Data Management Manual

Biodiversity Information for Development and Environmental Resilience in Southwestern Ethiopia (BIDERSE)

December 2017

This programme is funded by the European Union

Introduction

Data management‖ embraces the full spectrum of activities involved in handling data. It focuses on:

- **Policy and Administration**
  - data policy
  - roles and responsibilities
    - data ownership
    - data custodianship
- **Collection and Capture**
  - quality data
  - data documentation and organization
    - dataset titles and file names
    - file contents
    - metadata
  - data standards
  - data life-cycle control
    - data specification and modeling (database design)
    - database maintenance
    - data audit
    - data storage and archiving
- **Longevity and Use**
  - data security
  - data access, data sharing, and dissemination
  - data publishing

Biodiversity Data Policy and Administration

Data Policy

Ethiopia has no biodiversity data policy and sharing strategy. However the biodiversity professionals with the BIDERSE data initiative, biodiversity data policy will be developed and endorsed by the government. A sound data policy defines strategic long-term goals for data management in all aspects of a project, agency, or organization (Burley and Peine 2007). A data policy is a set of high-level principles that establish a guiding framework for data management. A data policy can be used to address strategic issues such as data access, relevant legal matters, data stewardship issues and custodial duties, data acquisition, and other issues (Burley and Peine 2007).

To provide a high-level framework, a data policy should be flexible and dynamic. This allows a data policy to be readily adapted for unanticipated challenges, different types of projects, and potentially opportunistic partnerships while still maintaining its guiding strategic focus (Burley and Peine 2007). Issues to be considered when establishing a data policy include:

- Cost: Consideration should be given to the cost of providing data versus the cost of providing access to data. Cost can be both a barrier for the user to acquire certain datasets, as well as for the provider to supply data in the format or extent requested.

- Ownership and Custodianship: Data ownership should be clearly addressed (Burley and Peine 2007). Intellectual property rights can be owned at different levels; e.g. a merged dataset can be owned by one organization, even though other organizations own the constituent data. If the legal ownership is unclear, the risk exists for the data to be improperly used, neglected, or lost. See below for more discussion of Data Owner and Data Custodian roles.

- Privacy: Clarification of what data is private and what data is to be made available in the public domain needs to occur. Privacy legislation normally requires that personal information be protected from others. Therefore clear guidelines are needed for personal information in datasets.

- Liability: Liability involves how protected an organization is from legal recourse. This is very important in the area of data and information management, especially where damage is caused to an individual or organization as a result of misuse or inaccuracies in the data. Liability is often dealt with via end-user agreements and licenses. A carefully worded disclaimer statement can be included in the metadata and data retrieval system so as to free the provider, data collector, or anyone associated with the dataset of any legal responsibility for misuse or inaccuracies in the data (Burley and Peine 2007).

- Sensitivity: There is a need to identify any data which is regarded as ―sensitive.‖ Sensitive data is any data which if released to the public, would result in an ―adverse effect‖ (harm, removal, destruction) on the taxon or attribute in question or to a living individual. A number of factors need to be taken into account when determining sensitivity, including type and level of threat, vulnerability of the taxon or attribute, type of information, and whether it is already publicly available (Chapman and Grafton 2008).

- Existing Law & Policy Requirements: Consideration should be given to laws and policies related to data and information that apply to agencies or multi-agency efforts. Existing legislation and policy requirements may have an effect on a project‗s data policy. A list of laws, policies, and directives related to data and information in the Federal Government is provided in Appendix B.

Roles and Responsibilities

Data management is about individuals and organizations as much as it is about information technology, database practices, and applications. In order to meet data management goals and standards, all involved in a project must understand their associated roles and responsibilities .

The objectives of delineating data management roles and responsibilities are to : clearly define roles associated with functions, establish data ownership throughout all phases of a project, install data accountability, and ensure that adequate, agreed-upon data quality and metadata metrics are maintained on a continuous basis.

