**Global Terrestrial Observing System**

# GTOS Central and Eastern European Terrestrial Data Management and Accessibility Workshop

## Edit Kovács-Láng

**Vácrátót, Hungary, 30 October – 4 November 2000**

**GTOS – 27**

# Contents

# Introduction

Long-term site-based ecological research and data collection are highly welcomed by ecologists in Central and Eastern Europe for several reasons:

- They compliment standard practices and existing research, especially in long term biodiversity studies.

- They are needed in joint efforts in preventing regional and global environmental degradation.

- They allow the collection of important long-term ecosystem and landscape dynamics data, which are needed for making predictions.

The main threats to the region's environment and biodiversity are eutrophication, acidification (due to air and water pollution), toxic pollution and changes in land use and climate. Although many regions of Central and Eastern Europe are heavily damaged or degraded there are still large areas of diverse ecosystem, which are rich in flora and fauna. The region has a tradition for long-term data collection, with meteorological, geological, soil, flora, fauna and vegetation surveys and mapping records dating back decades or centuries.

Site-based integrated research and monitoring have been given a considerable impetus by international scientific Programmes such as IBP (International Biosphere Program) and MAB (Man and Biosphere). Biological productivity, ecosystem function and human ecological impact are usually monitored at these sites. Data collected at terrestrial sites include biodiversity, primary and secondary productivity, nutrient cycling, and ecosystem management and restoration observations. Other terrestrial studies include vegetation dynamics, bird migration and the response of populations and communities to disturbances (e.g. pollution, climate and land use changes). The main research themes in aquatic systems are hydrobiology (plankton, macrophytes, benthos, fish stock, microbial loop), biodiversity, and water and sediment quality, as well as nutrient loading from catchments, eutrophication and acidification impacts, and the function and sustainable management of estuaries, wetlands, fishponds and reservoir ecosystems.

The research is usually programmed and carried out under extremely poor financial conditions by enthusiastic and keen members of scientific institutions (national academies and universities) and state institutes and services (natural history museums, state forestry institutes, national meteorological and hydrological services and national parks).

University involvement also creates the prospect of student participation in site research. Most sites have undergraduate and postgraduate students working on their diplomas and Ph.D. theses. There is a great quantity of data accumulated but in most cases the information is not computerized and properly managed making it partly or completely inaccessible. The region therefore needs training and development as well as financial support in information technology.

**The Workshop**

On the request of participants of the Central and East European (CEE) regional ILTER and GTOS meetings held in Budapest (Hungary, 1999) and Nitra (Slovakia, 2000) and with the support of ILTER, GTOS and the Hungarian Academy of Sciences a five-day workshop was organized in Vácrátót (Hungary) at the Institute of Ecology and Botany of the Hungarian Academy of Sciences, where a network of computers were made available to participants.

The workshop was a combination of lectures, demonstrations, discussions and hands-on computer lab experience focusing on data management and accessibility issues. It had both training and research objectives. The latter was the involvement of participants in integrating the CEE LTER (Long-Term Ecological Research) sites into the metadata networks of ILTER (International Long-Term Ecological Research) and the GTOS TEMS database.

Inclusion of the CEE LTER sites into these networks was a powerful demonstration to workshop participants of the value of metadata and the potential for international collaboration. Participants were asked to bring datasets from their research sites to use in laboratory exercises. The training was provided in a helpful and collaborative atmosphere.

Participants presented short informal reports and slide demonstrations on the characteristics of their sites. They provided information on their sites monitoring and research activities and in some cases provided information on the type (paper, computerized, etc.) and management of site data.

Data sources useful for both LTER and GTOS were identified and participants made proposals to finish developing the metadata and to bring databases online to meet GTOS objectives. After the workshop participants dedicated some time in completing GTOS questionnaires with their site data.

**Instructors**

The workshop was lead by three members of the US LTER Data Management Group, who are nationally recognized for their expertise. The lecturers were:

**Dr John Porter**
US Virginia Coastal Reserve LTER Data Manager
Department of Environmental Sciences
University of Virginia, USA

**Dr Peter McCartney**
US Central Arizona-Phoenix LTER Data Manager
Center for Environmental Studies
Arizona State University, USA

**Dr Kristin Vanderbilt**
US Sevilleta LTER Data Manager
Department of Biology
University of New Mexico, USA

**Participants**

Participants of the workshop were young scientists from CEE countries (Czech Republic, Hungary, Poland, Romania, and Slovakia) who are involved and/or responsible for the data sets of LTER sites or scientific institutes and Universities.


Czech Republic

**Ing. Zdenek Fajfr** (data manager), Krkonose National Park and Biosphere Reserve LTER site.

**Ing. Vaclav Hauser** (researcher), Trebon Basin Protected Landscape Area and Biosphere Reserve LTER site.

Hungary

**Gabor Varbiro** (data manager), Sikfokut Oak Forest LTER site.

**Sandor Barabas** (researcher), KISKUN forest- steppe LTER site.

**Barbara Lhotsky** (researcher), KISKUN forest- steppe LTER site.

**Janos Garadnai** (researcher), KISKUN forest- steppe LTER site.

Poland

**Przemyslaw Wasiak** (researcher), Bieszczady Mountains LTER site.

**Stanislaw Twerek** (researcher), Institute of Nature Conservation Polish Acad.Sci..

Romania

**Dan Cogalniceanu** (researcher), Department of Ecology, Bucharest University, Retezat Mountain System Biosphere Reserve.

**Ana Maria Benedek** (Ph.D.student at Bucharest University), Danube Delta Biosphere Reserve.

Slovakia

**Henrik Kalivoda** (researcher), Institute of Landscape Ecology of Slovak Acad. Sci. Biely Vah Research Site.

**Gabriel Bugar** (researcher), Institute of Landscape Ecology of Slovak Acad. Sci. Polana Research Site.

**Robert Kanka** (researcher), Institute of Landscape Ecology of Slovak Acad. Sci. Biely Vah Research Site.

# 1. Workshop Programme

The workshop Programme covered the most important themes of information management, including the organization of data systems, quality assurance and control, the use of metadata and data archiving. Information was also provided on GTOS and the TEMS database. All workshop participants were provided with a copy of the book Ecological Data: Design, Management, and Processing, Methods in Ecology Series by **W. K. Michener and J. W. Brunt, Eds. (2000)**, *Blackwell Science*.

This book was the source for the majority of the lectures given during the workshop and is the source of the tables and figures within this report.

The lecture topics were as follows:

**General Data Management Principles**

A sound philosophy for data management in ecology research is that it should be people-oriented. It must offer practical solutions to ecologists and place training and education above technical sophistication and complexity. It should provide well managed high quality ecological data which is easy to access but is also secure and durable in time.
Two basic principles can facilitate successful data management:

- Start small, keep it simple, and be flexible;

- Involve scientists in the data management process.

Ecologists have the responsibility for defining scientific objectives for the data management system. They must establish research priorities and determine resource allocation. The data management system at a site should be developed from a research/monitoring perspective and must reflect the objectives and priorities of the research/monitoring programme.

Data management can provide added value to a project's database by assuring that archived data are of an acceptable quality and can be retrieved and understood by future investigators. Data management systems for ecological research have mainly evolved over the last 30 years out of large projects like the IBP and the US Long-Term Ecological Research projects, with systems becoming broader and more complex. Many future advances in ecology are likely to hinge on the ability to integrate diverse computerized data sets. Carefully considered and applied data management practices are therefore required.

## 1.1 Data Management Implementation

The reasons for implementing a data management system are to:

- Formalize the procedures used to acquire and maintain the products (e.g. data) of a research project so that they may extend beyond the lifetime of the original investigator(s).

- Facilitate the resurrection of currently inaccessible historical data.

- Support the preparation of data sets for peer-review, publication in data journals, or submission to a data archive.

- Provide access to data sets that are commonly used by more than one investigator on a project.

- Provide access to data sets by the broader scientific community (e.g. via the world wide web).

- Reduce the time and effort spent by researchers in locating, accessing, and analysing data, thereby increasing the time available to generate results.

