

Supplement E. Data Management Plan for the Pima County Ecological Monitoring Program

**Supplement to the Pima County Ecological Monitoring Program:
Phase II Monitoring Plan Summary**

September 2010

**Brian Powell, Pima County Office of Conservation
Science and Environmental Policy**

This page intentionally left blank

1 Introduction

“Data and information are the basic products of scientific research. In ecological research, where field experiments and data collections can rarely be replicated under identical conditions, data represent a valuable and, often, irreplaceable resource . . . In long-term ecological studies, retention and documentation of high quality data are the foundation upon which the success of the overall project rests”

Brunt 2000

The main body of the report provides justification for a long-term monitoring program to support the County’s forthcoming MSCP and ongoing SDCP effort. This supplement presents a strategy for ensuring that the Pima County Ecological Monitoring Program (PCEMP) data are documented, secure, accessible, and useful into the future. This data management plan also refers to the need to develop other standards and steps for achieving data management goals. Because the PCEMP is now ready to be implemented (after Section 10 permit issuance), this plan provides an opportunity to articulate the level of detail and tools for data management that will be employed once the PCEMP is implemented. Therefore, this plan acts as a foundation upon which to build as new protocols are developed and advances in technology are adopted.

This data management plan describes how the PCEMP will:

- Support PCEMP and SDCP objectives
- Acquire and process data
- Assure data quality
- Document, analyze, and summarize data and information
- Integrate with other regional data management systems
- Disseminate data and information
- Maintain, store, and archive data

Revisions to this plan and associated data management documents (guidelines and procedures) will be made as needed, and the overall plan will be reviewed and revised as necessary every 3-5 years.

1.1 Data and Data Management: An Overview

The collection of scientifically credible, natural resource data is a critical step toward understanding and conserving natural resources. Though data are a set of discrete, objective facts, if they are to be meaningful or useful, they must be processed or transformed into information by adding context and appropriate interpretation. Thus, data management is more than simply inputting values into a table or spreadsheet or filing away data sheets. Rather, if the data management goals of the PCEMP are to be achieved, a modern information management infrastructure (i.e., staffing, hardware, software) must be developed. In addition, procedures must be established to ensure that relevant natural resource data collected by PCEMP staff, cooperators, researchers, and others are entered, quality checked, analyzed, reported, archived, documented, cataloged, and made available to others for management decision making, research and education. This endeavor requires planning.

Any good set of data must be accompanied by enough explanatory documentation (*e.g.*, how and why it was collected) so that any trained personnel (now or in the future) can understand and use it with confidence. Therefore, any data management system cannot simply attend to the tables, fields, and values that make up a data set. It must also provide a process for developing, preserving, and integrating the context that makes data interpretable and valuable. Although thoroughly documenting a data set is time-intensive, it results in clear preservation and presentation of the data.

The term 'data' has many meanings, and they fall into five general categories:

- **Raw data:** GPS rover files, raw field forms and notebooks, photographs and sound/video recordings, telemetry or remote-sensed data files, biological voucher specimens;
- **Compiled/derived data:** Relational databases, tabular data files, GIS layers, maps, species checklists;
- **Documentation:** Data collection protocols, data processing/analysis protocols, record of protocol changes, data dictionary, metadata, data design documentation, quality assurance report, catalog of specimens and photographs;
- **Reports:** Annual progress reports, final reports (technical or general audience), periodic trend analysis reports, publications
- **Administrative records:** Contracts and agreements, study plans, research permits/applications, other critical administrative correspondence

1.1.1 Sources of Natural Resource Data

There are many potential sources of important data and information about the condition of natural resources, including inventories, monitoring, and research projects. Because the PCEMP focuses on long-term monitoring, our first priority is to produce and curate high-quality, well-documented data for that effort. However, applying a set of data management standards, procedures, and infrastructure to other natural resource data should be a long-term goal of Pima County. As time and resources permit, we will work toward raising the level of data management for current projects, legacy data, and data originating from outside the program. These categories encompass one or more of the following data formats:

- Hard-copy documents (e.g., reports, field notes, survey forms, maps, references, administrative documents)
- Objects (e.g., specimens, samples, photographs, slides) Electronic files (e.g., Word files, email, websites, digital images)
- Electronic tabular data (e.g., databases, spreadsheets, tables, delimited files)
- Spatial data (e.g., shapefiles, coverages, geodatabases, remote-sensing data). (Pima County already has an excellent system for spatial data produced by the IT team).

1.2 Data Management Goals and Objectives

The data management activities for the PCEMP will provide scientifically and statistically sound data to support this goal. The data management objectives will be guided by five principles:

- *Quality*: ensure that appropriate quality assurance measures are taken during all phases of data development: acquisition, processing, summary and analysis, reporting, documenting, and archiving.
- *Interpretability*: ensure that complete documentation accompanies each data set so that users will be aware of its context, applicability, and limitations.
- *Security*: ensure that both digital and analog data are maintained and archived in a secure environment that provides appropriate levels of access to project leaders, technicians, network staff, and other users.
- *Longevity*: ensure that data sets are maintained in an accessible and interpretable format, accompanied by sufficient documentation.
- *Availability*: ensure that the data and information from our activities are made available and easily accessible to managers and other users.