Data Ownership

Data owners should involve in the critical data usage steps. A key aspect of good data management involves the identification of the owner(s) of the data. Data owners generally have legal rights over the data, along with copyright and intellectual property rights. This applies even where the data is collected, collated, or disseminated by another party by way of contractual agreements, etc. Data ownership implies the right to exploit the data, and in situations where the continued maintenance becomes unnecessary or uneconomical, the right to destroy it. Ownership can relate to a data item, a merged dataset or a value-added dataset.

It is important for data owners to establish and document the following (if applicable) : the ownership, intellectual property rights and copyright of their data, the statutory and non-statutory obligations relevant to their business to ensure the data is compliant, the policies for data security, disclosure control, release, pricing, and dissemination, and the agreement reached with users and customers on the conditions of use, set out in a signed memorandum of agreement or license agreement, before data is released.

Data Custodianship

Data custodians are established to ensure that important datasets are developed, maintained, and are accessible within their defined specifications. Designating a person or agency as being in charge with overseeing these aspects of data management helps to ensure that datasets do not become compromised. How these aspects are managed should be in accordance with the defined data policy applicable to the data, as well as any other applicable data stewardship specifications (Burley and Peine 2007). Some typical responsibilities of a data custodian may include (Burley and Peine 2007): adherence to appropriate and relevant data policy and data ownership guidelines, ensuring accessibility to appropriate users, maintaining appropriate levels of dataset security, fundamental dataset maintenance, including but not limited to data storage and archiving, dataset documentation, including updates to documentation, and assurance of quality and validation of any additions to a dataset, including periodic audits to assure ongoing data integrity.

Custodianship is generally best handled by a single agency or organization that is most familiar with a dataset‘s content and associated management criteria. For the purposes of management and custodianship feasibility in terms of resources (time, funding, hardware/software), it may be appropriate to develop different levels of custodianship service (Burley and Peine 2007), with different aspects potentially handled by different organizations.

Specific roles associated with data custodianship activities may include Project Leader, Data Manager

GIS Manager, IT Specialist, Database Administrator, and Application Developer.

Collection and Capture

Data Quality

Quality as applied to biodiversity data has been defined as fitness for use or potential use. Many data quality principles apply when dealing with species data and with the spatial aspects of those data.

These principles are involved at all stages of the data management process, beginning with data collection and capture. A loss of data quality at any one of these stages reduces the applicability and uses to which the data can be adequately put (Chapman 2005a). These include: data capture and recording at the time of gathering, data manipulation prior to digitization (label preparation, copying of data to a ledger, etc.), identification of the collection (specimen, observation) and its recording, digitization of the data, documentation of the data (capturing and recording the metadata), data storage and archiving, data presentation and dissemination (paper and electronic publications, web-enabled databases, etc.), and using the data (analysis and manipulation).

All of these affect the final quality or fitness for use‖ of the data and apply to all aspects of the data.
Data quality standards may be available for accuracy, precision, resolution, reliability, repeatability, reproducibility, currency, and relevance, ability to audit, completeness, and timeliness.

Quality control (QC) is an assessment of quality based on internal standards, processes, and procedures established to control and monitor quality, while quality assurance (QA) is an assessment of quality based on standards external to the process and involves reviewing of the activities and quality control processes to insure final products meet predetermined standards of quality (Chapman 2005a, National Land & Water Resources Audit 2008). While quality assurance procedures maintain quality throughout all stages of data development, quality control procedures monitor or evaluate the resulting data products.

Although a data set containing no errors would be ideal, the cost of attaining 95%-100% accuracy may outweigh the benefit. Therefore, at least two factors are considered when setting data quality expectations: frequency of incorrect data fields or records, and significance of error within a data field.

Errors are more likely to be detected when dataset expectations are clearly documented and what constitutes a significant ‘error is understood. The significance of an error can vary both among datasets and within a single dataset. For example, a two-digit number with a misplaced decimal point (e.g., 99 vs. 9.9) may be a significant error while a six-digit number with an incorrect decimal value (e.g., 9999.99 vs. 9999.98), may not. However, one incorrect digit in a six-digit species Taxonomic Serial Number could indicate a different species.