- Increase research capacity by allowing the analysis of broader scale questions, which would be impossible without the organization of a data management system (e.g. the need to integrate multiple dissimilar data sets).

- Incorporate data from automated acquisition systems into ongoing analytical efforts such as ecological modeling.

To manage ecological data a logical organized structure for the data management system is essential.

The primary goal of a system is to provide the best quality data possible within a reasonable budget. The second system requirement is that data is easily accessible by investigators. Thirdly (but just as important as access) short- and long-term security of data through data archiving needs to be provided. The fourth requirement is that computational support and quality assurance is given to users of the system.

The protocols and computational infrastructure required to achieve these goals vary considerably within the research community and can range from relatively simple to extremely sophisticated implementations. Technology affects the way data management is carried out but it should not affect the principles that are applied.

## 1.2 Data Management System Components

Research data management in ecology (as in other disciplines) involves acquiring, maintaining, manipulating, analysing, archiving and facilitating access to both data and results.

Six components or activities are fundamental in the implementation of a data management system:

i. An inventory of existing data and resources will have to be compiled, and priorities for implementation set.

ii. A logical structure needs to be developed so that data can be organized within data sets. This will facilitate data storage, retrieval and manipulation.

iii. Procedures for data acquisition, quality assurance and quality control (QA and QC) will need to be established.

iv. Data set documentation protocols will need to be developed. Including the adoption or creation of metadata content standards and procedures for the recording of metadata.

v. Measures for the storage and maintenance of printed and electronic data will have to be developed.

vi. Finally, an administrative structure will need to be developed so responsibilities are clearly defined.

## 1.3 Inventory of Existing Data Sets

It is essential to have an inventory of past, present and also future data sets and resources. Research programmes and activities, data type and quantity, staff, facilities and financial standing should all be included. Once the inventory is completed the objectives for each data set will need to be determined. These objectives should then be made available to the project investigators with as much supporting information as possible. This information will be invaluable for initiating the data management system and will be requested, used and revised countless times.

## 1.4 Data Design, Organization and Manipulation

Data design (the 'logical' organization of data into tables) should primarily reflect the experimental design in use. Decisions about data design will be necessary before data is collected to allow field and laboratory data sheets to be constructed. The completed design can be transferred directly to data entry tools to aid data collection, facilitate analysis by statistical software, support metadata development and to structure the data set for archiving. Designing the precise structure of data tables for use in database management systems (DBMS) is called normalization. The use of this design process on ecological data requires *detailed knowledge of the data* to avoid costly mistakes.

Organization of the Data Sets

Commercial DBMS software may be inappropriate for initial file organization and management. A file management systems can be initially used to organize data sets. The data, metadata, analytical and management tools are all stored in a logically built structure (see Figure 1).

Data manipulation and maintenance

Special software is needed to store, manipulate and maintain scientific data. There are several software packages available to data managers that are designed specifically for the manipulation and analysis of scientific data (e.g. SAS and S+). There are also other types of software that are designed for general database management systems (e.g. Foxpro, Oracle and MS SQL Server).

One of the most important decisions by the data manager will be whether to invest time and money in a DBMS. The decision will depend on many factors, including the type and the intended use of the data. Geographic information systems (GIS) represent a special type of DBMS that combines spatial mapping and analytical capabilities with relational database functions.
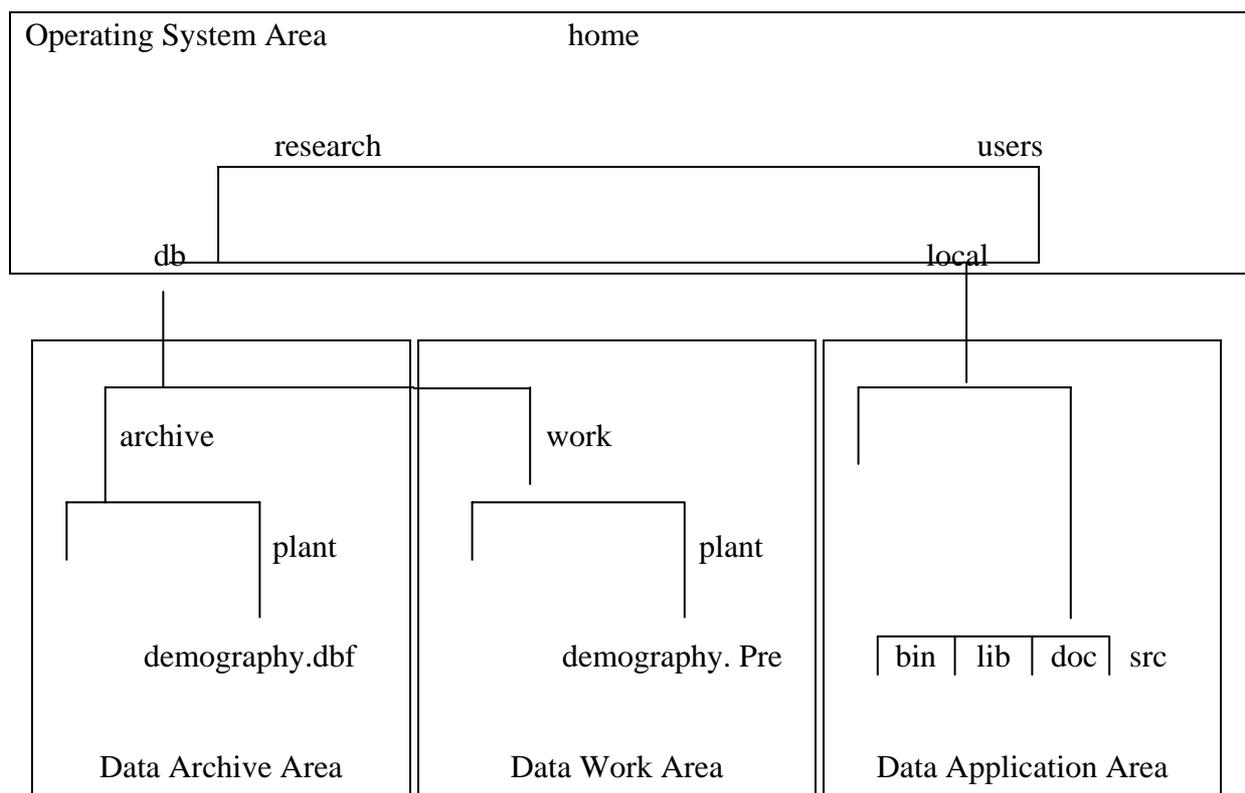
*Figure 1* *An example of the use of a computer system directory structure in the data management system. The dashed lines indicate the physical separation on the hard disk while the solid lines represent virtual links between directories.*

GIS should be viewed as part of the overall data management system, particularly if the research has a strong spatial component. Considerations, including the use of accepted practices in documenting and storing coordinates, should be made in all data management systems for the potential use of the data in GIS.

A Internet-based information system may be needed to deliver data and data derived products. Such a system is also likely to be required for advanced query, integration and analysis functions. Information systems expand the capabilities of data and database management systems by providing additional integrative services and access. A theoretical example is the development of a database that integrates other databases and then provides the integrated product via the world wide web.

**1.5 Data Acquisition and QualityAssurance (QA) and Quality Control (QC)**

Ecological data are usually collected using paper and then transferred on to computer for analysis and storage. There is also an increase in the use of automated data collectors which record data directly onto a computer. Normally, the fewer the times data are transferred, the lower will be the errors introduced. Ideally the transfer of data from one form to another should only occur once, with appropriate QA and QC procedures during the process.

## Paper data forms

Paper is one of the primary tools used to acquire data. The development of suitable data sheets is therefore important.

Prior to data collection each page should contain basic information on the study for which it is to be used. Each page should be numbered and contain a section for comments and metadata (like date, weather conditions, names of the collectors, etc). Where possible, each page should reflect the design structure of the data set. This single requirement will greatly reduce entry errors and greatly facilitate QA and QC procedures.

## Tape recorders

Tape recorders can provide a high quality, efficient method of collecting data that can be easily operated by a single person in the field. The recorded observations can then be transcribed to paper or entered into computer files under the more favourable conditions of the laboratory.

