade with its Chapter 40 specific to Information for Decision Makers, that noted:

> *"The gap in the availability, quality, coherence, standardisation and accessibility of data between the developed world and the developing world has been increasing, seriously impairing the capacities of countries to make informed decisions concerning environmental and development."*, and
>
> *"There is a lack of capacity, particularly in developing countries, and in many areas at the international level, for the collection and assessment of data, for their transformation into useful information and for their dissemination."*

To this day a number of barriers to more effective use of biodiversity information remain. Moving towards reducing these barriers must be integral to the design of bio-informatics systems. Principal barriers are:

- lack of consistency in nomenclature and vocabulary for describing biological entities and conditions
- parallel linguistic differences (exacerbated by the narrative form so common)
- overlapping and inconsistent reference systems for species and biological description (such as vegetation classification)
- project orientation – hence lack of archiving and re-use of data from one project to the next.

1.5 Issues in Bio-informatics

The emphasis in this Document is on the conditions that are peculiar to, or more pressing in bio-informatics than in information systems for other subject domains. These are summarised in the following paragraphs:

Meeting identified needs

Any Information System (IS) should be directed at meeting identified "business needs". In bio-informatics that translates to systems that are directed at addressing identified environmental issues, that is that address problems that need solving, decisions that need to be made, conditions that need to be improved. General objectives such as "improving availability of information" or "increased information sharing in the region" do not easily lead to useful functional specifications or effective means to assess success or failure (availability to whom and why?, information sharing for what shared decisions or goals?)

Need for harmonisation

Where a common business goal is identified (such as curtailing the spread of invasive species), it is clear that biodiversity information sharing between institutions will be essential. This then leads to the necessity to place considerable effort on information harmonisation (comparability and compatibility of content) as world-wide standards seldom exist. These issues will often be more onerous than technical issues of communications and database "interoperability".

Need for Sustainability

Much of the value of biodiversity information comes from its consistency over time in the form of a time series – to give early-warning of deteriorating conditions, or monitoring the impact (hopefully positive) of policies and decisions. It is therefore essential to consider the

long-term stability and sustainability of bio-informatics projects. This raises issues of absorptive capacity of institutions, provisions for on-going technical maintenance and support, and proper facilities for archiving datasets for potential future use.

CHAPTER 2 EFFECTIVE USE OF INFORMATION

2.1 Introduction

That information is essential to good biodiversity decision-making would seem to go without saying. Central to effective actions is the formulation of **policies, strategies** and **action plans** to promote the conservation and sustainable use of living resources. The performance of these policies depends on the degree to which they reflect the perspectives of different groups of people and, consequently, on the extent to which government, the private sector and society at large work in partnership. Experience has shown that the best policies result from a process of consultation, consensus building and, if need be, reconciliation amongst affected stakeholders, leading to solutions which balance economic and social goals with the need to safeguard the environment. When pursued intelligently, a commitment to such policies can improve the economic performance of companies and allow longer-lasting benefits to be delivered by governments.

Good policies have other features in common, including the effective use of information throughout the policy's lifetime. Indeed, the transition from exploitation of living resources to conservation and sustainable use requires a major investment in information and monitoring, otherwise it is not possible to judge progress.

In general, donor-financed projects are addressing policy issues that have been identified as priorities, and bio-informatics information management should ultimately support good decision-making promoting conservation and sustainable use of biological resources. This may involve development of policy instruments such as national conservation strategies, action plans, sectoral master plans, and so on, as well as measures to implement obligations to multi-national international environmental agreements (MEAs). Many MEAs are becoming increasingly target-oriented and thus require information systems to monitor progress through indicators.

2.2 Participation and Consensus Building

Individuals, communities, companies, nations and international bodies all make decisions which affect the sustainability of living resources and, consequently, all have a role to play in their conservation and sustainable use. No segment or level of society can be left out, since decisions made by one group can affect the livelihoods of others. Thus, wherever possible, policies aiming to conserve living resources should reflect the perspectives and needs of all stakeholders who stand to win or lose.

Even small-scale environmental challenges tend to arouse the interest of many stakeholder groups, perhaps more so than for industrial or economic development. Such groups are

typically politicians, civil servants, natural resource managers, local government officials, non-governmental and community-based organisations, business leaders, industry representatives, professional associations, scientific researchers, teachers, the general public, media and the international community.

Despite their apparent diversity, most of those involved fall into one of three categories: **government, private sector** and **civic society**. Together, these are referred to as the development 'triad' (Sandbrook 1994), since they represent the three core interests shaping development policy. In the long term, policies which do not represent all three broad interests are destined to falter, stall or fail. Figure 1 depicts the development triad in terms of its constituent stakeholders. Clearly, the term 'civic society' represents all of those who are not actively engaged with government or commercial activities, including non-governmental organisations, community-based organisations and individuals.

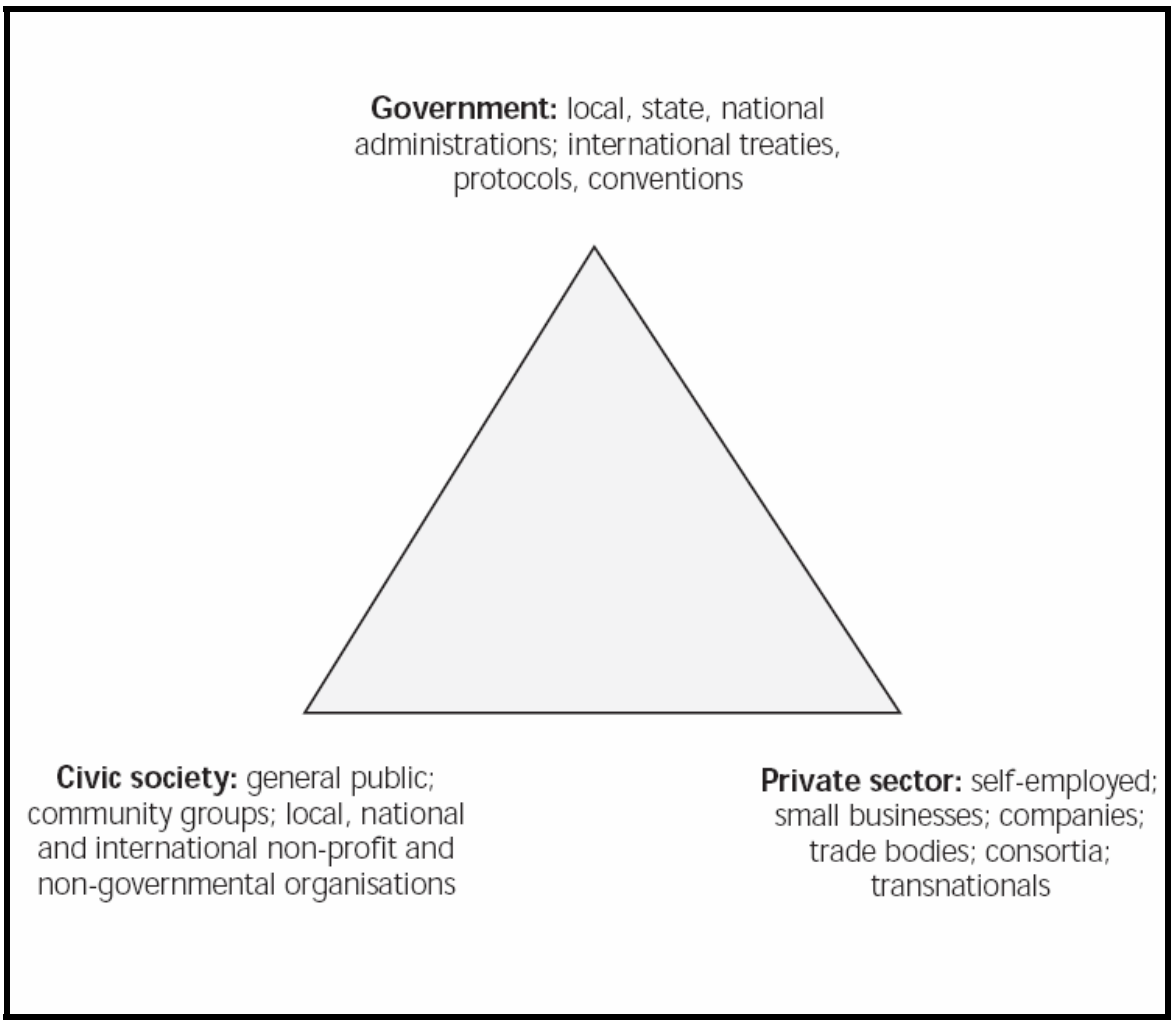

Figure 1: The development triad

2.3 Policy Development and Implementation

The development of policy (and associated decision and actions) is an iterative process requiring information and information systems support at various stages. Figure 2 illustrates a generic 'loop' of policy development processes: **plan, implement, monitor** and **review**. By making extensive use of information, the four processes enable policy goals to be achieved in a progressive manner through successive iterations of the loop.