QA/QC mechanisms are designed to prevent data contamination, which occurs when a process or event introduces either of two fundamental types of errors into a dataset:

1) Errors of commission include those caused by data entry or transcription, or by malfunctioning equipment. These are common, fairly easy to identify, and can be effectively reduced up front with appropriate QA mechanisms built into the data acquisition process, as well as QC procedures applied after the data has been acquired.

2) Errors of omission often include insufficient documentation of legitimate data values, which could affect the interpretation of those values. These errors may be harder to detect and correct, but many of these errors should be revealed by rigorous QC procedures.

Data quality is assessed by applying verification and validation procedures as part of the quality control process . Verification and validation are important components of data management that help ensure data is valid and reliable. The US EPA defines data verification as the process of evaluating the completeness, correctness, and compliance of a dataset with required procedures to ensure that the data is what it purports to be. Data validation follows data verification, and it involves evaluating verified data to determine if data quality goals have been achieved and the reasons for any deviations (US EPA 2002). While data verification checks that the digitized data matches the source data, validation checks that the data makes sense. Data entry and verification can be handled by personnel who are less familiar with the data, but validation requires in-depth knowledge about the data and should be conducted by those most familiar with the data.

Principles of data quality need to be applied at all stages of the data management process (capture, digitization, storage, analysis, presentation, and use). There are two keys to the improvement of data quality – prevention and correction. Error prevention is closely related to both the collection of the data and the entry of the data into a database. Although considerable effort can and should be given to the prevention of error, the fact remains that errors in large datasets will continue to exist and data validation and correction cannot be ignored (Chapman 2005a).

Documentation is the key to good data quality. Without good documentation, it is difficult for users to determine the fitness for use of the data and difficult for custodians to know what and by whom data quality checks have been carried out. Documentation is generally of two types and provision for them should be built into the database design. The first is tied to each record and records what data checks have been done and what changes have been made and by whom. The second is the metadata that records information at the dataset level. Both are important, and without them, good data quality is compromised (Chapman 2005b).

An in-depth overview of data quality principles including quality assurance, quality control, and data cleaning -- is provided by Chapman (2005a, b). Additional information is also found in the National Park Service‘s (2008) data management guidelines.

Data Documentation and Organization

Data documentation is critical for ensuring that datasets are useable well into the future. Data longevity is roughly proportional to the comprehensiveness of their documentation. All datasets should be identified and documented to facilitate their subsequent identification, proper management and effective use, and to avoid collecting or purchasing the same data more than once.

The objectives of data documentation are to : ensure the longevity of data and their re-use for multiple purposes, ensure that data users understand the content, context, and limitations of datasets, facilitate the discovery of datasets, and facilitate the interoperability of datasets and data exchange.

One of the first steps in the data management process involves entering data into an electronic system. The following data documentation practices may be implemented during database design and data entry to facilitate the retrieval and interpretation of datasets not only by the data collector, but also by those who may have future interest in the data.

Dataset Titles and File Names

Dataset titles and corresponding file names should be descriptive, as these datasets may be accessed many years in the future by people who will be unaware of the details of the project or program. Electronic files of datasets should be given a name that reflects the contents of the file and includes enough information to uniquely identify the data file. File names may contain information such as project acronym or name, study title, location, investigator, year(s) of study, data type, version number, and file type. The file name should be provided in the first line of the header rows in the file itself. Names should contain only numbers, letters, dashes, and underscores – no spaces or special characters. In general, lower-case names are less software and platform dependent and are preferred (Hook et al. 2007). For practical reasons of legibility and usability, file names should not be more than 64 characters in length and, if well-constructed, could be considerably less (Hook et al. 2007); file names that are overly long will make it difficult to identify and import files into analytical scripts (Borer et al. 2009). Including a data file creation date or version number enables data users to quickly determine which data they are using if an update to the data set is released (Hook et al. 2007).