## Hand-held computers

Acquiring data with hand-held computers is probably the best way of obtaining high quality data especially if combined with point-of-entry data quality checks.

## Automated data acquisition systems

After the data is acquired automatically it is subsequently downloaded or transferred to another computer for processing. Small data loggers are also widely used and are an efficient way of collecting continuous sensor data. On the negative side, they usually require considerable programming and electronic expertise and must be routinely downloaded as their physical memory is usually limited.

## Quality assurance and quality control

QA and QC mechanisms are designed to prevent data contamination (the introduction of errors into a data set). Commission and omission are the two fundamental types of errors that can occur. Commission errors include incorrect or inaccurate data. These can be derived from a variety of sources including malfunctioning instrumentation and data entry and transcription errors. Such errors are common and are relatively easy to identify. Errors of omission, on the other hand, may be much more difficult to identify. Omission errors frequently include inadequate documentation of legitimate data values (affecting the way a given data value is interpreted).

QC procedures can be very effective at reducing errors of commission. Control mechanisms are usually constructed in advance and applied during the data acquisition and transcription process to prevent corruption and contamination.

QA procedures are used to identify errors of omission and commission. QA mechanisms can be applied after the data have been collected, transcribed, entered in a computer and analysed.

Combined quality control and assurance procedures for ecological data include four actions which range from relatively simple and inexpensive to sophisticated and costly. The four actions are:

i. Defining and enforcing standards for formats, codes, measurement units and metadata.

ii. Checking for unusual or unreasonable patterns in the data.

iii. Checking for comparability of values between data sets.

iv. Assessing overall data quality.

Most QA/QC is typically in category 1. The most basic element of QA/QC begins with data design and continues through to data acquisition, metadata development, and preparation of data and metadata for submission to a data archive. Examples of QA/QC for each of these stages are listed in Table 1.

## 1.6 Data Documentation (Metadata)

Metadata or 'data about data' describe the content, quality, condition, and other characteristics of data.

Without supporting metadata, the data would be meaningless. Metadata can easily be more extensive and complex than the data it is associated with.

The metadata needed to understand a observation should be compressed to fit under a single attribute in a table. The experimental design, sampling methods and other supporting documentation are a fundamental and logical component of each data record (Figure 2)**.**

## 1.7 Archival Storage and Access

Data stored in a computerized information system may be of tremendous value. This value is determined not only by what is stored, but *how* it is stored. The type of storage device, the format of the stored data and the types of access available to the data have a significant influence on the ultimate value of the stored data. Ideally (once entered and verified) data are transferred to an archive file or format. The archive file then becomes the reference version of the data, regardless of whether it exists locally or in a formal data archive. All subsequent alterations should be carried out on the archived file or a copy of the archived file, which will then replace the original. This helps avoid the proliferation of
 'offspring' files containing different versions of the data.

Tape or optical disk backups should be made of data that are archived online. Backups should be placed in two or more locations off the premises to protect against data loss. It is important to keep these and on-site copies up-to-date.

| Quality assurance and quality control (QA/QC) | Design | Acquisition | Metadata | Archive |
|---|---|---|---|---|
| Data sheets represent experimental design | + | | | |
| Measurement units are defined on the data sheet | + | | | |
| Attribute names meet project standards | + | | | |
| Date, site, and coded values meet project standards | + | | | |
| Attribute names and descriptions are provided | + | | | |
| Data are complete | | + | | |
| Data entry procedures were followed | | + | | |
| Information such as time, location, and collector(s) was included | | + | + | + |
| Measured data is within the specified range | | + | | |
| Data values or codes are represented correctly | | + | | |
| Data is formatted correctly for further use | | + | + | + |
| Data table attributes are reasonable | | + | + | + |
| Data table design reflects experimental design | | + | + | + |
| Values for each attribute are represented in one way | | + | + | + |
| Errors and corrections are recorded | | + | + | + |
| Metadata are present | | | + | + |
| Check metadata for content (accuracy and completeness) | | | + | + |
| Data glossary is present and accurate | | | + | + |
| Measurement units are consistent | | + | + | + |
| Data and metadata are complete | | | | + |

**Table 1.** *Quality assurance and quality control procedures that are associated with data design, data acquisition, metadata development and data archival in a comprehensive data management system.*
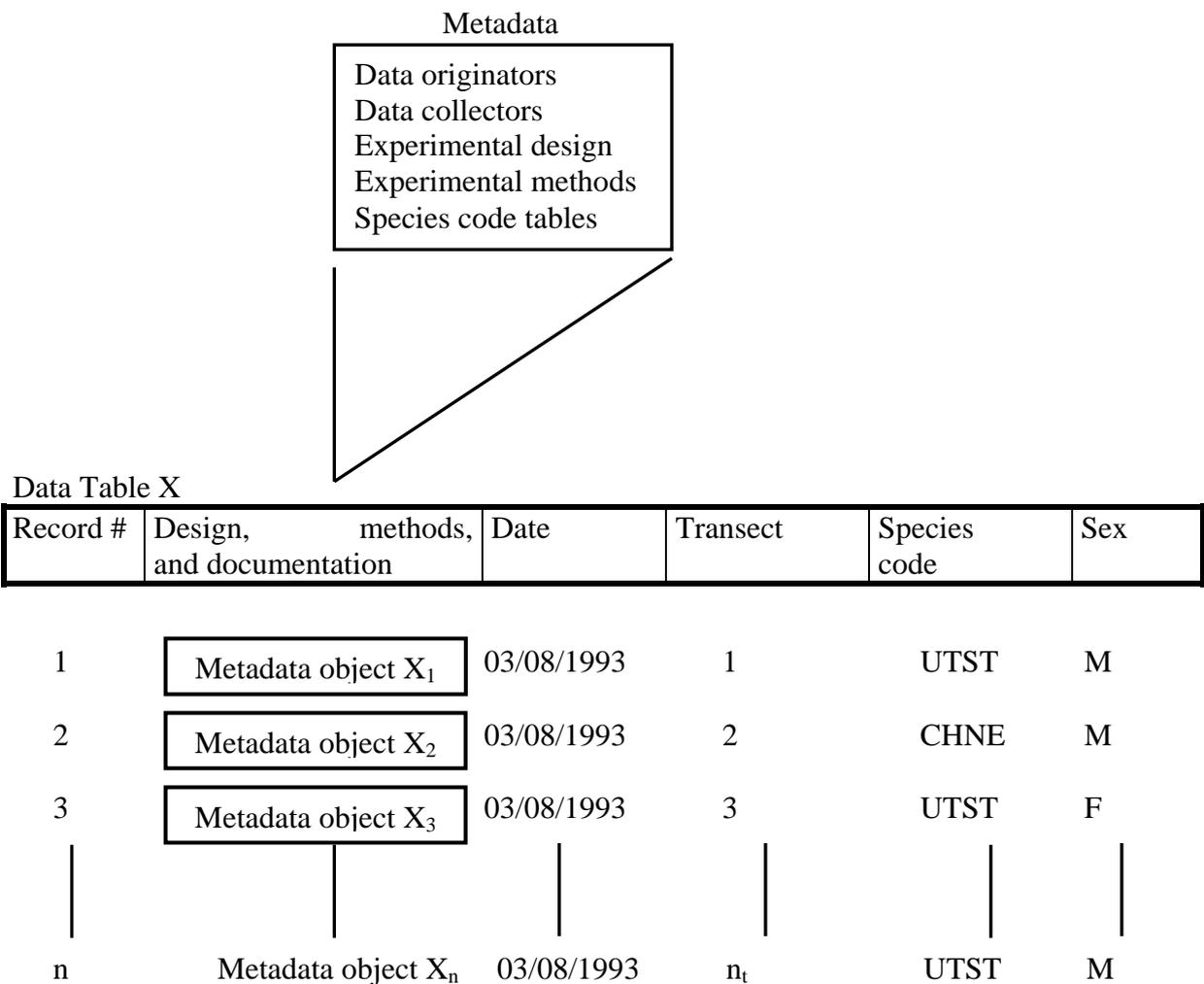
Metadata

```
Data originators
Data collectors
Experimental design
Experimental methods
Species code tables
```