Joining the loop at the implementation stage, activities are underway to meet agreed targets for conservation and sustainable use, set out in the planning stage. The loop proceeds to monitor the performance of the policy by, for example, obtaining data directly from measurements of biophysical variables, or reviewing the achievement of policy targets. Performance is then reviewed, leading to the production of clear and concise recommendations for policy-makers on how the policy should be refined in future.

Finally, the loop is 'closed' for another cycle by planning how the recommendations will be implemented, in terms of objectives, targets, roles and responsibilities. It should be noted that monitoring, review and planning activities would very often proceed in parallel with implementation. This enables continuous, rather than intermittent, feedback and policy-refinement.

Figure 2 simplifies what, in reality, is a complex, many-faceted process. For instance, the policy being developed may address an issue of national importance, such as the loss of crop genetic variability in an important agricultural zone, or a local concern, such as the restoration of a single eroded hillside. In both cases, successful implementation depends on maintaining a steady course around the management loop so that the four components flow into one another. To achieve this it is necessary to concentrate on the core objectives of the policy at all stages and to make sure that **all** actions contribute to these in some way.

An open, participatory approach encourages stakeholders from different levels and sectors of society to involve themselves in implementing the management loop. For instance, policy goals may be developed by consensus at fora established for this purpose; community groups or industry representatives can be asked to help monitor policy performance; and the success or failure of policies, plus options for future refinement, can be determined as a group.

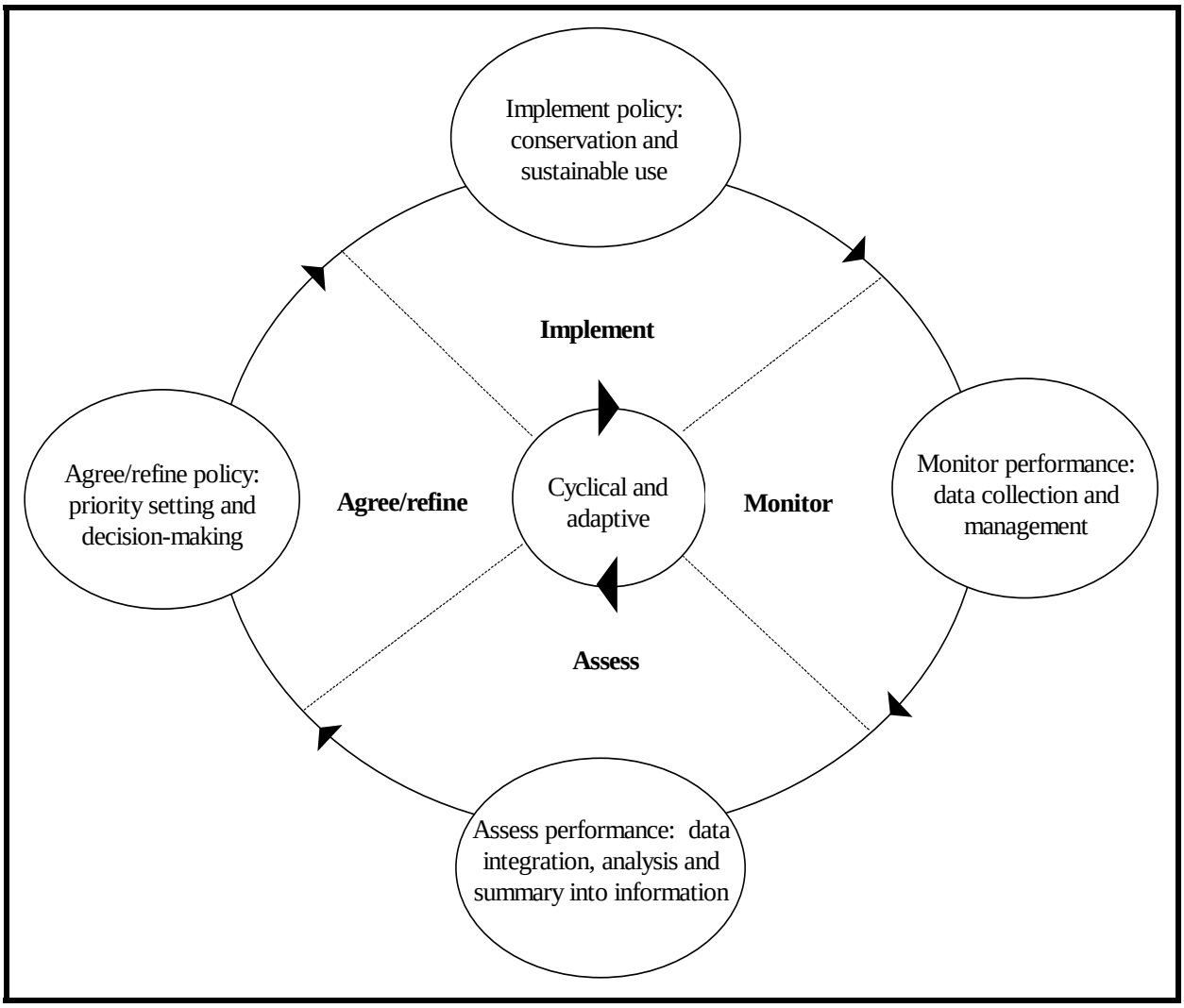

Figure 2: Management loop for policy-development

When stakeholder perspectives differ greatly, participation may not always be easy to manage. In such cases considerable effort is needed to keep stakeholders focused on policy goals, and not let implementation be diffused. Information, openly shared amongst all participants can often be the key to consensus and good decision-making. Common constraints on implementation include lack of financial resources, expertise and time, political interference and lack of political will. Clearly, the management loop is not an easy approach to follow, but it does offer a powerful means for developing and implementing policies capable of effectively responding to changing environmental conditions in a controlled, transparent fashion.

2.4 Information for Decision-Making

The term “Information Management” refers to organising, processing, analysing, storing, retrieving and disseminating information in order to support improved understanding (better decision-making) and thus, improved effectiveness and efficiency of an enterprise’s programmes. In abbreviated form, it is sometimes said that information management converts “data” into “information”. Information scientists often make a clear distinction between “data” (facts that result from measurements or observations of a phenomenon) and “information” (derived from data through assembly, analysis, interpretation or summarisation into a meaningful form). In day-to-day usage the distinction is much less clear. In the context of information systems it is common to use “data” for the input to any process and call the output “information” - which may then subsequently be the “data” that is input into the next process and so on. One agency’s information (or “information product”) is another’s data, even though it may be far removed from the initial raw scientific measurement.

Figure 3 illustrates this, with data at the base of the triangle and, moving towards the apex, information is generated from data as they are processed, manipulated, summarised, etc. At any level, do you have data or information? The figure also illustrates that in moving "up" the triangle -

- the data (or information) volume is likely to decrease
- the nature of the user will change
- subjectivity increases (increased intellectual interpretation and analysis)
- it will take time and resources to move from data to information.