The PCEMP Data Management Plan outlines how Pima County will implement and maintain a system that will adhere to these principles. This plan reflects the program's commitment to the acquisition, maintenance, documentation, accessibility, and long-term availability of high-quality data and information.

2 Data Management Resources: Infrastructure and Systems Architecture

Pima County has expended considerable resources toward computer resource architecture (applications, database systems, repositories, and software tools) and this infrastructure is an important foundation for the program. The PCEMP will rely on existing IT personnel and resources to maintain its computer resource infrastructure. These resources include, but are not limited to, hardware replacement, software installation and support, security updates,

virus-protection, telecommunications networking, and backups of servers. This chapter will describe the infrastructure that is likely to be needed and which is often central to data management activities.

2.1 Computer Resources Infrastructure

An important element of a data management program is a reliable, secure network of computers and servers. The current system of county servers is administered by the Information Technology group in downtown Tucson, and the local IT specialist in the NRPR building. All project data will be stored on the secure servers administered by the Information Technology group. Because of the recent service transfer from NRPR to the IT group, it may be necessary to invest in an additional (redundant) data backup system.

Among the types of data that will need backup include:

- Master project databases: compiled data sets for monitoring projects and other multi-year efforts.
- Common lookup tables: *e.g.*, projects, personnel, species
- Program digital library: repository for finished versions of products from program projects (*e.g.*, reports, methods documentation, data files, metadata, etc.)
- GIS files: base spatial data, imagery, project-specific themes
- Working files: working databases, draft geospatial themes, drafts of reports, administrative records

2.1.1 Maintaining Digital Files

The PCEMP local area network (LAN) will be set up to accommodate a hierarchical directory structures for storing digital files. This is critical to establish and maintain early in the program's development because the large number of files that can quickly accumulate.

Below are six categories of directory structure sections in which digital files will be maintained:

- Admin: documents related to program administration.
- Databases: all lookup tables, and back-end databases.

- **Libraries:** read-only storage of cataloged photographs and other reference documents.
- **Working:** workspace where groups and individuals can maintain draft material and other files as arranged by projects. The layout of folders and subfolders is more flexible here than elsewhere, but these areas must be cleaned out once per year to maintain order.
- **GIS:** base spatial data, imagery, and project-specific themes. Later in the program we will develop a system for separating those GIS files that are created by the Geographic Information Services group (Pima County Public Works), those created by the GIS specialist in NRPR, and those created by PCEMP staff.
- **Project Archive:** read-only storage of finished project products.

In general, this file management strategy has a number of advantages. First, working files are kept separate from finished products, the latter are typically read-only. Further, standards such as naming conventions and hierarchical filing will be enforced within the Libraries, Project Archive, Database, and GIS sections.

2.1.2 National Park Service Inventory and Monitoring Program Applications

The National Park Service (NPS) has devoted considerable resources to data management and as a result have developed many database applications that will be valuable to our effort. The PCEMP has a close working relationship with the Sonoran Desert Network I&M program, based in Tucson. They have committed to provide the PCEMP with databases and related products. Among those that might be useful for our efforts include databases for landbirds and an integrated aquatics database for plants, geomorphology, and water quantity and quality.

For those protocols that will differ from those collected by the Sonoran Desert Network, Pima County will have full access to the Vital Signs Monitoring Protocol Database (*i.e.*, **Protocol Database**), a web-based clearinghouse of sampling protocols used in other National Parks to monitor the condition of selected natural resources. The database provides allows the user to download a digital copy of sampling protocols that have been developed by the prototype monitoring parks or other well-established protocols used in National Parks. The

Protocol Database also makes it possible to download database components (*e.g.*, tables, queries, data entry forms) consistent with the Natural Resource Database Template (see below) that have been developed for a particular protocol in MS Access.

The Natural Resource Database Template (NRDT) is a flexible, relational database in MS Access developed by the NPS I&M program for storing inventory and monitoring data (including raw data collected during field studies). This relational database can be used as a standalone database or in conjunction with GIS software (*e.g.*, ArcView or ArcGIS) to enter, store, retrieve, and otherwise manage natural resource information. The template has a core database structure that can be modified and extended based on the needs of the PCEMP. The NRDT includes separate modules detailing different aspects of monitoring project implementation, from sampling design to data analysis and reporting, and include data management components that describe database table structure, data entry forms, and quality checking routines. Approved monitoring protocols, including the databases that are based on the Database template, are made available through a web-based protocol clearinghouse (see above).