Data Table X

| Record # | Design,          methods, and documentation | Date | Transect | Species code | Sex |
|----------|----------------------------------------------|------|----------|--------------|-----|
| 1 | Metadata object $X_1$ | 03/08/1993 | 1 | UTST | M |
| 2 | Metadata object $X_2$ | 03/08/1993 | 2 | CHNE | M |
| 3 | Metadata object $X_3$ | 03/08/1993 | 3 | UTST | F |
| n | Metadata object $X_n$ | 03/08/1993 | $n_t$ | UTST | M |

**Figure 2.** *Hypothetical relationship of the design, methods, and metadata (represented as a single object) associated with the ecological observation. The metadata object becomes a logical and obligate part of the primary key to each observation.*

## 1.8 Data Administration

Data can originate from a variety of sources and be stored in many locations in a number of formats. This variation makes administration of the data management system an important task. The purpose of data administration is to coordinate the storage and retrieval of computerized information and to oversee the implementation of data management policies and procedures in larger projects. An important task of the data administrator is to establish measures for data protection, for example, by having data backup and recovery procedures.

Data access policy

Guidelines need to be established on the level and type of access allowed by the different users. Access rules can be controversial and may have to be dealt with on a case-by-case basis.

Project data management has high resource and labour demands. In large projects, a data manager, information manager or a data administrator directs this effort. There is an important distinction between a database manager and a data manager. A database manager deals with the technical issues involved in storage and retrieval of data from the database management system. While a data manager, data administrator or information manager deals with the administration and management of all project data, not just the data stored in a database management system.

## 2. Developing a Scientific Database

A scientific database is a computerized, organized collection of related data, which can be accessed for scientific inquiry and long-term stewardship. Scientific databases allow the integration of dissimilar data sets and allow data to be analysed in new ways, often across disciplines, making new types of scientific inquiry possible.

There are several advantages in developing and using scientific databases. First, databases lead to an overall improvement in data quality. The rise in user numbers increases the frequency of detecting and correcting problems in the data. A second advantage is the cost. It generally costs less to save than to recollect data. Also, in many cases uncontrollable factors (such as weather, population influences and ecosystem processes) make data recollection impossible, at any cost.

The development of a database is an evolutionary process. A database will serve a dynamic community of users during its lifetime and will need to change to meet their changing requirements. For this reason four questions need to be asked.

i. Why is database *needed*?

ii. Who will be the *users* of the database

   It is an important question because-
    a) If a user group for the database cannot be identified, then the need for the database should to be re-examined.

    b) Knowledge on the potential users will provide important information on the crucial functions that will be needed to make the database a success.

iii. What types of *questions* should the database be able to answer?

   This will determine how the data will be structured within the database. The data could be structured to maximize the efficiency of the system for the most common types of queries, or multiple indices of the data could be provided to allow different types of searches.

iv. What *incentives* will be available for data providers?

   All database are dependent upon one or more sources of data. The current scientific environment provides few rewards for individuals who contribute data to databases.

### 2.1 Types of Database Systems

Database systems have a central structure that governs the basic working and function of the database. This structure is either part of the database management system software or is defined within the code of more homegrown systems. Today, most systems use a relational database structure but other systems also exist which may be more suited to certain types of data (see Table 2).

### 2.2 Relational Databases

Relational databases are by far the most used database models and are widely used for scientific databases. This is probably because relational databases allow data to be structured in a similar way as hierarchical and network databases but it then allows inter-relationships to be specified based on key values of the data themselves. It is therefore relatively easy to revise the structure of a relational database by changing or adding links between data (Figure 3). These features have made the system popular with users.

| Database Type | Characteristics |
|---|---|
| File-system-based | Uses files and directories to organize information.<br>Examples: Gopher information servers (not typically considered a DBMS) |
| Hierarchical | Stores data in a hierarchical system.<br>Examples: IBM IMS database software, phylogenetic trees, satellite images in Hierarchical Data Format (HDF) |
| Network | Stores data in interconnected units with few constraints on the type and number of connections.<br>Example: Cullinet IDMS/R software, airline reservation database |
| Object-oriented | Stores data in objects each of which contains a defined set of methods for accessing and manipulating the data.<br>Examples: POSTGRES database |
| Relational | Stores data in tables that can be linked by key fields.<br>Examples: Structured Query Language (SQL) databases such as Oracle, Sybase and SQL server, PC databases such as DBASE and FoxBase |

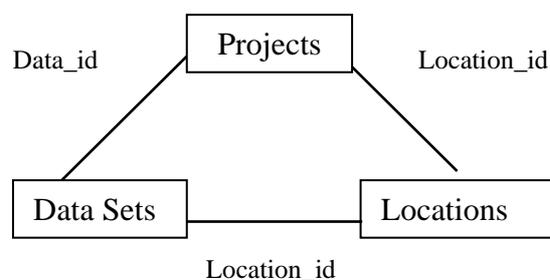**Table 2** *Database system types and characteristics.*

**Figure 3.** Relational database structure

## 2.3 DBMS Software Considerations

The choice of database software will be governed by the tasks that the software is expected to accomplish (e.g. input, query, sorting and analysis). Simplicity is the key.

Many useful functions including the ability to sort, index and search data are built into DBMS software. Large relational databases also include extensive integrity and redundancy checks. These databases can also support transaction processes with 'rollback' capabilities, which allows the recreation of the database as it existed at a particular time.

## 2.4 Interacting with the World Wide Web

An important innovation to DBMS is the introduction of software that enables DBMS to interact with word wide web (www) information servers. This allows a whole range of dynamic www pages to be produced. These pages, called web applications, allow users to retrieve and contribute data and metadata through a common and familiar user interface.

## 2.5 Data Modelling and Normalization

In database creation, the DBMS constitutes the canvas, but the data model is the painting. The purpose of a data model is to explicitly define the entities represented in a database and to spell out the relationships among these different entities (Figure 4). Ultimately, the data model will be used as the road map for the definition of tables, objects and relations.
Normalization is the process where a data model is reduced to its essential elements. The aim of normalization is to eliminate redundancies and potential sources of inconsistency. During the normalization process, it is not unusual to define new entities and attributes or to eliminate old ones from a data model.
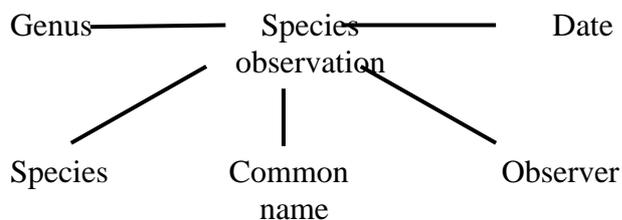


**Figure 4** *Example for an entity-relationship diagram.*

| 'Deep' databases | 'Wide' databases |
|---|---|
| • Specialize on one or a few types of data. | • Contain many different kinds of data. |
| • Contain large amounts of observations on one (or a few) types of data. | • Contain many different kinds of observation, but relatively few data of each type. |
| • Provide sophisticated data query and analysis tools. | • May provide tools for locating data, but typically do not have tools for analysis. |
| • Tools operate primarily on data content. | • Tools operate primarily on metadata content. |

**Table 3** '*Deep' vs. 'wide' databases.*

### 2.6 Examples of Scientific Databases

Based on their content, scientific databases can be placed into one of two categories, deep or wide (see Table 3). 'Deep' databases specialize in a single or a few types of data and implement sophisticated search and analytical capabilities.

'Wide' databases are data collections that attempt to contain all the data related to a specific field of science.

Databases that can be classified as wide are project-based databases. These databases support a particular multi-disciplinary research project and may include a wide range of data specific on a particular site or research question. Databases at Long Term Ecological Research (LTER) sites are a good example. These databases contain a wide range of ecological data (e.g. weather and climate, primary productivity, nutrient movements, organic matter, trophic structure, biodiversity and disturbance data), along with site management information (e.g. research directories, bibliographies and application proposals).