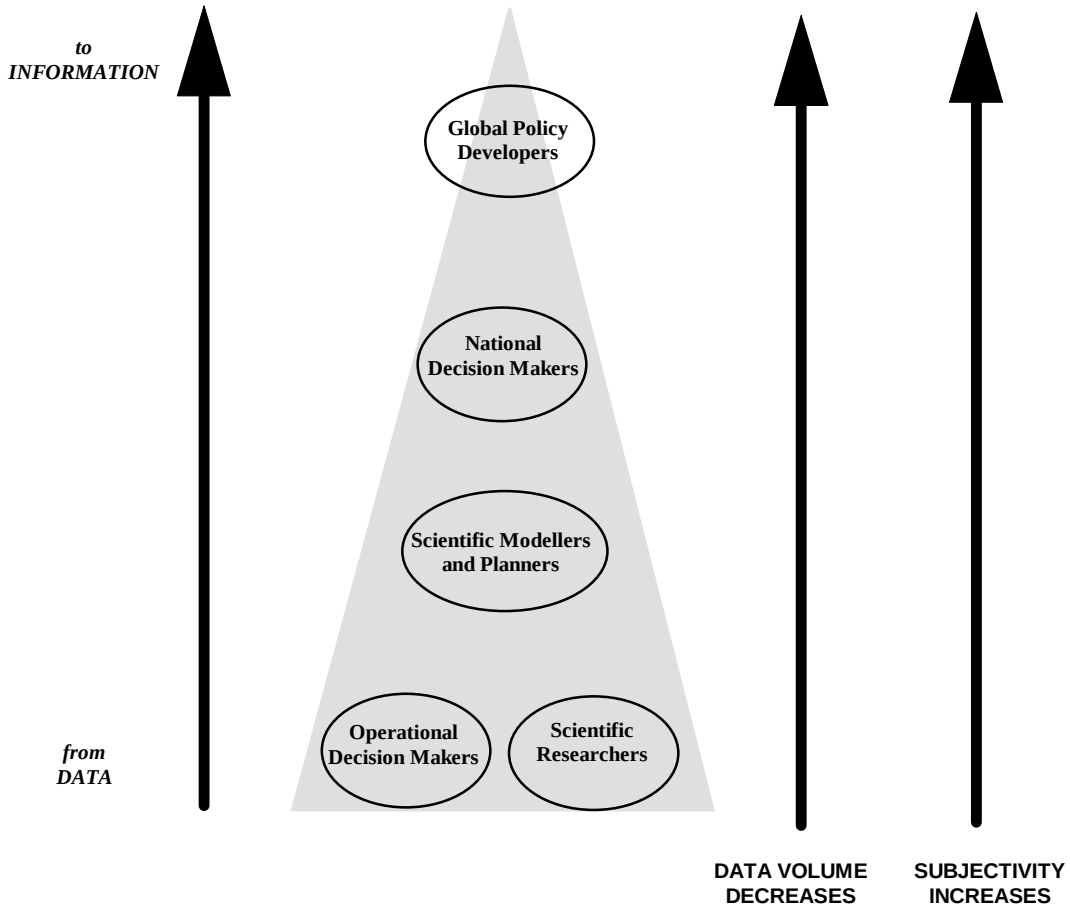

Figure 3: Information for decision-makers – the information triangle

In any information system intended to serve multiple users, outputs may come from any level, i.e. from basic data sets of the variables observed, through transformed, derived and generated data that are in forms more easily handled by specific user groups. Decision-makers are seldom able to use unprocessed data; they require the results of analysis and processing, often involving integration with other data and expert assessment. The production of good quality information at that high level almost always requires considerable resources and a mixture of expertise.

2.5 Determining Information Needs

Solutions to environmental concerns are usually complex and it is not always obvious how to determine what information is needed to achieve conservation goals. This is particularly true when decision-makers have only a general idea of their requirements, (for example, "to maintain the quality of the ecosystem"). Without the 'right' information, there is a risk that

stakeholders in environmental decisions will select inappropriate options, with potentially damaging consequences for living resources, and yet it may be difficult to determine just what key information should be monitored. At a lower level, in terms of information management requirements within the project, a clear definition of the data and information requirements is vital to the success of this component.

An analysis of "information needs" is therefore a crucial element in the early stages of the development of any information system. Such an analysis will result in specifications for data requirements, for the processing and analysis to be done (transforming data to information), for the form of products to be delivered, and so on. These are essential for the design of an information system, and incomplete or erroneous specifications are the most common cause of cost over-runs in systems implementation. Changes and additions to information needs become more and more costly as implementation progresses.

The analysis of needs must involve the stakeholders since they will be the primary users of the ultimate information system. Inviting too many stakeholders to participate in the analysis process can lead to high overall costs, prolonged consultation periods, introduction of extraneous issues, and the possibility of conflicts arising. At the same time it is important that no concerned group, whether potential data providers, subject matter experts, decision-makers, etc, be overlooked. The higher-level stakeholder analysis for the project as a whole should provide guidance on circumscribing the scope.

There are many tools and methods that can be applied to information needs analysis. Any particular analysis may require only a subset of these, the most appropriate methods depending on its depth, the nature of the issues being addressed, and the range of stakeholders involved.

Structured approaches are suitable in situations where information needs are already broadly defined, and the goal is to elaborate these in more detail. Questionnaires are particularly useful in situations where organisations are mandated to prepare information in a prescribed form, and feedback is required from users on the quality of the information supplied, or ideas for future improvements. Structured interviews provide an opportunity to engage users in free-flowing discussions, yet keep to an agenda with a fixed set of questions.

Where new information capacities are being developed, perhaps by a series of collaborating organisations, more sophisticated techniques may be employed to engage stakeholders in consultation. Less structure may be preferable, leading to the use of alternative, participatory approaches, such as visioning exercises, brainstorming and problem tree analysis.

Finally, process models can be employed at any stage in the information needs analysis to illustrate the relationship between information sources and selected processes in an operation. They serve to simplify and consolidate otherwise complex data flows.

2.6 Use of Existing Information Sources

As outlined in Section 1.2, the full breadth of biodiversity information encompasses a number of major categories that in most countries are the responsibility of different established agencies and institutions. For example, a plant species, database may be maintained by the national herbarium, whereas data on protected areas may be managed by the national parks agency. In addition, information other than the "core" biodiversity types is needed e.g. administrative boundaries, infrastructure (roads, railways, etc,), soils, population, and each will be kept by the agency with the relevant mandate.

Data collection is becoming an increasingly expensive and time-consuming operation so it makes practical sense to use existing sources not only for current data but potentially, for continuing monitoring activities using the expertise and procedures in place. In cases where there may be deficiencies or problems, it may be better to invest in improving these existing conditions rather than developing something separate.

The concept of having specific agencies recognised as definitive sources for different data types is widely used and promoted by both donors and national governments, and information "networks" are common. These may be in place to provide data that is commonly required for multiple purposes, needed by many different types of project. For example, an official in a ministry of agriculture may require a map showing the distribution of wild relatives of crops in a specific location. This need differs greatly from that of a forest officer wishing to know the sustainability of logging operations in the same area. However, much of the baseline data required to build the maps (e.g. administrative boundaries, rivers, vegetation and topography) may be the same. Networks also may be created to allow common access to specialised types of data, such as species taxonomy. Biodiversity-related projects tend to need both kinds of information network, and the latter is especially common in this field (see Ch 4).

A project addressing a regional issue needs a very different level of information from one that addresses more local impact. The sources of the same types of information may therefore vary depending upon the level being addressed, and range from international bodies down to offices of local authorities. In many cases, activities need to be coordinated at various levels to ensure standards are adhered to, overlap and duplication are minimised, and local-scale datasets can be smoothly integrated and generalised to support national and regional-level decision-making. Achieving this is both assisted and complicated by the vast array of global and regional biodiversity-related information services and networks, many of which have exaggerated claims to be comprehensive or definitive.

CHAPTER 3 GOOD PRACTICE IN INFORMATION MANAGEMENT

3.1 Information Management Framework

There are a number of general principles to guide information management activities in a project.

- Information assembly, communication and maintenance should be planned as an integral part of the program planning and budgetary process.
- Information should be recognised as an asset and actively and professionally managed.
- Information holdings should augment, not duplicate, information managed and available from recognised authoritative sources.

Figure 4 shows an "end-to-end" information management process reflecting these. The overall project objectives are the driving force behind the requirements which determine the data needed, either gathered from existing sources or from data collection programs. The data is assembled and integrated – metadata is reviewed, data is processed, additional metadata supplied, datasets are merged, etc. Products of various kinds (reports, books, maps, newsletters, briefings, datasets) are generated and distributed to users. Throughout this process is the need to ensure data quality and implement procedures for appropriate preservation and archiving.

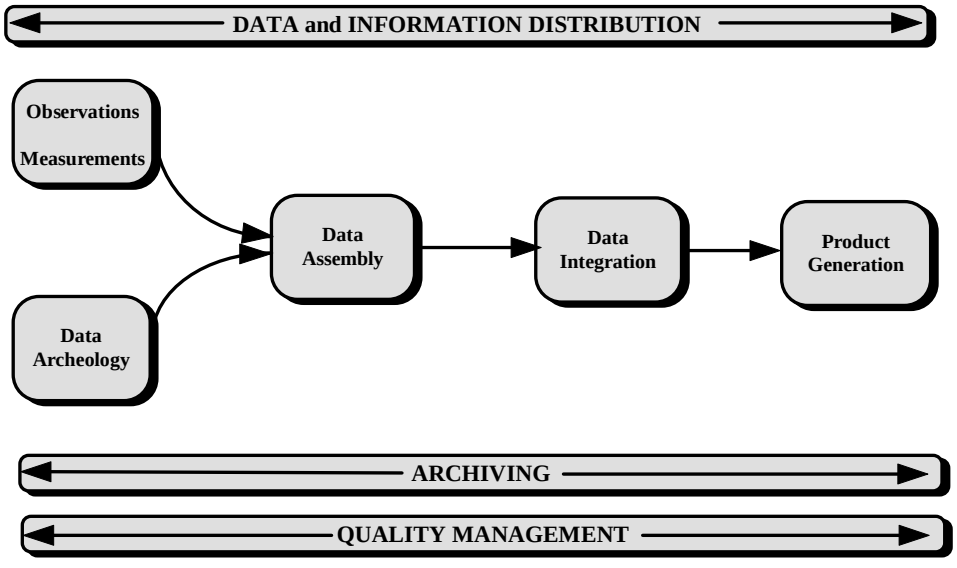

Figure 4: "End-to-End" Information Management Process

This process is intended to ensure effective and efficient information management within a project, specifically to:

- Avoid re-doing data collection because previous data has been or was poorly documented
- Reduce costs of data assembly and integration
- Ensure a consistency of presentation and compatibility of information
- Improve the quality of information
- Prevent data loss
- Reduce any time lag to prepare and present information in support of decision-making
- Make efficient multiple use of datasets.