File Contents

In order for others to use your data, they must understand the contents of the dataset, including the parameter names, units of measure, formats, and definitions of coded values. At the top of the file, include several header rows containing descriptors that link the data file to the dataset; for example, the data file name, dataset title, author, today‘s date, date the data within the file was last modified, and companion file names. Other header rows should describe the content of each column, including one row for parameter names and one for parameter units (Hook et al. 2007). For those datasets that are large and complex and may require a lot of descriptive information about dataset contents, that information may be provided in a separate linked document rather than as headers in the data file itself.

Parameters: The parameters reported in datasets need to have names that describe their contents and their units need to be defined so that others understand what is being reported. Use commonly accepted parameter names (Hook et al. 2007). A good name is short (some software is limited in the size parameter name it can handle), unique (at least within a given dataset), and descriptive of the parameter contents. It is recommended that you select parameter names that are unique in their first 7 characters (even if they are longer) (Porter 1997). Column headings should be constructed for easy importing by various data systems. Use consistent capitalization and use only letters, numerals, and underscores – no spaces or decimal characters – in the parameter name. Choose a consistent format for each parameter and use that format throughout the dataset. When possible, try to use standardized formats, such as those used for dates, times, and spatial coordinates (Hook et al. 2007).

All cells within each column should contain only one type of information (i.e., either text, numbers, etc.). Common data types include text (alphanumeric strings of text), numeric, date/time, Boolean (also called Yes/No or True/False), and comments (for storing large quantities of text) (Borer et al. 2009).

Coded Fields: Coded fields, as opposed to free text fields, often have standardized lists of predefined values from which the data provider may choose. Data collectors may establish their own coded fields with defined values to be consistently used across several data files. Coded fields are more efficient for the storage and retrieval of data than free text fields (Hook et al. 2007).

Missing Values: There are several options for dealing with a missing value. One is to leave the value blank, but this poses a problem as some software do not differentiate a blank from a zero; or, a user might wonder if the data provider accidentally skipped a column. Another option is to put a period where the number would go. This makes it clear that a value should be there, although it says nothing about why the data is missing. One more option is to use different codes to indicate different reasons why the data is missing (Porter 1997).

Metadata

Metadata, defined as data about data, provides information on the identification, quality, spatial context, data attributes, and distribution of datasets, using a common terminology and set of definitions that prevent loss of the original meaning and value of the resource. This common terminology is particularly important to biodiversity datasets because: different biodiversity projects collect dissimilar types of data and record them in various ways; occur at a variety of scales; and are dispersed globally. Without descriptive metadata, discovering that a resource exists, what data was collected and how it was measured and recorded, and how to access it would be a monumental undertaking (Kelling 2008).

Metadata in the biodiversity information domain provide (Kelling 2008): an accurate description of the data itself; a description of spatial attributes, which should include bounding coordinates for the specific project, how spatial data was gathered, limits of coverage, and how this spatial data is stored; a complete description of the taxonomic system used by the project, with references to methods employed for organism identification and taxonomic authority; and a description of the data structure, with details of how to access the data and/or how to access tools that can manipulate the data (i.e., visualizations, statistical processes, and modelling).

Several initiatives are underway that are developing discovery resources for biodiversity data and monitoring programs. These initiatives can be identified as open-ended (encompassing all biodiversity resources), or domain specific (only organizing the resources within a specific area of interest); and their foci range from a description of data generated by monitoring programs to a description of the projects or programs themselves (Kelling 2008).

Metadata standards for database content documentation and other types of biodiversity information are provided in Appendix C of this document.

Data Standards

Data standards describe objects, features, or items that are collected, automated, or affected by activities or the functions of organizations. In this respect, data need to be carefully managed and organized according to defined rules and protocols. Data standards are particularly important in any co-management, co-maintenance, or partnership where data and information need to be shared or aggregated.

Benefits of data standards include: more efficient data management (including updates and security); increased data sharing; higher quality data; improved data consistency; increased data integration; better understanding of data, and improved documentation of information resources.

When adopting and implementing data standards, consideration should be given to the different levels of standards such as, international , national, regional, local, and where possible, adopt the minimally complex standard that addresses the largest audience.