There are a number of essential requirements that a database must have to be successful with users.

i. The data it contains must be wanted by a group of users.

The data must also be up to date and complete. The datasystem must therefore be easy to update and data providers should be given suitable incentives to update the information.

ii. The data must be presented in an attractive format to the user.

iii. The technological expertise needed to use the system must be compatible with the users.

iv. A mechanism of dialog is needed to deal with user inquiries and for identifying unmet user needs.

A database to be successful must meet *all* of these requirements.

## 3. Data Quality Assurance

It cannot be overstated how important data quality is to ecological research data. Data contamination occurs when a process or phenomenon, other than the one of interest, affects a variable value. Prevention through quality control is the first step in eliminating data contamination and is by far more preferable than 'cure'. Prevention is primarily a data management issue, not a statistical one. Many of the quality problems encountered are due to construction and data management.

### 3.1 Preventing Data Contamination

Sources of data contamination due to data entry errors can be eliminated or greatly reduced by using quality control techniques. One very effective strategy is to have the data independently keyed in by two technicians and then computer-verified for agreement. This practice is commonplace in professional data entry services and in some service industries.

**3.2 Illegal Data Filters**

Illegal data are variable values or combinations of values that are literally impossible given the actual phenomenon observed. A simple and widely used technique for detecting this type of contamination is an illegal data filter. This is a computer programme that simply checks a data set with a 'laundry list' of variable value constraints and then creates an output with the identity and details of each violation. The filter programme can be updated and/or enhanced to detect new types of illegal data that may not have been anticipated earlier in the study.

**3.3 Outlier Detection**

An outlier is a unusually extreme value for a variable, given the statistical model in use. What is meant by 'unusually extreme' is a matter of opinion, but the operative word here is 'unusual'. Infact, some extreme values are to be expected in any data set.

Outlier detection is part of the process of checking the assumptions of the statistical models (a process that should be intergral to any formal data analysis).

Elimination of outliers should not be a goal of data quality assurance. Many ecological phenomena naturally produce extreme values, and to eliminate these values simply because they are extreme is equivalent to pretending the phenomenon is 'well-behaved' when it is not.

**3.4 Checking Test Assumptions with Normal Probability Plots**

The normal distribution patterns of uncontaminated data sets of a given size need to be known before outliers can be detected in contaminated data. This is usually achieved by assuming that uncontaminated measurements follow a given probability distribution, usually the normal (or Gaussian) distribution. Most outlier tests also assume that the measurements of interest (the 'errors' in a regression or ANOVA model) follow a normal distribution. An old means of checking this normal distribution, that is gaining increased popularity in the computer age, is the normal probability plot.

A bell-shaped curve is obtained in an idealized histogram of normal distribution data. If at each potential variable value (Y) the cumulative area under the bell curve to its left is calculated, and then these are plotted, we obtain a sigmoid curve. These (Y) values are in fact the values provided in normal probability distribution tables. A normal probability plot essentially re-spaces the vertical axis so that points following this particular sigmoid curve, when plotted against the re-spaced axis, fall on a straight line. If sample points deviate substantially from a line when plotted in this way, they from a non-normal distribution.

**3.5 A Formal Outlier Test: Grubbs' Test**

One of the oldest and most widely used procedures for detecting contamination in a sample is the Grubbs' test. This test assumes that once contamination is removed data will follow a normal distribution. The test is very sensitive to this assumption and should therefore not be used if 'cleaned' data is known not to have a normal distribution.

**3.6 Diagnostic Measures for Leverage Points and Outliers**

Leverage points and outliers are influential data in multiple linear regression. In simple linear regressions, formal diagnostic measures are almost unnecessary, since leverage points and outliers can usually be detected by eye in a plot of the response variable versus the regressor. In multiple linear regression this is no longer true. In this case, the diagnostic checks using leverage values and studentized residuals can help a data analyst find influential observations that are well hidden in scatter plots and other simple analysis tools.

Prevention and detection of contamination in samples and in regression.

The Grubbs' test can be adapted to meet the requirements of repeated small samples, as would often be the case in water quality studies. This can be achieved by using a pooled variance estimator over several samples. Very large data sets are becoming increasingly commonplace, and will require new or the creative adaptation of existing quality assurance methods.

# 4. Metadata and their Importance to Information Technology

Metadata contain information needed to understand and effectively use the data. This includes documentation of the data set contents, context, quality, structure and accessibility.

Metadata are receiving increasing attention from the scientific community. Ecologists, scientific societies and state and federal agencies are recognizing the importance of high quality, well-documented and securely archived data for addressing long-term and broad-scale environmental questions. Ecological and environmental data (e.g. collected at field stations, marine laboratories, national parks, and natural reserves), represent a significant institutional, regional, national and international resource. This data is essential to understanding and monitoring the health of the dynamically changing environment. Comprehensive metadata is the key to 'unlock' these resources, thereby allowing the broad and long-term use of this data.

**4.1 Benefits and Costs Associated with Metadata**

The rows and columns of numeric and textual observations contained within a data set are frequently referred to as raw data. Raw data are usually considered of value if they can be used within the scientific framework of the study that generated the data. Interpreting and using raw data to investigate a study's underlying theoretical or conceptual model(s) requires an understanding of the types of variables measured. The measurement units, the data quality, the conditions under which the variables were measured and other relevant facts are all needed and are provided in metadata. Information is then generated from the combination of raw data and metadata.

Information content can be lost through the degradation of the raw data or the metadata. Such loss is unavoidable and has been referred to as information entropy.

Although metadata loss and degradation can occur throughout the period of data collection and analysis, the rate of loss frequently increases after project results have been published or the study has been terminated.

*Benefits*

At least three major benefits can be obtained from investing adequate time and money into metadata development.

i. *Data entropy is delayed and, correspondingly, data set longevity is increased*

As a consequence of data complexity, time and funding constraints, and information entropy, the life span of a typical ecological data set may be very short, possibly lasting only from data set conception to publication. Even data which is properly archived and maintained is often found to become useless because relevant metadata is missing or unavailable. Development and maintenance of comprehensive metadata can counteract this natural tendency of data to degrade in information content through time.

ii *Data reuse by the originator and data sharing with others is facilitated*

With the rare exception of extremely simple data sets that are immediately analysed after collection, even data collectors need some form of metadata for subsequent analysis and processing. Furthermore, scientists require highly detailed instructions or documentation in order to interpret and analyse accurately unfamiliar research data and complicated experimental designs.

iii *Well-documented data may be used to expand the scale of ecological inquiry and support valid comparisons in space and time*

For example, short-term investigations may evolve or be integrated into long-term studies. Metadata will then be essential for maintaining historical records of such long-term data sets. This is because inconsistencies in documenting data and changes in personnel, methods and instrumentation are likely to occur during ongoing long-term projects. Furthermore, metadata are critical for combining physical, chemical and biological data sets containing different parameters but sharing common spatial or temporal domains. Comprehensive metadata can therefore enable data sets (which were designed for a single purpose) to be used repeatedly for other objectives and over long periods.

*Costs*

High costs, mainly in terms of staff time, can be associated with developing and maintaining metadata. For relatively simple, short-term experiments the size and the effort employed to create the metadata may exceed the size and effort needed to create the data file. This is not unusual in disciplines such as chemistry and physics where understanding the experimental conditions is critical for reproducing data. In such cases, metadata may be scrutinized to the same extent or even more than the data itself. High costs are associated with editing and publication (in paper or electronic formats) of data and metadata. Furthermore, costs associated with developing and distributing metadata are rarely included in project budgets. Long-term stewardship and maintenance of data and metadata represent real cost burdens which are seldom calculated. Its is also difficult to anticipate the numbers of potential secondary users which may require comprehensive metadata for a particular aspect of the data set.

## 4.2 Metadata Content 'Standards' Relevant for Ecology

All ecological data have a spatial or geographic component. The spatial component of the data may range from being central to being relatively unimportant to the success of a project. Geospatial data, for example, are explicitly associated with multiple geographical locations. In such cases, both environmental attributes associated with each sampling point and the specific location of the points are of scientific interest. So far most metadata standardization efforts have focused on data with a strong geospatial component. In contrast, non-geospatial data might include data from laboratory experiments or other ecological data collected at a limited number of locations.