3.2 Implementation of Information Systems

Informatics and Communication Technologies (ICT) are almost inevitably used in projects that have an information management component, whatever the sector. Reference has been made in previous sections to a "Task Managers' ICT Toolkit" which is designed to give guidance on ICT components in World Bank operations. It is in two parts:

- ***A Route Map for ICT Components***, intended to raise the awareness of task managers as to what ICT can do, good practices, limitations, risks and common pitfalls, and
- ***Good Practice for Planning, Delivering and Sustaining ICT Products***, which goes into greater detail giving guidelines for effective planning, management and delivery of ICT components.

The materials are well organised and well-reasoned and fully applicable to ICT components targeting biodiversity information management and should be used extensively. This top-level guidance means that there is little to add in this document by way of general guidance and comment on the application of technologies, and the following sections deal with three specific aspects of information management. These are technology-related, but have institutional or organisational aspects that are particularly of relevance to bio-informatics.

3.3 Custodianship

An important aspect of the end-to-end information management regime is the concept of custodianship. A custodian is the body **responsible** for the development, maintenance and quality of a dataset, and for the arrangements for access to it. The most important aspect of a custodian is that they should have the scientific and technical knowledge and expertise to be in

the best position to assess and ensure data quality, and indicate the appropriate uses and limitations of the data.

The custodianship concept is designed to ensure the availability of the highest quality data, while reducing redundancy of data collection and maintenance - such as the maintenance of multiple copies or various "versions" of the same data. It thus serves the dual purpose of promoting quality and cost-efficiency.

The dataset custodian is responsible for:

- building the dataset (primary or secondary data entry)
- maintaining the dataset (up-date and additions)
- ensuring quality (data validation, error correction)
- ensuring integrity (back-up and security precautions)
- providing access (possibly with appropriate restrictions)
- providing directory level metadata
- maintaining full and accurate additional information (metadata) needed to correctly use the data
- providing advice on appropriate uses and interpretation of the dataset
- providing for the archiving of the dataset.

Some form of active "management" of the custodianship functions must be put in place, including clear identification and designation of custodians of any data set. A custodian may choose to delegate some of the specified functions - such as archiving or day-to-day operational management of the data - but the custodian must maintain the responsibility of "ownership".

A custodian should be designated at an early stage in the creation of a dataset. All existing datasets may have an "owner" who is (implicitly) the custodian and this should be acknowledged, or a custodian assigned retroactively. As has been noted earlier, in biodiversity, a number of national institutions may be involved, for instance museums and herbaria, environment and resource ministries, universities and NGOs. Their custodian roles need to be explicitly recognised.

3.4 Metadata

Metadata are "data about data", describing such things as the location, sources, general content, condition, format, etc. of existing datasets. Although not explicit in the end-to-end information

management figure above, they are an essential component constituting documentation covering all aspects of the process. Metadatabase systems are systems specifically designed to manage metadata i.e. to provide facilities for input, update, retrieval and reporting of data about data. Such systems may be used within a single institution to organise and maintain their own data holdings. They are also used on a broader level and can then provide a mechanism through which data producers can ensure that potential users are made aware of existing data, their nature, and how they might be obtained.

In general, metadata are at two levels. The first, referred to as "directory level", identifies the dataset through such items as a general description (subject, geographic coverage, dates, collection methods, processing done,), details of availability (access conditions, costs...), contact point (for further information and/or ordering). These are items that are essentially common to all types of dataset, regardless of the subject matter. The second or "dataset level" is subject matter specific, for instance, instrument settings, calibration data, adjustment factors, classification systems and legends, reference standards, taxonomies, etc.

In summary, directory level metadata enables a potential user to judge whether a dataset might be useful for the intended purpose and how to obtain it; the dataset level metadata allows the data to be used correctly, once obtained.

Metadata must exist to enable an organisation to use datasets effectively in their own projects, and to deliver data and information for external use. They are used specifically to facilitate data access, to enable quality assessments to be made, and in archiving.

Considerable international work has been done relating to metadata at the directory level and, although there is no single accepted global standard, there is commonality across some in wide use. The requirements for dataset level metadata are dependent upon the subject matter. Relevant standards for both areas are discussed in the next chapter.

3.5 Archiving

The preservation of data and information to enable use over the long-term is intrinsic to the concept that data and information holdings are a corporate resource. Management of that resource requires measures to prevent loss or damage, and to maximise potential for return on investment. Furthermore, in the case of biodiversity information, archiving is essential to meet the identified need to build a time-series of measurements for monitoring impacts and status.

Archiving is an essential element of the end-to-end data management framework and there are potentially several points at which material should be archived. These points vary depending upon the dataset(s) and processes but, should be clearly defined and documented in the overall

management system implemented. At all stages and in all cases, relevant metadata must be included in the archive material to ensure that the data and information can indeed be used in a meaningful fashion at some later date.

CHAPTER 4 EFFECTIVE USE OF INTERNATIONAL STANDARDS

4.1 Introduction

As identified in Chapter 1, one of the essential needs in bio-informatics is the sharing and integration of information between institutions in order to understand inter-relationships and obtain the "big picture" for effective decision-making. This requires that information networks and systems must be able to cross-communicate. Even if a Bank project appears to stand alone, it is imperative that the component being developed **adds** to the national and regional capacity to share information.

Although this section is dealing with international standards relating to the use of technologies, there are significant factors of a non-technical nature that influence technology choices and need to be considered in planning implementation of information systems. These include:

- When deciding on the standards and practices to adopt for sharing biodiversity information, it is important to consider the needs and capabilities of the contributors and end-users. If the technologies and standards employed do not allow easy integration or assist in their programme of work, then their use is of little value.
- Special consideration also needs to be given to the development of policies for the management and distribution of information provided by partners and third parties. Data distribution policies need to be developed to allow these providers to include their data in the network, whilst maintaining ownership and control over how the data is distributed and used within other systems. This will also have an effect on the type of system that is employed to store and manage the data within the network.
- Data may be incorporated into bio-informatics information systems either by compiling the information in some kind of central repository or by linking to information held in the different organisations. There are advantages and disadvantages to each with regard to accuracy, infrastructure within the collaborating centres, and the currency of data provided. In practice it is advisable to construct the system in such a way that it is possible to include data held both locally and remotely, thus allowing as many organisations as possible to collaborate.
- Data ownership is a growing issue of concern to many specialised reference sources. Considerable time, effort and money are required to compile taxonomic information, hence, institutions and individuals can be reluctant to make data freely available and accessible so that others can 'reap the rewards' of their labour. Where data collection is specifically funded through governments or foundations, public access to the resulting data is often a condition of funding, so these issues may be of lesser concern.

In addition, many scientific journals now insist, as a condition of publication, that the data used in the paper should be made freely available (usually over the Internet). However, where data are gathered in a voluntary capacity, as part of the core activities of an institution, or in any commercial context, problems with data ownership are likely to persist. To address this concern, many reference sources have data "user agreements" requiring full acknowledgment of the source database.

There are two main areas of concern for use of international standards: technical "interoperability" and semantic (content) compatibility. "Interoperability" while having various meanings, normally refers to the technical ability of databases with heterogeneous technology to be connected and interact as if one. In this way it is possible for a single query to extract and integrate data from multiple databases. This is an improvement on the use of "data exchange formats" by which data are transformed to a common format for communication and subsequent integration with other datasets and re-conversion to local formats. By and large, interoperability provides for the **theoretical** ability to integrate, but ignores the data content issues. It is a necessary but not sufficient condition to ensure that the data so integrated is compatible, and hence effectively useable. To ensure this usability, it is further necessary to address the identified semantic barriers, such as:

- lack of consistency in nomenclature and vocabulary for describing biological entities and conditions
- parallel linguistic differences (exacerbated by the narrative form so common in biodiversity)
- overlapping and inconsistent reference systems for species (taxonomy), and biological description (such as vegetation classification).

Both the interoperability and content harmonisation issues can greatly benefit from the use of international standards (or commonly used practices of respected international organisations even if not formalised by official standards bodies). In all cases, a review should be made of the applicable international standards and practices, and wherever feasible such accepted standards should be adopted in order to maximize the potential for information sharing, and minimize the proliferation of divergent "standards" and practices.

The following sections emphasize the standards and practices most relevant and specific to the subject domain of biodiversity, with the understanding that accepted general IT standards are to be applied in the first instance, and form the basis for more particular considerations. Issues of interoperability of heterogeneous databases are not unique to biodiversity, and are therefore only touched on briefly in this document (see also the ICT Toolkit). GIS technology is frequently used in bio-informatics and so some elaboration of the related issues and standards is provided. However, the proviso must be made that rate of change of technology is very

rapid (and with it related standards and protocols) so any specific tools mentioned may be obsolete in a few years.

International standards are discussed with respect to content harmonisation in key areas such as taxonomy and species reference archives, biological vocabularies, and other reference sources.