Be aware that standards are continually updated, so the necessity of maintaining compliance with as few as possible is desirable.

Data Life-cycle Control

Good data management requires the whole life cycle of data to be managed carefully. This includes : data specification and modelling, processing, and database maintenance and security, ongoing data audit, to monitor the use and continued effectiveness of existing data, archiving, to ensure data is maintained effectively, including periodic snapshots to allow rolling back to previous versions in the event that primary copies and backups are corrupted

Data Specification and Modelling

The majority of the work involved in building databases occurs long before using any database software. Successful database planning takes the form of a thorough user requirements analysis, followed by data modelling.

Understanding user requirements is the first planning step. Databases must be designed to meet user needs, ranging from data acquisition through data entry, reporting, and long-term analysis. Data modelling is the methodology that identifies the path to meet user requirements. The focus should be to keep the overall model and data structure as simple as possible while still adequately addressing project participants' business rules and project goals and objectives (Burley and Peine 2007).

Detailed review of protocols and reference materials on the data to be modelled will articulate the entities, relationships, and flow of information. Data modelling should be iterative and interactive.

The conceptual design phase of the database life cycle should produce an information/data model. An information/data model consists of written documentation of concepts to be stored in the database, their relationships to each other, and a diagram showing those concepts and their relationships. In the database design process, the information/data model is a tool to help the design and programming team understand the nature of the information to be stored in the database, not an end in itself. Information/data models assist in communication between the people who are specifying what the database needs to do (data content experts) and the programmers and database developers who are building the database (and who speak wholly different languages). Careful database design and documentation of that design are important not only in maintaining data integrity during use of a database, but are also important factors in the ease and extent of data loss in future migrations (including reduction of the risk that inferences made about the data now will be taken at some future point to be original facts). Therefore, information/data models are also vital documentation when it comes time to migrate the data and user interface years later in the life cycle of the database (Morris 2005).

Information/data models may be as simple as a written document or drawing, or may be complex and constructed with the aid of software engineering tools. Appendix D provides information about different types of data models used in the database design process.

Database Maintenance

Technological obsolescence is a significant cause of information loss, and data can quickly become inaccessible to users if stored in out-of-date software formats or on outmoded media. Effective maintenance of digital files depends on proper management of a continuously changing infrastructure of hardware, software, file formats, and storage media. Major changes in hardware can be expected every 1-2 years, and in software every 1-5 years. As software and hardware evolve, datasets must be continuously migrated to new platforms, and/or they must be saved in formats that are independent of specific platforms or software.

A database or dataset should have carefully defined procedures for updating. If a dataset is live or ongoing, this will include such things as additions, modifications, and deletions, as well as frequency of updates. Versioning will be extremely important when working in a multi-user environment (Burley and Peine 2007).

Management of database systems requires good day-to-day system administration. Database system administration needs to be informed by a threat analysis, and should employ means of threat mitigation, such as regular backups, highlighted by that analysis (Morris 2005).

Data Audit

Good data management requires ongoing data audit to monitor the use and continued effectiveness of existing data. According to Henczel (2001), data or information audit is a process that involves:

- identifying the information needs of an organization/program and assigning a level of strategic importance to those needs,
- identifying the resources and services currently provided to meet those needs,
- mapping information flows within an organization (or program) and between an organization and its external environment, and
- analysing gaps, duplications, inefficiencies, and areas of over-provision that enable the identification of where changes are necessary.

An information audit not only counts resources but also examines how they are used, by whom, and for what purpose. The information audit examines the activities and tasks that occur in an organization and identifies the information resources that support them. It examines, not only the resources used, but how they are used and how critical they are to the successful completion of each task. Combining this with the assignment of a level of strategic significance to all tasks and activities enables the identification of the areas where strategically significant knowledge is being created. It also identifies those tasks that rely on knowledge sharing or transfer and those that rely on a high quality of knowledge (Henczel 2001).