*Geospatial metadata*

Significant effort has been put into developing geospatial metadata standards during the past decade. Recently the Federal Geographic Data Committee (1994, 1998) completed the Content Standards for Digital Geospatial Metadata. The Content Standards contain more than 200 metadata fields that are categorized into seven classes of metadata descriptions. Efforts are underway to add extensions to the Content Standards, creating metadata supersets appropriate to biological, cultural, demographic, and other types of data. For more information on emerging metadata standards in Europe contact MEGRIN (http://www.megrin.org/). MEGRIN is an organization that represents and is owned by a number of European National Mapping Agencies. MEGRIN also maintains a Geographic Data Description Directory on the word wide web that has informational on digital map data of European countries.

*Non-geospatial ecological metadata*

Metadata standards for non-geospatial ecological data do not exist currently in any accepted format beyond those used for individual studies, projects, or organizations. Ecological studies often require large amounts of variable data related to the chemical and physical attributes of the environment, as well as information on the individual organisms, populations, communities and ecosystems which make up the biotic part of the environment. It is unlikely that a single metadata standard, no matter how comprehensive, could be developed to cover all types of ecological data. As a result, a generic set of non-geospatial metadata descriptors have been recently introduced for ecological sciences. This list of metadata descriptors was proposed as a template for more refined project-specific metadata procedures. Five classes of metadata descriptors were defined.

i. Data set descriptors: Basic information of the data set (e.g. data set title, associated scientists, abstract and keywords).

ii. Research origin descriptors: All the relevant metadata that describe the research that generated the data set (i.e. hypotheses, site characteristics, and experimental design and methods).

iii. Data set status and accessibility descriptors: The status of the data set and associated metadata, as well as information related to data set accessibility.

iv. Data structural descriptors: All attributes related to the physical structure of the data file.

v. Supplementary descriptors: All other related information that may facilitate secondary usage, publishing and auditing of the data sets.

These metadata descriptors were formulated to answer five basic questions that might arise when an ecologist attempts to identify and use a specific data set:

i.      What relevant data exists?

ii. Why was the data collected and is it suitable for a particular use?

iii. How can the data be obtained?

iv. How was the data organized and structured?

v. What additional information is available that would facilitate data use and interpretation?

It can be especially difficult to identify and document all supplemental information that may be required for specific data uses. For this reason, it may be beneficial to design metadata that can also serve as a vehicle for user feedback and data anomaly reporting. A 'data set usage history' for example, may add value to data sets and facilitate their long-term use.

*Metadata standards*

Information scientists have developed several generic metadata 'standards' to facilitate cataloguing and discovery of electronic resources. Some examples are the Dublin Core, NASA's Directory Interchange Format and the Government Information Locator Service format.

## 4.3 Software and Resources

Guidelines for metadata structure and supporting technology (i.e. user-friendly software for metadata generation and management) are being discussed and developed by numerous organizations. One particularly promising approach to metadata management is embodied in Web-based metadata search and data retrieval systems. Mercury, developed at the Oak Ridge National Laboratory Distributed Active Archive Centre is an example of such a Web-based system. Mercury supports searches of metadata to identify data of interest and then delivers the data to the user. To make data and metadata available, data providers make them 'visible' in an area on their computer and Mercury periodically harvests the metadata and automatically constructs an index and a relational database that subsequently reside at a central facility. Web-based metadata management programmes like Mercury have several benefits, including control of 'data visibility' by the scientist, high levels of inherent automation and computer platform independence.

*Metadata generation tools*

When selecting a metadata generation tool it is important to consider whether the software meets the specified objectives (especially, metadata completeness), and whether it conforms to industry-wide or discipline-specific guidelines. In some cases, it may be necessary to use more than one metadata generation tool.

For example, an institution's spatial data may be incorporated into a GIS vendor-supplied metadata programme that conforms to Federal Geographic Data Committee (1994, 1998) standards and is well integrated into the GIS environment. Their water quality data, in contrast, may be incorporated into a specific metadata programme that meets other requirements established by a state or federal funding agency.

Some of the most important metadata attributes (e.g. natural history observations) are often recorded and maintained in unstructured formats (often in the form of paper notes). These attributes may be critical for correct data interpretation and analysis. Field notes and other unstructured metadata can either be archived in paper files or converted into digital format (e.g. scanning, transcription to text or word processing files). These types of unstructured metadata may be suitable for exchange with expert colleagues but are inadequate for electronic data set publication and sharing with the broader scientific community. Although existing or proposed metadata generation tools may fill most of a project's metadata needs, consideration should be given to how maps, field notes and other unstructured data will be archived and managed, as well as referenced in the metadata.

*Metadata structure*

Increasing amounts of supplementary metadata are being added to meet the demands of secondary data users. The utility of this data can be improved by adding a structure to the metadata. A highly structured and fully searchable metadata record would include a sophisticated database management system (DBMS). Minimal metadata structure should not be confused with low content.

An increased metadata structure can be beneficial for several reasons. The structured metadata character checklist often provides a memory-aid on the information needed to facilitate subsequent data processing and interpretation. The increased structure will facilitate the development of searchable catalogues and database interfaces. This will potentially allow the data to be available to a larger number of users and processing software. High levels of structure may be a good practice or, in some cases, may be required for specific projects (e.g. those requiring periodic data audits). However, highly structured metadata may be excessive where low levels of secondary usage are anticipated. Thus, the benefits of incorporating metadata into highly structured DBMS format should be considered in relation to software, programming, development and maintenance costs.

The choice of metadata media and structure will often be dictated by the availability of metadata generation tools, available trained personnel, time and funding constraints, and projected rates of metadata usage. When specific metadata tools are inadequate or unavailable, metadata may be incorporated into word processing files (or free-flowing ASCII text), analytical programmes (e.g. Statistical Analysis System Programmes), or more structured DBMS Programmes. Satisfying high levels of demand for metadata may necessitate making the metadata DBMS-accessible via the word wide web.

During the mid-1990s, a number of organizations with moderate to large holdings of data began implementing metadata schemes. Format descriptions or more sophisticated DBMS software that increased metadata structure and often directly linked data and metadata were introduced. The primary objective of these efforts was to initiate the standardization of the data content and structure so as to facilitate search and retrieval.

Several tools are available for implementing word wide web-based metadata applications, including Hypertext Markup Language (HTML) and extensible Markup Language (XML). XML is a standardized text format that represents a subset of the Standard Generalized Markup Language (SGML; ISO standard 8879). XML currently is the most useful representation language for documenting the content and semantics of Web-based resources. It was specifically designed for transmitting structured data to web applications, and its utility is further increased by its relative ease of expansion by having a flexible structure that supports arbitrary nesting, and the potential for automated validation.

## 4.4 Metadata Implementation

Objectives for metadata implementation include facilitating identification and acquisition of data for a specific theme, time period and/or geographical location. It is also needed to help determine the data's suitability for specific objectives, analysis, modeling and processing. Three major issues warrant consideration during metadata planning and implementation: desired data longevity, projected rate of use and sharing of responsibility.

All data should be accompanied by some form of metadata (even if minimal). The level of metadata provided will determine the extent and time that the data can be reused by the original investigator(s), scientists, resource managers, decision-makers and other potential users. Metadata development and maintenance can be a costly enterprise. It may therefore be worthwhile attempting to match the metadata content and structure to the needs of the anticipated users. Dedicating project resources to metadata design and implementation costs money and personnel effort and can result in fewer publications in the short term. The rewards are the production of high quality data and metadata that can be 'mined' for many years or even decades. The balance of short-term costs against potential long-term benefits is an issue warranting considerable thought and discussion by data collectors, data users, institutions, and funding agencies.

The key step of metadata implementation is to assess the site or project needs. The objectives for the data need to be identified (e.g. the desired longevity of the project or data or the potential reuse needs of the data). Guidelines and procedures for data sharing and data ownership need to be established. The available infrastructure (e.g. hardware, personnel, funds, etc.) will need to be assessed, and finally metadata activities need to be prioritized and categorized.