4.2 Interoperability

4.2.1 General Systems Interoperability

The technology for interoperability is complex and, as noted above, evolving rapidly. The full potential of any new technology for biodiversity can only be realised through a sound analysis of information requirements and careful design of services to address these requirements. Bank projects should promote the use of appropriate technologies to provide the required services and avoid the temptation of technology for the sake of technology.

The successful development of interoperable systems for data sharing requires that the organisations that are the data providers have well-established data management infrastructures that apply accepted industry-wide ICT standards for system design and management, for instance, data modelling standards that define formal methods to determine data structures and ensure efficient storage and retrieval. The general adoption of the relational model and standardisation on Structured Query Language (SQL) has resulted in a degree of interoperability between systems. Current database design may make use of object-oriented (rather than relational) techniques. Whatever methods are used, standardised design and development of systems are vital so that valuable data has a long useful life and can be shared.

Two key aspects of achieving technical interoperability (independent of the subject matter content) are the technical protocols for information transfer (mainly using the World Wide Web), and metadata – the data about data that are necessary to align the database entities from one system to another.

With regard to standards in these areas, the three most relevant international standards organisations are:

- The World Wide Web Consortium (W3C) - concerned with common protocols to ensure an interoperable WWW
- The International Organization for Standardisation (ISO) - with a range of standards for both spatial and non-spatial data transfer and metadata
- The Open GIS Consortium (OGC) - in connection with geo-spatial metadata standards.

4.2.2 Web Services Technology

With regard to technology standards for Web Services, the following generic standards and protocols are of increasing use:

- JDBC and ODBC – both enable access to heterogeneous database systems
- XML – a simple meta-language recommended by W3C, becoming an essential tool in encoding information
- Web Services (WS) – based on XML and SOAP, a framework for distributed information networks

(Refer also toAnnex 3.)

It is noted that the WS architecture and supporting technologies have been endorsed by major hardware and software vendors and are being taken up by commercial software developers, standards organisations and biodiversity information networks. However, they are still very new and there are obstacles to overcome to realise their full potential.

4.2.3 Metadata Standards

To make metadata a successful data sharing and information discovery tool, its compilation and management must be part of the institutional data management culture. Establishing metadata is not a trivial task, either in terms of understanding it role, its utility, the tools used to manage it, or in its compilation. There is often little recognition of the effort required and it is commonly an under-resourced area.

Also there are several different levels of metadata to be considered, such as that intended for "discovery" (directory level - describing an information resource) and for "semantics" (dataset level - concerning the meaning of keywords for example).

At the higher level, there are several generic metadata standards of relevance. The "Dublin Core" (DC or DCMI) is intended for use in describing any type of electronic information resource and is relatively simple, consisting of only 15 elements. The US Federal Geographic Data Committee (FGDC) is responsible for developing a metadata standard for spatial data and US Federal Agencies are required to use this in documenting geo-spatial data holdings. This, with others, has been adapted by the ISO in their development of spatial metadata standards.

4.2.4 Standards More Specific to Bio-informatics

More specific to bio-informatics, the Biological Data Working Group (of the FGDC) have defined a Biological Data Profile (BDP). This includes additional elements to fully describe biological data but also, since biodiversity data may not always be spatial, removes the mandatory requirement for geo-spatial elements where applicable. The Knowledge Network for Biocomplexity has developed a set of tools to "discover, access, interpret, integrate and

analyse complex ecological data from a distributed set of field stations, laboratories, research sites and individual researchers". One component is a metadata standard – Ecological Metadata Language (EML) – that encompasses the elements of the standards mentioned above.

Four other standards, developed specifically to aid biodiversity information management and exchange are:

- The Darwin Core – developed as part of The Species Analyst as an extension of the generic Dublin Core
- Access to Biological Collections Data (ABCD) – from a joint working group of TDWG and CODATA
- Distributed Generic Information Retrieval (DiGIR)
- Xanthoria – a metadata query system associated with EML.

Controlled vocabularies are an important adjunct to metadata systems and multi-lingual vocabularies provide a simple means to cross language barriers. These are essential and must be specific to the subject matter domain (see Section 4.4 below).

4.3 Standards and Practices for Sharing of GIS-based Information

Geographic Information Systems (GIS) are designed to collate, manage, analyse and present information with a geographical (locational) component. Such systems are therefore potentially of great value in managing biodiversity information and are frequently implemented as part of bio-informatics projects. Increasing numbers of networks and international institutions are now making available compilations of spatially referenced biodiversity information.

In a bio-informatics project, there is frequently a need to both integrate biodiversity data of different types, from disparate sources, and to link this information with other non-biological data. Spatial data standards provide the architecture for the integration of GIS data from multiple sources and in multiple data formats, along with providing the ability to more fully integrate spatial data with non-spatial data. Of particular note is the Open GIS Consortium (OGC) which links companies, government agencies and universities concerned with interoperability standards for geo-processing.

Three areas of spatial data standards need to be considered.

GIS information exchange standards: These standards allow for the exchange of data between different organisations, so that the same information can be combined with other data held on different software platforms and in different formats. One approach is to use specific format

conversion extensions to GIS software packages, but this can add some undesirable complexity. To overcome this, various standard interchange and "open" file formats have been developed that allow data to be viewed (or used) within multiple software systems without the need for data converters. Probably the most significant of these file formats is GML (Geography Markup Language), a specialised form of XML (Extensible Markup Language), developed as part of the OGC initiatives. GML has advantages in that it is a text-based standard and not reliant on any specific software. However, there can be problems when dealing with large datasets.

GIS information inter-operability standards: These standards allow for the integration of data between different organisations, which may or may not be working on the same GIS software platform. The data can be linked interactively through either the organisation's internal network or through the Internet. The development of web-based technologies has allowed for the evolution of Interactive Mapping Services (IMS), which give users the capability to view GIS data in an Internet browser environment. The OGC has been instrumental in the creation of standards to enable IMSs to interact with each other and with multiple GIS packages. Currently there are standards for Web Map Services (WMS) and Web Feature Services (WFS), relating to IMS and Internet Data Services (IDS) respectively. Essentially, with IMS, processing and map creation are done at the server and the map image is delivered to the user; with IDS, the server delivers the data needed to create the map, and the map is produced at the user workstation.

Spatial metadata standards: Standards for GIS metadata have developed in line with the advances in technology and data formats. A wide range of metadata standards now exists covering different data types and user needs. International Organization for Standardization have developed standard ISO 19115 that has combined aspects of several other metadata standards to create a universal standard for the storage and distribution of metadata.

As mentioned in the introduction (Section 4.1), the ideal is for true interoperability, meaning that the information remains in the possession of the best-qualified custodians and is incorporated into other organisations through Web applications. By avoiding physical exchange through "exchange formats" it is ensured that the end-user is always working with the most up-to-date version of the dataset.

To integrate GIS data stored in different formats in different locations across the web, access must be provided through an Internet map or data server. To be fully interoperable between software systems, this server needs to provide information that complies with the OGC standards for WMS or WFS. Such compliant data can be viewed in many software packages, and also can be imported into applications running on different software such as the OGC compliant map viewer used in the FGDC data clearinghouse.

With regard to spatial information, for practical reasons the use of information exchange formats is a viable and convenient option for the sharing of data, at least initially. Although the GML approach is desirable, formats commonly used include ArcINFO export files, "shape" files and DXF (AutoCAD Digital Exchange Format).

4.4 Species and Taxonomy Reference Archives

4.4.1 Introduction

The species name is key to the management and dissemination of biodiversity information. Taxonomic Authority Archives (TAA), that is, species lists with taxonomic hierarchy, accepted names, synonyms and common names, are the cornerstone for many types of biodiversity information systems. Such authority files (a type of metadata) allow for the unambiguous linkage of species to related information such as distribution, species population, trends, threats, habitats and legal protection status, and are almost always needed in the development of such information systems. Being name-based (including synonyms), TAAs can extract and positively link information from narrative descriptions, as well as specific database fields. Hence, despite some debate and uncertainty, consistently organised taxonomic information is essential to manage information on species, an important component of biodiversity.

Many TAAs are specialised and deal only with specific taxonomic groups, which may be on a global, regional or national scale (e.g. the global FishBase and Index Kewensis). Some aim to be comprehensive and eventually include all taxonomic groups. The Species 2000 initiative estimates that the existing global species databases presently account for some 40% of the total known species.