Data Storage and Archiving

Data storage and archiving address those aspects of data management related to the housing of data. This element includes considerations for digital/electronic data and information as well as relevant hardcopy data and information. Without careful planning for storage and archiving, many problems arise that result in the data becoming out of date and possibly unusable as a result of not being property managed and stored (Burley and Peine 2007). Some important physical dataset storage and archiving considerations for electronic/digital data include the following.

- Server Hardware and Software: What type of database will be needed for the data? Will any physical system infrastructure need to be set up or is the infrastructure already in place? Will a major database product be necessary? Will this system be utilized for other projects and data? Who will oversee the administration of this system?

- Network Infrastructure: Does the database need to be connected to a network or to the Internet? How much bandwidth is required to serve the target audience? What hours of the day does it need to be accessible?

- Size and Format of Datasets: The size of a dataset should be estimated so that storage space can properly be accounted for. The types and formats should be identified so that no surprises in the form of database capabilities and compatibility will arise.

- Database Maintenance and Updating: A database or dataset should have carefully defined procedures for updating. If a dataset is live or ongoing, this will include such things as additions, modifications, and deletions, as well as frequency of updates. Versioning will be extremely important when working in a multi-user environment.

- Database Backup and Recovery Requirements: To ensure the longevity of a dataset, the requirements for the backing up or recovery of a database in case of user error, software / media failure, or disaster, should be clearly defined and agreed upon. Mechanisms, schedules, frequency and types of backups, and appropriate recovery plans should be specified and planned. This can include types of storage media for onsite backups and whether off-site backing up is necessary.

Archiving of data should be a priority data management issue. Organizations with high turnovers of staff and data stored in a distributed manner need sound documenting and archiving strategies built into their information management chain. Snapshots (versions) of data should be maintained so that rollback is possible in the event of corruption of the primary copy and backups of that copy. Additionally, individuals working outside of a major institution need to ensure that their data is maintained and/or archived after they cannot store it anymore or cease to have an interest in it. Similarly, organizations that may not have long-term funding for the storage of data need to enter into arrangements with appropriate organizations that do have a long-term data management strategy (including archiving) and who may have an interest in the data (Chapman 2005a).

Data archiving has been facilitated in the past decade by the development of the DiGIR/Darwin Core, BioCASE/ABCD, and TAPIR protocols. These provide a way for an organization, program, or individual to export their database and store it in XML format, either on their own site, or forwarded to a host institution. These methods facilitate the storage of data in perpetuity and/or its availability through distributed search procedures once a host institution is identified (Chapman 2005a).

Sustained Use

Data Security

Security involves the system, processes, and procedures that protect a database from unintended activity. Unintended activity can include misuse, malicious attacks, inadvertent mistakes, and access made by individuals or processes, either authorized or unauthorized. For example, a common threat for any web-enabled system is automated software designed to exploit system resources for other purposes via vulnerabilities in operating systems, server services, or application. Physical equipment theft or sabotage is another consideration. Accidents and disasters (such as fires, hurricanes, earthquakes, or even spilled liquids) are another category of threat to data security. Efforts should be made to stay current on new threats so that a database and its data are not put at risk. Appropriate measures and safeguards should be put in place for any feasible threats (Burley and Peine 2007).

The consensus is that security should be implemented in layers and should never rely on a single method. Several methods should be used, for example: uninterruptible power supply, mirrored servers (redundancy), backups, backup integrity testing, physical access controls, network administrative access controls, firewalls, sensitive data encryption, up-to-date-software security patches, incident response capabilities, and full recovery plans. Where possible, any implemented security features should be tested to determine their effectiveness (Burley and Peine 2007).

Risk management is the process that allows Information Technology (IT) managers to balance the operational and economic costs of protective measures with gains in mission capability by protecting the IT systems and data that support their organizations' missions. Risk management encompasses three processes: risk assessment, risk mitigation, and evaluation and assessment. Minimizing negative impact on an organization and the need for a sound basis in decision making are the fundamental reasons organizations implement a risk management process for their IT systems (Stoneburner et al. 2002).