After data categories have been prioritised, it is necessary to adopt an existing metadata standard (e.g. geospatial metadata standard, FGDC 1994,1998) or identify a set of minimal or optimal metadata descriptors that meet perceived needs. It is also recommended that a pilot project using one or more relatively 'simple' data sets be carried out. Project successes and difficulties should be used to re-evaluate site needs and objectives. Formal metadata standards and supporting software can then be developed. While these standards are being established metadata descriptors can be used to develop metadata for the individual scientists, laboratories and projects. Small groups of scientists focused on a specific research objective, such as synthesizing data on a particular topic, may benefit significantly from efforts to implement metadata.

- Keep it simple! Start small and build upon successes. For example, the time and effort spent on a pilot project are usually paid back several-fold in the long run.

- Build consensus among scientists and data managers from the start. Data management initiatives, regardless of their potential benefits, are often unsuccessful when the 'user community' is excluded from the process.

- Data longevity is roughly proportional to metadata comprehensiveness. However, establishing a goal of complete metadata that can meet all future needs may be exorbitantly expensive and, ultimately, unattainable.

- Data and metadata should ideally be platform independent. Hardware and software change frequently. Today's 'standard' may be gone tomorrow. Thus, it pays to avoid administrative storage formats whenever possible.

- The stewardship of secure archived and accessible ecological data for future research will depend on the way its promoted and rewarded.

Both basic and applied ecological research depends upon the availability of data including the ability to locate and use the data. If a greater effort is made to develop high quality data sets and accompanying metadata, then individual scientists and organizations can focus their valuable time on the analyses. Comprehensive metadata will also allow individual scientists and organizations to reuse data intended for other applications. Flexible metadata generation and management tools that support entry, search and retrieval are essential for facilitating metadata implementation. There is a significant need for research and development in this area.

## 5. Archiving Ecological Data and Information

A data archive (usually in electronic form) is a collection of data sets with accompanying metadata, stored in such a way that a variety of users can locate, acquire, understand and use the data. Archived data is secure against natural and man-made disasters, and are conserved in a form that will continue to be accessible as technology changes.

As ecology research moves toward regional and global multi-disciplinary investigations, mechanisms are needed to promote the sharing of data among many disciplines (meteorology, hydrology, soil science, forestry, agriculture, botany, etc.) and to insure future data accessibility. Official archives for ecological data will someday make this process possible. Submitting data to archives and acquiring data from archives will become an integral part of tomorrow's scientific process. However, archiving of data has not yet been given the attention, resources, or recognition required to become a routine part of the research and publication cycle. Although there is a limited number of official data archives for ecological data, ecologists can manage their data today in ways that meet local needs for data security and access. This will prepare their data for eventual inclusion into data archives.

A data archive is established for the 'preservation with understanding' of ecological data. Optimally, it is a permanent collection of data sets with accompanying metadata such that a

variety of users can readily acquire, understand, and use the data. The design of an archive is defined by three constraints:

i. Scope of the data to be stored;

ii. Capabilities for searching and access;

iii. Resources for operation and maintenance.

# 6. Demonstrations and Hands-on Computer Experience

Morning lectures were followed by topic related laboratory exercises. The aim of which was to familiarize participants with the available and relevant software, and demonstrate the different procedures for software selection. The exercises were arranged around creating and managing databases and making them available via the Internet.

### 6.1 The Tool for the Data Entry and Basic Analyses: Microsoft Excel

Most of the data collected in the field will be entered into the computer in a spreadsheet file. Excel is one, if not the most, common spreadsheet programme. Primary processing and simple statistical analyses can be carried out on the entered data. Most workshop participants were found to have already used this software. So only a short introduction was given after which outlier (unusually extreme value) detection was carried out on the software. Participants first created distribution graphs of the data. They then used the Grubbs' test to determine if suspected data were true outliers. Since the Grubbs' test assumes a normal distribution, the participants first checked this assumption using normal probability plots, and if needed, they transformed the data before running the Grubbs' test.

### 6.2 A Useful Database Managing Software: Microsoft Access

A large set of data usually results from a research or a monitoring project. After analysing them to get an answer to the special question at hand and after the publication of the results, the basic data still represent a high value. They can be used for example, for comparative studies, long-term studies, they can serve as reference data for other researches, or as input parameter to ecological models. To make the data accessible to other scientists or the wider public it is inevitable to arrange them into an easy-to-handle, logically constructed and well managed database.

There are several database management programmes in the market, but one of the most widely used tools is Microsoft Access, so the participants of the data management course became acquainted with this tool in a seminar. Although this software was developed mainly for handling business datasets, it can be used for handling ecological datasets as well.

Over several hours participants gained some experience with this software. They learnt how to create and modify tables for storing data, forms for entering data into the computer and queries, with the help of which the sought information can be gained from the database. After a short basic training the organizers of the course presented the data sheets of two projects, a bird and a vegetation monitoring project, and the participants had to choose one and create a database for storing the collected data, and make queries to get some information from the database. This practice enabled the participants to face the difficulties, to find the strengths

and weaknesses of the programme and seek solutions. The exercise showed that even in the case of the same set of data every participant created a different database, found different problems and different solutions even for the same problem. This demonstrated that before arranging the data into a database the aim of the database, the possible users, and the questions which the database has to be able to answer must first be considered.

### 6.3 The Role of the Internet in the Data Networks

The fundamental aim of the workshop was to establish the base for a Central Europe LTER research network information system. The main purpose of such a system would be to increase the ease of information exchange. The wide use and availability of Internet in nearly all research institutes favours it use in the implementation of such a network. To construct research results into databases which can then be accessed via the web needs at least some basic knowledge of Internet tools. Web oriented knowledge was therefore given great importance in the workshop.

### 6.4 The Language of the World Wide Web: HTML

The www (world wide web) uses the HTML (HyperText Markup Language) to create the web pages. This is a descriptive computer language for the web browser programmes (MS Internet Explorer, Netscape, Opera, etc.) which determines how the pages should appear on the display to the visitors. It is quite simple to use at a basic level, the knowledge of a few commands (here called "tags") is enough to create a useful web page. We can embed not only objects (pictures, sounds), but also links to other related pages or programmes, which are able, for example, to search in our database. HTML is the base for the design of the network interface of our databases. It is advisable to consider that although the HTML offers many tempting opportunities to make the pages more attractive, but these tools will increase the size and can make the availability of the page difficult. It is therefore worth striving to create simple and well-arranged pages.

To edit HTML pages there are several types of easy-to-use software on the market. The participants of the workshop became acquainted with the Netscape composer. The participants created their own simple web page in the HTML seminar, which were collected by the trainers, to make them available soon via the Internet. At the end of the course a decision was made to create an international web page, where the information heard in the workshop and several connected information would be collected for those who would be interested in the programme.

### 6.5 Web Servers

To make web pages public via the Internet a computer with a web server programme is needed. Several types of software are available and most of the new operating systems often have them built in. In the workshop the participants used the Microsoft Internet Information Server and the MS Personal Web Server. This has a good connection with the other Microsoft software. Of course, servers working under other operating systems (e.g. Unix/Linux Apache Web Server) are perfectly suitable. To install these usually the help of the system operator of the research institute is needed, but after installation everybody can work with them.

**6.6 Web Service Tools**

Since the HTML determines only the outlook of the web pages, to communicate with the databases other associated programmes are essential. These programmes assure the link between the database tools and the users, execute user requests, and create a mainly HTML based page of the results and returns it back to the user. For these aims CGI and JAVA programmes are used most often. The JAVA is a computer language specially written for HTML, which usually runs on the computer of the user. CGI scripts do the job on the server, and several programming languages are able to write them. The use of both programmes requires some experience, but a huge amount of scripts can be downloaded from the Internet for a number of purposes freely or quite cheaply. These must be only slightly modified for our special use.