4.4.2 Comprehensive Reference Sources

Four international comprehensive reference sources of note are described below:

- *The Global Biodiversity Information Facility (GBIF).* Established in 2001, its mission is to make the world's primary data on biodiversity freely and universally available via the Internet. It provides digital access to information on taxonomic hierarchies with links to further information where available. A number of areas of emphasis have been identified by GBIF, namely: data access and interoperability; digitisation of natural history collections; electronic cataloguing of names of known organisms; and outreach and capacity building. According to their website "*In the near term, GBIF will provide a global metadata registry of the available biodiversity data with open interfaces. Anyone can then use it to construct thematic portals and specialised search facilities. Building on the contents of this registry, GBIF will provide its own central portal that enables simultaneous queries against biodiversity databases held by distributed,*

worldwide sources." Of particular note is the GBIF programme ECAT (Electronic Catalogue of Names of Known Organisms) which is working towards an electronic catalogue of the names of known organisms. It aims to provide content infrastructure to enable searches across multiple information domains, to make seed-money awards to speed progress of Catalogue development, and to develop the Taxonomic Name Service function of GBIF information architecture.

- *The Integrated Taxonomic Information System (ITIS)* aims to provide authoritative taxonomic information on plants, animals, fungi, and microbes. Although originally established with a North American focus, ITIS has since expanded and includes information on species from around the globe. The goal is to create an easily accessible database with reliable information on species names and their hierarchical classification. ITIS includes documented taxonomic information on flora and fauna from both aquatic and terrestrial habitats. Currently coverage for some taxa is global, and for others it is as yet confined to North America, and there are many gaps in ITIS in the coverage of South American taxa. Each of the ITIS countries (US, Canada and Mexico) has a separate portal in which data can be queried. In addition, the ITIS data can be queried through the GBIF portal.
- *Species 2000* is a "federation" of database organisations working closely with users, taxonomists and sponsoring agencies. Species 2000 aims to provide an index of all known species in the world through an array of participant global species databases covering each of the major groups of organisms. Each such database will cover all known species in the group, using a consistent taxonomic system.
- *The All Species Foundation* was established to catalogue every living species on earth. This inventory would need to enlist the support and cooperation of scientific organizations around the world. All Species is intended to be a temporary endeavour which will cease to exist in 25 years when its mission to compile a list of all species is completed. Information in All Species is sourced from both comprehensive and specialist TAAs around the world. At the moment, the activities of All Species as a TAA appear to have been surpassed by the activities of the other major initiatives. Funding constraints have led to a scaling down of its activities.

Relationships between these TAAs are complex. ITIS and Species 2000 together produce the Catalogue of Life, a uniform and validated index to the world's known species collated by taxonomists throughout the world. It is available on a CD and can also be downloaded from the web. ITIS and Species 2000 signed a Memorandum of Understanding in November 2003 to further enhance collaboration.

GBIF has recently signed a three-year Memorandum of Cooperation with the Catalogue of Life Partnership. The Memorandum provides a basis for mutual co-operation and a

framework for GBIF to access the Catalogue of Life and to use it in its services. The synonymic species checklists provided by the Catalogue of Life partnership will be made available to GBIF, and it is anticipated that they will play a key role in the name-service and indexing functions of the GBIF portal. The role of GBIF differs from that of the Catalogue of Life partners. While ITIS and Species 2000 will provide a checklist of species of the world, GBIF will be a portal both to that information as well as to large amounts of other information on species collections from museums and herbaria throughout the world. GBIF will contain not just the Catalogue of Life species names but also names that have long since gone out of use, that have been misused, misspelled etc and location and other information linked to each such specimen.

The initial focus of GBIF appears to be on museum specimens, which should provide new sources of information and will complement work already undertaken by ITIS and Species 2000.

Overall, GBIF is emerging as a leading global player in the field of comprehensive reference material with strong support from, and connections to, the Catalogue of Life Partnership. It is therefore the recommended "first stop" in the search for standardised species and taxonomic reference data and it can then lead to connections to the more specialised services (see below).

4.4.3 Specialised Reference Sources

In addition to the four comprehensive systems, a number of more specialised taxonomic authority archives exist, including Fishbase, MammalBase, GloBIS (butterflies), Index Kewensis (seed-bearing plants), ILDIS (legumes), CABI (chiefly fungi), CGIAR SINGER (germplasm), UNEP-WCMC Species Database (species protected by international conventions), IUCN Red List (taxa that are facing a higher risk of global extinction), Zoological Record (citations to scientific literature on animals) and EUNIS (European Species of Conservation Concern). Relationships between them are uneven and there are gaps and overlaps, although many have formal or informal cooperative understandings with the comprehensive TAAs.

The value of these reference sources and associated information management tools should not be under-rated where the subject matter domain is relevant.

4.5 Biodiversity Thesauri

Controlled vocabularies are an essential adjunct to metadata, and are subject domain dependent. In bio-informatics such vocabularies are very useful in key wording of descriptive information, and for helping to bridge linguistic barriers. Thesauri and controlled vocabulary lists can primarily be used to assist in two key areas of knowledge management - information cataloguing and information discovery.

Six internationally recognized environmental vocabularies and thesauri of relevance are:

- *CBD Controlled Vocabulary:* The CBD Controlled Vocabulary was developed by the Convention on Biological Diversity Secretariat to provide a list of terms that could be used as descriptors for the Convention's web site including the Clearing-House Mechanism (CHM). The list is also recommended for use by CHM National Focal Points to describe the contents of their national CHM web sites. It is intended to facilitate the searching, locating and retrieval of information by linking similar documents and resources with a unique term. It would also standardize descriptions of web sites, and so assist in efforts to make information interoperable within the CHM network, and with other websites related to the CBD. The CBD Controlled Vocabulary is regularly updated with new terms as needed.
- *UNEP EnVoc:* Developed by the United Nations Environment Program, EnVoc is a multilingual thesaurus with a controlled and structured vocabulary for use in indexing, storing and retrieval of environmental information. The latest edition contains categorised and alphabetical lists of subjects, together with a KWIC (KeyWords in Context) list. This thesaurus is available in the six official United Nations languages. EnVoc supersedes the former INFOTERRA Thesaurus of Environmental Terms. Available for purchase as a printed document, this thesaurus has also been accessible for on-line querying although the service seems frequently unavailable.
- *FAO AgroVoc:* The AgroVoc thesaurus of the Food and Agriculture Organisation (FAO) is designed to cover the terminology of all subject fields of agriculture, forestry, fisheries, food and related domains, in order to catalogue documents. AgroVoc is currently at version 4 and is available for on-line browsing. It supports seven languages: Arabic, Chinese, Czech, English, French, Spanish and Portuguese.
- *GEMET 2001:* The General Multilingual Environmental Thesaurus (GEMET) was developed by the European Environment Agency (EEA), with the co-operation of international experts, to serve the needs of environmental information systems. Analysis and evaluation work led to a core terminology of 5,400 generalised environmental terms and their definitions. This vocabulary ensures validated indexing, cataloguing and retrieval within environmental information services as well as harmonised translations in the multilingual European network. GEMET 2001 is provided as a polyhierarchically structured thesaurus and is now available in 19 languages. It provides a complete numerical equivalence (all descriptors have an equivalent) with the included languages. The semantic equivalence (correct correspondence of meaning between languages) has been separately ensured.
- *CIESIN (Center for International Earth Science Information Network) Indexing Vocabulary:* The CIESIN Indexing Vocabulary was developed to index data resources

and datasets related to the human interactions in global change. The Vocabulary comprises two elements: CIESIN Indexing Terms and CIESIN Location Indexing Terms. The former is a controlled thesaurus of socio-economic and environmental terms arranged in nine Science Data Domains with all of the terms organised in a hierarchical relationship (of broader to narrower terms). The Location Indexing Terms is a controlled vocabulary developed to represent the geographical, geopolitical, and spatial coverage of socio-economic and environmental data resources. At the time of writing, these had been last revised in 2002 but the Indexing Terms have not been revised since 1997. The database is available for on-line interrogation. All terminology is in English.

- *NBII/CSA Biocomplexity Thesaurus:* The Biocomplexity Thesaurus was developed in 2002-3 through a partnership between the US National Biodiversity Information Infrastructure (NBII) project and CSA, a leading bibliographic database provider. The thesaurus was developed through the merger and reconciliation of five thesauri – the CSA Aquatic Sciences and Fisheries Thesaurus, the CSA Life Sciences Thesaurus, the CSA Pollution Thesaurus, the CSA Sociological Thesaurus and the CERES/NBII Thesaurus. The thesaurus is overseen by the NBII Thesaurus Working Group, which considers its expansion and addition/modification of terms. The thesaurus holds its terms with relationships including Subject Categories (SC). It can be accessed on-line, and only provides results in English on the website.

Some convergence in biodiversity thesauri towards GEMET is evident and it is advisable to follow progress. UNEP recently (April 2004) convened a meeting of major participants in the field of multi-lingual thesauri, that for the first time brought together the major providers of environmental terminologies to discuss the status of their terminologies, and how these resources can be integrated using new technologies. The meeting examined many of the overlapping thesaurus initiatives underway, and opportunities to bring them together using the available web-based collaborative technologies for developing a global multi-lingual system. Representatives from many of the major thesaurus initiatives participated. Further annual meetings are planned to allow all parties to be apprised of recent developments and to foster collaborative efforts, essentially based on GEMET.