Risk assessment is the first process in the risk management methodology. Organizations use risk assessment to determine the extent of the potential threat and the risk associated with an IT system throughout its system development life cycle. The output of this process helps to identify appropriate controls for reducing or eliminating risk during the risk mitigation process. Risk is a function of the likelihood of a given threat-source's exercising a particular potential vulnerability, and the resulting impact of that adverse event on the organization. To determine the likelihood of a future adverse event, threats to an IT system must be analyzed in conjunction with the potential vulnerabilities and the controls in place for the IT system. Impact refers to the magnitude of harm that could be caused. The level of impact is governed by the potential mission impacts and in turn produces a relative value for the IT assets and resources affected (e.g., the criticality and sensitivity of the IT system components and data) (Stoneburner et al. 2002).

Risk mitigation, the second process of risk management, involves prioritizing, evaluating, and implementing the appropriate risk-reducing controls recommended from the risk assessment process. Because the elimination of all risk is usually impractical or close to impossible, it is the responsibility of senior management and functional and business managers to use the least-cost approach and implement the most appropriate controls to decrease mission risk to an acceptable level, with minimal adverse impact on the organization's resources and mission (Stoneburner et al. 2002). It seems likely that the most prudent and cost-effective approach for ensuring the security of biodiversity data (which is not particularly time-sensitive) is to maintain regular snapshots of the data in secure, offline (and off-site) repositories.

In most organizations, the information system itself will continually be expanded and updated, its components changed, and its software applications replaced or updated with newer versions. In addition, personnel changes will occur and security policies are likely to change over time. These changes mean that new risks will surface and risks previously mitigated may again become a concern. Thus, the risk management process is ongoing and evolving (Stoneburner et al. 2002).

Data Access, Sharing, and Dissemination

Data and information should be readily accessible to those who need them or those who are given permission to access them. Some issues to address with access to data and a database system include (Burley and Peine 2007):

- Relevant data policy and data ownership issues regarding access and use of data
- The needs of those who will require access to the data
- Various types and differentiated levels of access needed and as deemed appropriate
- The cost of actually providing data versus the cost of providing access to data
- Format appropriate for end-users
- System design considerations, including any data (if any) that requires restricted access to a subset of users
- Issues of private and public domain in the context of the data being collected

Liability issues should be included in the metadata in terms of accuracy, recommended use, use restrictions, etc. A carefully worded disclaimer statement can be included in the metadata so as to free the provider, data collector, or anyone associated with the data set of any legal responsibility for misuse or inaccuracies in the data.

The need for single-access or multi-user access, and subsequent versioning issues associated with multi-user access systems

Intentional obfuscation of detail to protect sensitive data (e.g. private property rights, endangered species) but still share data

Whether certain data is made available or not, and to whom, is a decision of the data owner(s) and/or custodian. Decisions to withhold data should be based solely on privacy, commercial-in-confidence, national security considerations, or legislative restrictions. The decision to withhold needs to be transparent and the criteria on which the decision is made need to be based on a stated policy position (ANZLIC Spatial Information Council 2004).

An alternative to denying access to certain data is to ―generalize‖ or aggregate it to overcome the basis for its sensitivity. Many organizations will supply statistical data which has been derived from the more detailed data collected by surveys. Some organizations will supply data that has lower spatial resolution than the original data collected to protect sensitive data. It is important that users of data be made aware that certain data has been withheld or modified, since this can limit processes or transactions they are involved in and the quality or utility of the information product produced. One remedy is for data custodians to make clear in publicly available metadata records and as explicit statements on data products that there are limitations applied to the data supplied or shown which could affect fitness for use (ANZLIC Spatial Information Council 2004).

Various national and global initiatives are currently underway to facilitate the discovery and access to data via the use of metadata (description of data), data exchange schemas (descriptions of database content structure), and ontologies (formal specifications of terms in an area of knowledge and the relationships among those terms). Appendix E provides a brief overview of some of those national/global data discovery and access initiatives as they relate to biodiversity data. Participation in these initiatives by organizations that maintain biodiversity data will contribute to increasing access and dissemination of this data for its use in conservation.