# 7. Information on GTOS and the TEMS Database

Kristin Vanderbilt using GTOS documentation informed the participants of the workshop on the organization, goals, and links GTOS has with other observing networks. The emphasis on the improvement of data exchange and accesss to information by wide circle of users was highlighted. As well as the need to harmonize methods, data classification and data dissemination.

The major issues addressed by GTOS were described as follows:

• Changes in land quality;

• Availability of freshwater resources;

• Loss of biodiversity;

• Climate change;

• Impacts of pollution and toxicity.

Due to the ongoing revision process the TEMS metadatabase was not available on the Internet during the period of the workshop. However a copy of the TEMS 97 database was available to illustrate how the database can be used to find information about participating sites. Using reports of workshop attendees and literature sources a preliminary compilation of metadata on the types and length of data sets of selected sites in the CEE Region was produced.

# Information On-Site Data

The type and time duration (where available) of data sets from selected sites in CEE countries

**Czech Republic**

Krkonose National Park and LTER site

Complex mountain system, subarctic mires, Ramsar site

- thorough inventories, monitoring and research since 1960s
- permanent plots: vascular plants since 1967
- nonvascular plants since 1981
- invertebrates since 1982
- vertebrates since 1983
- hydrology data since 1904
- climate, soil, geology since 1932
- GIS (ARC/INFO- more than 300 layers)
- a 4-member data management team exists.

Contact person: Jiri Flousek, Krkonose NP Administration, posta@kmap.cas.cz

Trebon Basin Biosphere Reserve and LTER site:

mosaic of wetlands, and temperate broad leafed and mixed forests, 2 Ramsar sites since 1990 and 1993
- temperature since 1876
- hydrology since 1945
- bird census since 1960
- water chemistry
- fishponds monitoring and management over 20 years, waterfowl
- population fluctuations, mammals, insects, forestry plots since1938.

Contact person: Miroslav Hatle, Trebon Basin PLA Administration, chkot@envi.cz

**Hungary**

Sikf   kút Oak Forest LTER site:

temperate deciduous forest

- continuous data on tree, shrub and herb layer
- dynamics and meteorological elements since 1972
- non- continuous data for primary production, secondary production
- zoomass, litter production and decomposition, soil bacteria and fungi
- nutrient cycling in the air- soil- plant systems
- oak decline since 1979

Contact person; Janos Attila Toth, Debrecen University, tja@tigris.klte.hu

Lake Balaton LTER site:

temperate shallow lake

- meteorological and hydrological data since 1900
- water chemistry at 12 sample points since 1975
- phytoplankton and zooplankton since 1965
- fish yields since 1905
- benthic and littoral invertebrates since 1975
- macrophytes since 1975

Contact person: Sandor herodek Balaton Limnological Inst.HAS, intezet@tres.blki.hu

KISKUN sand forest- steppe LTER site:

mosaic of semiarid grasslands, wetlands, Juniper- Poplar woodland in agricultural landscape

- meteorological and ground water records since 1950
- historical and recent vegetation maps from 1783-84, 1883-84,1980ies, and 1997-2000 years
- flora database for the vascular flora since 1995
- grassland phytomass since 1969 (not continuous)
- insect records from light traps since 1962,
- insect dynamics in apple orchards 1986-1994
- plant- herbivore interactions since 1991

Contact person: Edit Kovacs-Lang, Inst.Ecol.Bot.HAS, lange@botanika.hu

**Poland**

Bieszczady Mountains LTER site:

Carpathian beech- fir forest, alpine meadows, reservoirs

- data on climate
- hydrology, forest structure since 1960
- history of land use since 1850
- rodent populations since the mid-1950s
- large mammals since 1980
- invertebrates since 1960
- hydrochemistry since 1973
- benthic fauna since 1989
- ozone studies since 1997 (Regional Network on the Carpathians)

Contact person: Kajetan Perzanowski Internat.Centre of Ecol.PAS icepas@mikrotech.com.pl

**Romania**

Danube Delta Biosphere Reserve:

complex of natural and seminatural ecosystems like river branches, channels, lakes, swamps, wetlands, dunes, forests

- early data on hydrology, sediment formation, structure and dynamics of
- plant communities fish and bird populations
- since the 1980-ies. records on water chemistry, heavy metals
- phytoplankton, macrophytes, primary and secondary production

Contact person: Angheluta Vadineanu, Bucharest University, anvadi@bio.bio.unibuc.ro

Retezat Mountains Biosphere Reserve:

complex mountain system (the site belongs to the GEF network)

- earlier surveys produced data on geology, geography and the flora of the region
- since the 1980-ies floristic, faunistic, phytocoenological data, water
- chemistry, plankton data from monutain lakes

The site is a member of the "Carpathian Regional Ozone and Air pollution network"

Contact person: Dan Cogalniceanu, Bucharest University, danco@bio.bio.unibuc.ro

**Slovakia**

Bab Forest LTER site:

temperate deciduous oak- hornbeam forest

- data on biodiversity primary and secondary productivity
- meteorological elements
- nutrient cycles
- soil characteristics

Contact person: Pavol Elias, Agricultural University of Nitra, elias@afnet.uniag.sk

Biely Vah LTER site:

temperate coniferous forest

- data on biodiversity
- meteorological elements
- soil characteristics
- structure of vegetation

Contact person: Julius Oszlányi, Institute of Landscape Ecology SAS, director@uke.savba.sk

<u>Polana LTER site:</u>

temperate mixed mountaineous forest

- data on biodiversity
- structure of vegetation
- primary productivity
- meteorological elements
- soil characteristics

Contact person: Julius Oszlányi, Institute of Landscape Ecology, SAS, [director@uke.savba.sk](mailto:director@uke.savba.sk)

<u>A Regional Monitoring Network for</u>

- ozone, sulphur dioxide and nitrogen dioxide concentrations and their effect on forest was established in the Carpathian Mountains. Czech Republic, oland, Romania, Slovakia, and Ukraina are involved.

Project leaders: A. Bytnerowicz (USDA) and K. Grodzinska (Institute of Botany, Polish Acad. Sci.)

Contact person:Krystyna Grodzinska, [grodzin@ib-pan.krakow.pl](mailto:grodzin@ib-pan.krakow.pl)

The status of the data at the different sites is very different. Many of the data only exist in the form of publications or are only stored on the computers of the individual scientists. In a few cases the data are organized into accessible databases.

Information management levels are also very different between sites. The Krkonose (Czech Republic) site has had a four-membered management team for the last four years. However most sites lack human and financial resources preventing the development of information management. A sound initiative of providing special grants to be used in developing database from existing "latent" long-term data was launched by the Czech Republic in 1999 and was followed by Hungary in 2000.

Two main problems have to be overcomed if advances are to be made:

i. Attracting informatic experts away from industry and economy, where they earn much higher salaries, to research institutions.

ii. Scientists are expected to produce scientific publications and not databases. There is, therefore a need to change the approach and attitudes of scientific assessment. Financial backing is also needed to overcome these problems. Similar based workshops will help change scientific approaches and attitudes.

**Workshop Homepage**

At the end of the workshop, participants agreed to develop a web homepage of the workshop as a constructive utilization of what they had learnt and to serve as a basis for further cooperation among the sites in the region.

The coordinator for the web page is Zdenek Fajfr, Czech Republic. Workshop participants will supply information on their sites and Barbara Lhotsky and Gabor Varbiro from Hungary will provide the additional material.

The planned content of the web page:

i. Workshop summary;

ii. List of instructors;

iii. List of participants and sites they represented;

iv. Themes of the lectures;

v. The documented material of the lectures;

vi. Links to the homepages of the sites and institutions of participants;

vii. Links to access different kinds of tutorials e.g. databases, metadata standards, recommended softwares, etc.

It was also agreed that an electronic forum should be established.

Vácrátót, 14. 12. 2000
Edit Kovács-Láng (organizer of the workshop)

**GTOS Secretariat**
c/o FAO, SDRN
Viale delle Terme di Caracalla
I-00100 Rome, Italy

Tel: +39-06 5705 3450
Fax: +39-06 5705 3369
E-mail: gtos@fao.org
Web: www.fao.org/gtos