4.6 Other Reference Systems and Sources

Beyond species taxonomy and biodiversity vocabulary several other areas of international standardisation are relevant. This mainly involves standardized structures for classifying or typifying biophysical parameters and conditions – for instance for soil, vegetation and ecosystem mapping. The adoption of such formalisms across a country or region greatly

enhances the capacity to usefully integrate information. Most such classification systems are maintained by specialised international organisations or NGOs, through scientific cooperation.

Some of the more relevant are:

FAO - for soil and vegetation ("land cover") classification

The Ramsar Convention Bureau – for wetland classification

IUCN – protected area management categories, species threat categories

World Commission on Protected Areas – standard specifications ("core datasets") for describing protected areas

UNEP-WCMC – forest and coral reef classification

CHAPTER 5 EXPERIENCES AND APPROACHES OF OTHER DONORS

5.1 Donor Approaches to ICT

In recent years, most bilateral and multilateral donors have recognised the role that ICT can play in development in a range of sectors, and have responded with varying strategic approaches towards funding and delivery of projects with significant ICT components. The emphasis varies – developing ICT infrastructure, human resource capacity, e-governance, policy development – within the general framework of creating an enabling environment (DAC, 2005). Most bilateral donors either have ICT-specific programmes or indicate that ICT is "mainstreamed" in development projects. Major multi-lateral donors such as The World Bank and the Asian Development Bank have developed strategic approaches for this sector (as reflected in, for example, the ADB documents "*Toward E-development in Asia and the Pacific - A strategic approach to information and communication technology*", and "*Information and Communication Technology for Development in the Pacific - The role of information and communication technology (ICT) in fostering poverty reduction efforts and socioeconomic development in the Pacific region*". The UNDP has through its Sustainable Development Network been fostering ICT for development for many years and has strategic guidance documents.

> *"UNDP's ICTD strategy focuses on upstream policy advice to help countries design a strategic approach to ICT as an enabler for development and link it to Poverty Reduction Strategies (PRS) and related development focus processes. This is complemented by support to the implementation of ICTD priority programmes based on a multi- stakeholder approach and innovative national and global partnerships to secure additional resources and expertise.*
>
> *To accomplish the above goals, UNDP has identified, in consultation with developing country stakeholders, five strategic areas for ICTD related interventions. They are:*
>
> - *National ICT for Development Strategies*
> - *Capacity development through strategy implementation*
> - *E-governance to promote citizen participation and government transparency*
> - *Bottom-up ICTD initiatives to support civil society and SMMEs*
> - *National awareness and stakeholder campaigns."* (quoted from UNDP web-site)

At the highest level there is a UN ICT Task Force with particular emphasis on applying technology to assist in meeting the Millennium Development Goals.

Apart from these general strategies some of the donors have more ICT guidelines and implementation approaches. The World Bank "Task Managers ICT Toolkit" is an exemplary

document of its kind, providing practical and sensible guidance and case study examples. Amongst bilateral donors the DAC indicates that NORAD (Norway) has *"Guidelines for assessing how ICT should be integrated"* and the Royal Danish Department of Foreign Affairs has the *"Use of ICT integrated in aid management guidelines"*. The UK Department for International Development has published in 1998 *Good Practice in developing sustainable information systems* (and supporting guides).

These various guidance documents deal with ICT as a general practice independent of areas of implementation. It is further understood that contractors and executing agencies will apply internationally recognised ICT standards or good practices particularly in information systems development, and of course in implementation of communications networks where national and international regulation and standards apply.

5.2 Bio-informatics Guidance of Donors

No major donor seems to have guidelines that are specific to biodiversity information management or "bio-informatics". Some environmental agencies and NGOs have from time to time developed some guidance materials such as the "Guidelines for Biodiversity Information Management" prepared under the UNEP "Biodiversity Data Management" (BDM) project in the late 1990s, and a suite of recommendations on standards for environmental data content and data exchange formats that can be found in reports of the ADB Sub-regional Environmental Management Information System (SEMIS) project (for the Mekong Subregion). General guidance has also been published by UNEP-WCMC as the Handbooks on Biodiversity Information Management Series, and rather similar materials from the Biodiversity Conservation Information System (BCIS) Consortium. Various "toolkits" related to the implementation of biodiversity Clearing House Mechanisms have relevant content, for example:

- The CBD Clearing-House Mechanism and Biosafety Clearing-House toolkits
- The Netherlands CHM National Focal Point toolkit
- The European Community Clearing-House Mechanism toolkit
- The Global Biodiversity Information Facility Portal toolkit.

Reference should also be made to the others of the suite of documents that have arisen out of the GEF Funded Building IABIN Project, particularly:

Document 3 - Linking biodiversity information with non-biological networks

Document 4 - Standards and practices for sharing GIS-based information

Document 6 - National strategies for effective biodiversity information management

Document 7 - Taxonomic authority archives, networks and collections

Document 8 - International initiatives in biodiversity vocabularies and thesauri

Document 10a - Review of international initiatives in metadata management

Document 10b - Review of experience in developing interoperable systems for international data management and sharing

These along with several separately commissioned studies (e.g. McClarty, 2003, and Abreu, 2000), while aimed generally at IABIN have relevant recommendations on standards and approaches that are widely applicable to biodiversity information networks and projects in other settings.

ANNEX 1 – References

DAC 2005 Financing ICTs for Development – Efforts of DAC Members, OECD Development Assistance Committee, 2005

Abreu 2000 Abreu, V.J. Final Report Harmonizing Metadata Initiatives throughout IABIN, October 2000

McClarty 2003 McClarty, D. IABIN Portal Architecture, IABIN GEF PDF Project Report, Version 0.5, July 2003

ICT Toolkit 2003 Task Mangers' ICT Toolkit, The World Bank Washington DC, 2003

Sandbrook 1994 Sandbrook, R. Annual Report 1994-1995, International Institute for Environment and Development, London, UK, 1994

ANNEX 2 – List of Abbreviations

ABCD	Access to Biological Collections Data
ADB	Asian Development Bank
BCIS	Biodiversity Conservation Information System
BDM	Biodiversity Data Management
BDP	Biological Data Profile
CABI	CAB International
CBD	Convention on Biological Diversity
CERES	California Environmental Resources Evaluation System
CGIAR	Consultative Group on International Agricultural Research
CHM	Clearing-House Mechanism
CIESIN	Center for International Earth Science Information Network
CODATA	ICSU Committee on Data for Science and Technology
CSA	[A company specialising in bibliographic and text data services]
DAC	OECD Development Assistance Committee
DC or DCMI	The “Dublin Core”
DiGIR	Distributed Generic Information Retrieval
DXF	AutoCAD Digital Exchange Format
ECAT	Electronic Catalogue of Names of Known Organisms
EEA	European Environment Agency
EML	Ecological Metadata Language
EUNIS	European Nature Information System

FAO	Food and Agriculture Organisation
FGDC	US Federal Geographic Data Committee
GBIF	Global Biodiversity Information Facility
GEMET	General Multilingual Environmental Thesaurus
GEMS	Global Environmental Monitoring System
GIS	Geographic Information Systems
GloBIS	Global Butterfly Information System
GML	Geography Markup Language
IABIN	Inter-American Biodiversity Information Network
ICSU	International Council for Science
ICT	Information and Communication Technologies
ICTD	Information and Communications Technologies for Development
IDS	Internet Data Services
ILDIS	International Legume Database & Information Service
IMS	Interactive Mapping Services
INFOTERRA	[A UNEP global information exchange programme]
IS	Information System
ISO	International Organization for Standardisation
IT	Information Technology
ITIS	Integrated Taxonomic Information System
IUCN	The World Conservation Union
JDBC	Java Database Connectivity

KWIC	KeyWords in Context
MEAs	Multinational environmental agreements
NBII	US National Biodiversity Information Infrastructure
NGO	Non Governmental Organization
NORAD	Norwegian Agency for Development Cooperation
ODBC	Open Database Connectivity
OECD	Organisation for Economic Co-operation and Development
OGC	Open GIS Consortium
PDF	Adobe Portable Document Format
PRS	Poverty Reduction Strategies
SEMIS	ADB Sub-regional Environmental Management Information System
SINGER	CGIAR System-wide Information Network for Genetic Resources
SOAP	Simple Object Access Protocol
SQL	Structured Query Language
TAA	Taxonomic Authority Archives
TDWG	Taxonomic Databases Working Group
UNDP	United Nations Development Programme
UNEP	United Nations Environment Programme
UNEP-GRID	UNEP Global Resource Information Database
UNEP-WCMC	UNEP World Conservation Monitoring Centre
W3C	World Wide Web Consortium
WFS	Web Feature Services

WMS	Web Map Services
WS	Web Services
WWW	World Wide Web
XML	Extensible Markup Language

ANNEX 3 – Annotated References to Standards and Practices

This listing of references to standards and practices was assembled in early 2005. It should be noted that ICT technology and related standards evolve rapidly, and the associated URLs change very frequently, so this list should be considered a starting point, but will become out-of-date rapidly.