Data Publishing

Information publishing and access need to be addressed when implementing integrated information management solutions. Attention to details, such as providing descriptive data headings, legends, metadata/documentation, and checking for inconsistencies, help ensure that the published data actually makes sense, is useable to those accessing it, and that suitable documentation is available so users can determine whether the data may be useful and pursue steps to access it.

Conclusion

In Ethiopia, biodiversity data handling and sharing policy is critically important. Data management is increasingly recognized as an important component of effective data use in biodiversity conservation. Methods, best practices, and standards for management of biodiversity data have been developed by the bioinformatics community over the past fifteen years to facilitate electronic data access and use. These methods and best practices range from defining policies, roles, and responsibilities for data management; organizing, documenting, verifying, and validating data to enhance its quality; managing for the entire data life-cycle from design of a database to storage and archiving of data; to disseminating data by providing appropriate access while maintaining security of the data. As best data management practices and standards become more widely used in the management of bird monitoring data, their adoption and implementation will increase utility of this data in providing the information needed for research, management, and conservation of birds.

References

ANZLIC Spatial Information Council. 2004. Discussion Paper: Access to Sensitive Spatial Data. Online: Accessed June 2009. <http://www.anzlic.org.au/pubinfo/2399972232.html>

Borer, E.T., Seabloom, E.W., Jones, M.B. and M. Schildhauer. 2009. Some Simple Guidelines for Effective Data Management. Bulletin of the Ecological Society of America 90(2): 205-214. Online: Accessed October 2009. <http://www.esajournals.org/doi/abs/10.1890/0012-9623-90.2.205>

Burley, T.E. and J.D. Peine. 2007. NBII-SAIN Data Management Toolkit. Online: Accessed May 2009. < http://pubs.usgs.gov/of/2009/1170/>

Chapman, A. D. 2005a. Principles of Data Quality, version 1.0. Report for the Global Biodiversity Information Facility, Copenhagen. Online: Accessed May 2009. < http://www2.gbif.org/DataQuality.pdf>

Chapman, A. D. 2005b. Principles and Methods of Data Cleaning – Primary Species and Species Occurrence Data, version 1.0. Report for the Global Biodiversity Information Facility, Copenhagen. Online: Accessed May 2009. <http://www2.gbif.org/DataCleaning.pdf>

Chapman, A.D. and O. Grafton. 2008. Guide to Best Practices for Generalising Sensitive Species Occurrence Data. Copenhagen: Global Biodiversity Information Facility, 27 pp. ISBN: 87-92020-06-2. Online: Accessed May 2009. <http://www2.gbif.org/BPsensitivedata.pdf>

Henczel, S. 2001. The Information Audit as a First Step Towards Effective Knowledge Management. Information Outlook, Vol. 5, No. 6. Online: Accessed June 2009. <http://www.sla.org/content/Shop/Information/infoonline/2001/jun01/Henczel.cfm>

Hook, L.A. Beaty, T.W., Santhana-Vannan, S., Baskaran, L. and R.B. Cook. 2007. Best Practices for Preparing Environmental Data Sets to Share and Archive. Oak Ridge National Laboratory Distributed Active Archive Center, Oak Ridge, Tennessee, U.S.A. Online: Accessed October 2009. <http://daac.ornl.gov/PI/bestprac.html>

Jones, M.B. 2008. A Proposal for a Distributed Earth Observation Data Network. Presentation at: TDWG 2008, Freemantle, Australia. Online: Accessed June 2009. <http://www.tdwg.org/fileadmin/2008conference/slides/Jones_14_02_DataNetOne.ppt>

Jones, S., Ross, S. and R. Ruusalepp. 2008. Data Audit Framework: A data management toolkit for research led institutions. Presentation at: CNI Task Force Meeting, Washington DC, USA, 8th December 2008. Online: Accessed June 2009. <http://www.data-audit.eu/docs/DAF_CNI.pdf>