Short Name	Full Name	Description	URL for more information
ICT – general			
ANSI Z39.50	American National Standards Institute, Standard Z39.50	A protocol defining a standard way for two computers to communicate for the purpose of information retrieval.	www.niso.org/z39.50/z3950.html
JDBC	Java Database Connectivity	A standard SQL database access interface.	java.sun.com/products/jdbc
ODBC	Open Database Connectivity	A (Microsoft) database programming interface to access databases on a network.	www.webopedia.com
SQL	Structured Query Language	A standard language for accessing and manipulating databases.	www.w3schools.com.soap
XML	Extensible Markup Language	A markup language used to describe documents to enable interoperability between computers.	www.w3.org/XML
SOAP	Simple Object Application Protocol	A protocol to let applications exchange information over HTTP.	www.w3schools.com.soap

Short Name	Full Name	Description	URL for more information
DCMI	Dublin Core Metadata Initiative	An open forum engaged in the development of interoperable online metadata standards.	http://www.dublincore.org
FGDC	US Federal Geographic Data Committee Content Standard for Digital Geospatial Metadata	A standard for specifying the description (metadata) of geo-spatial datasets (maps). (The FGDC also maintains standards for geo-spatial data exchange)	www.fgdc.gov
Geographic Information Systems			
GSDI	Global Spatial Data Infrastructure Association	The purpose of the organization is to promote international cooperation and collaboration in support of local, national and international spatial data infrastructure developments that will allow nations to better address social, economic, and environmental issues of pressing importance.	www.gsdi.org
OGC	Open Geospatial Consortium	A non-profit, international voluntary organisation concerned with development of geospatial standards.	www.opengeospatial.org
GML	Geography Markup Language	XML based, it is an extension designed to facilitate the exchange and interoperability of geo-spatial documents.	opengis.net/gml

Short Name	Full Name	Description	URL for more information
ISO 19115	Interantional Organization for Standardization – Standard 19115	A geospatial metadata standard.	www.isotc211.org
IMS	Interactive Mapping Services	Generic term for services that are compliant with OGC standards and guidelines.	www.opengeospatial.org
WMS	Web Map Services	Part of the more general OGC Web Services Common Specification defining ways to share geo-spatial data.	www.opengeospatial.org
WFS	Web Feature Services	Part of the more general OGC Web Services Common Specification defining ways to share geo-spatial data.	www.opengeospatial.org
E00	ArcINFO export files	Commonly used geospatial data exchange format – originally the “export” formerly from proprietary ESRI Arc/INFO.	www.esri.com
.shp	“shape”	Commonly used geospatial data exchange format – originating from proprietary ESRI ArcView, but now openly used.	www.ci.bakersfield.ca.us/gis/tutorials/tutorial_04/
DXF	AutoCAD Digital Exchange Format	Commonly used drawing or cartographic data exchange format used by the proprietary AutoCAD system.	www.autodesk.com/techpubs/autocad/acadr14/dxf/

Short Name	Full Name	Description	URL for more information
Biodiversity Information			
BDP	FGDC Content Standard for Digital Geospatial Metadata - Part 1: Biological Data Profile	Prepared by the Biological Data Working Group of the FGDC it extends the FGDC metadata standard to be more useful for biological map data.	www.fgdc.gov
EML	Ecological Metadata Language	A metadata standard produced for the Knowledge Network for Biocomplexity, based on prior work done by the Ecological Society of America. EML is implemented as a series of XML document types that can by used in a modular and extensible manner to document ecological data.	knb.ecoinformatics.org/software/eml
Xanthoria		A query system for ecological metadata (EML) This metadata query system provides a web-accessible method for querying a number of structurally different, remote metadata storage facilities simultaneously.	ces.asu.edu/bdi/Subjects/Xanthoria
DwC	The Darwin Core metadata standard	A metadata standard for describing biological museum specimen data. It is an extension of the "Dublin Core", used by "The Species Analyst" search tool.	speciesanalyst.net/docs/dwc/

Short Name	Full Name	Description	URL for more information
ABCD	Access to Biological Collections Data	An evolving comprehensive standard for the access to and exchange of data about specimens and observations. It is being developed by a joint initiative of the Taxonomy Database Working Group (TDWG) and CODATA.	www.bgbm.org/TDWG/CODATA/Schema
DiGIR	Distributed Generic Information Retrieval	DiGIR is a protocol and a set of tools for linking a community of independent databases of natural history collection data, into a single, searchable "virtual" collection. It has replaced Z39.50 in The Species Analyst and has been adopted by The Global Biodiversity Information Facility (GBIF), and, the European Network for Biodiversity Information (ENBI).	www.specifysoftware.org/Informatics/informaticsdigir
ECAT	Electronic Catalogue of Names of Known Organisms	A programme of GBIF which is working towards an electronic catalogue of the names of known organisms. It aims to provide content infrastructure to enable searches across multiple information domains.	www.gbif.org/prog/ecat

Short Name	Full Name	Description	URL for more information
ITIS	Integrated Taxonomic Information System	ITIS aims to provide authoritative taxonomic information on plants, animals, fungi, and microbes - an easily accessible database with reliable information on species names and their hierarchical classification. Currently coverage for some taxa is global, and for others it is confined to North America.	www.itis.usda.gov
Species 2000	Species 2000	Species 2000 has the objective of enumerating all known species of organisms on Earth (animals, plants, fungi and microbes) as the baseline dataset for studies of global biodiversity. It works through an array of participant global species databases covering each of the major groups of organisms, using a consistent taxonomic system.	www.sp2000.org/

Short Name	Full Name	Description	URL for more information
CBD Controlled Vocabulary	Convention on Biological Diversity Controlled Vocabulary	The CBD Controlled Vocabulary was developed by the Convention on Biological Diversity Secretariat to provide a list of terms that could be used as descriptors for the Convention's web site including the Clearing-House Mechanism (CHM).	www.biodiv.org/doc/cbd-voc.aspx
UNEP EnVoc	United Nations Environment Programme Multilingual Thesaurus of Environmental Terms	EnVoc is a multilingual thesaurus with a controlled and structured vocabulary for use in indexing, storing and retrieval of environmental information.	www.earthprint.com/cgi-bin/ncommerce3/ProductDisplay?prrfnbr=25383&prmenbr=27973
FAO AgroVoc	Food and Agriculture Organization, Agricultural Thesaurus	The AgroVoc thesaurus of the Food and Agriculture Organisation (FAO) is designed to cover the terminology of all subject fields of agriculture, forestry, fisheries, food and related domains, in order to catalogue documents.	www.fao.org/agrovoc/
GEMET 2001:	The General Multilingual Environmental Thesaurus	GEMET was developed by the European Environment Agency (EEA) to serve the needs of environmental information systems. It has a core terminology of 5,400 generalised environmental terms and their definitions in 19 European languages.	www.eionet.eu.int/GEMET

Short Name	Full Name	Description	URL for more information
CIESIN Indexing Vocabulary	CIESIN Indexing Vocabulary	The CIESIN Indexing Vocabulary was developed to index data resources and datasets related to the human interactions in global change. The database is available for on-line interrogation in English only.	sedac.ciesin.columbia.edu/metadata/vocab
Biocomplexity Thesaurus	NBII/CSA Biocomplexity Thesaurus	The Biocomplexity Thesaurus was developed in 2002-3 through a partnership between the US National Biodiversity Information Infrastructure (NBII) project and CSA. It represents a merger and reconciliation of five biology-related thesauri.	http://knb.ecoinformatics.org/index.jsp
WCPA core dataset	World Commission on Protected Areas core data set specification	A controlled common definition for core data required for describing protected areas for the World Database on Protected Areas.	parksdata.conserveonline.org/website/gis_prod/data/IMS/WDPA_viewer/English/WDPA2005.html
IUCN categories	IUCN – protected area management categories	A widely recognized standard way of classifying the management level of a protected area.	www.unep-wcmc.org/protected_areas/data/un_eintro2.htm
BCIS - Framework	Biodiversity Conservation Information System, Framework for Information Sharing	An eight volume handbook series providing general guidance on biodiversity information management.	www.biodiversity.org

Short Name	Full Name	Description	URL for more information
GBIF Metadata Standard	GBIF Metadata Standard, GBIF Secretariat	Describes the metadata standards for publication of data within the GBIF Network.	www.gbif.org
WCMC Handbooks	World Conservation Monitoring Centre Handbooks on Biodiversity Information Management	A seven volume handbook series providing general guidance on biodiversity information management (plus an overview volume), published in 1998.	www.unep-wcmc.org

Attention is drawn to the work of the Taxonomy Database Working Group (TDWG) Subgroup on Biological Collection Data that has compiled an extensive "*Reference List of Standards, Information Models, and Data Dictionaries for Biological Collections*". This listing provides references to a number of more specialized standards relevant of biological specimens and taxonomy data exchange. It can be found at www.bgbm.org/TDWG/acc