2.1.3 Other National-Level Information Management and GIS Applications

STORET is an interagency water-quality database developed and supported by the Environmental Protection Agency (EPA) to house local, state, and federal water-quality data collected in support of managing the nation's water resources under the Clean Water Act. STORET is currently used by Pima Association of Governments for data collected at Cienega Creek Preserve.

2.1.4 Other Related County Databases

Some departments and programs within Pima County have Pima County is also developing an open-space database that would store information critical to open space management. This database would relate the "Preserves" GIS layer to tabular data about property rights, management plans, buildings, wells, and other infrastructure located on County open space parcels. As in the current Mapguide system, users would view preserve lands in relation to

various reference layers, but custom applications would be developed so that departments could update information.

2.2 PCEMP Systems Architecture

Rather than developing a single, integrated database system for all parameters in the monitoring program, it will be best to develop modular, stand-alone project databases that share design standards and links to centralized data tables. This way, individual project databases are developed, maintained, and archived separately. There are numerous advantages to this strategy. First, because data sets are modular in format, they allow greater flexibility in accommodating the needs of each project area. Individual project databases and protocols can be developed at different rates without a significant cost to data integration. In addition, one project database can be modified without affecting the functionality of other project databases. Also, by working up from modular data sets, a large initial investment in a centralized database can be avoided.

2.2.1 Project Database Standards

Project database standards are necessary for ensuring compatibility among data sets, which is vital given the often unpredictable ways in which data sets will be aggregated and summarized. When well thought out, standards also help to encourage sound database design and facilitate interpretability of data sets. As much as possible, network standards for fields, tables, and other database objects will mirror those conveyed through the Natural Resource Database Template. In addition, documentation and database tools (*e.g.*, queries that rename or reformat data) will be developed to ensure that data exports for integration are in a format compatible with current national standards. Databases for the PCEMP should all contain the following main components:

- **Common lookup tables**: links to entire tables that reside in a centralized database, rather than storing redundant information in each database. These tables typically contain information that is not project-specific (*e.g.*, lists of reserves, personnel, and species).
- **Core tables and fields based on templates**: these tables and fields are used to manage the information describing the ‘who, where and when’ of project data. Core tables are distinguished from common lookup tables in that they reside in each individual project

database and are populated locally. These core tables contain critical data fields that are standardized with regard to data types, field names, and domain ranges.

- Project-specific fields and tables: the remainder of database objects can be considered project-specific, although there will typically be a large amount of overlap among projects. This is true even among projects that may not seem logically related. For example, a temperature field will require similar data types and domain values.

Certain key information is not only common to multiple data sets, but to the organization as a whole – lists of contacts, projects, reserves, and species that are often complex and dynamic. It is a good strategy to centralize this information so that users have access to the most updated versions in a single, known place. Centralizing also avoids redundancy and versioning issues among multiple copies. Centralized information is maintained in database tables that can be linked or referred to from several distinct project databases. Program applications for project tracking, administrative reporting, or budget management can also link to the same tables so that all users in the Network have instantaneous access to edits made by other users.

The three types of database objects also correspond to three acknowledged levels of data standards. Because common lookup tables are stored in one place and are referred to by multiple databases, they represent the highest level of data standard because they are implemented identically among data sets. The second level of standards is implied by the core template fields and tables, which are standardized where possible, but project-specific objectives and needs could lead to varied implementations among projects.

3 Data Management Process and Work Flow

This section considers the general work flow characteristics of PCEMP projects that produce natural resource data, and then gives an overview of how natural resource data are generated, processed, finalized, and made available. Data management activities that relate to the various stages of a project are highlighted. By describing the progressive stages of a

project and the life cycle of the resulting data, one can more easily communicate the overall objectives and specific steps of the data management process. In addition, this awareness helps manage staffing resources needed to produce, maintain, and deliver quality data and information. More details about data acquisition, quality assurance, documentation, dissemination, and maintenance can be found in later sections of this plan. In the next phase of this plan, tasks will be assigned to specific individuals, but because staff structure has not been finalized, it is too early in the program to assign such tasks.

3.1 Project Work Flow

Projects can be divided into five primary stages:

1) Planning and Approval. During the initial phase of the program, many of the preliminary decisions regarding project scope and objectives are made, and funding sources, permits, and compliance are all addressed.

2) Design and Testing. All of the details are worked out regarding how data will be acquired, processed, analyzed, reported, and made available to others. An important part of this phase is the development of the data design and data dictionary, where the specifics of database implementation and parameters that will be collected are defined in detail. Devoting adequate attention to this aspect of a project is possibly the single most important part of assuring the quality, integrity, and usability of the resulting data. Once the project methods, data design, and data dictionary have been developed and documented, a database can be constructed to meet project requirements.

3) Implementation. In this phase, data are acquired, processed, error-checked, and documented. This is also when products such as reports, maps, GIS themes, and others are developed and delivered. Throughout this phase, data management personnel function primarily as facilitators by providing training and support for database applications, GIS, GPS, and other data processing applications; facilitation of data summarization, validation, and analysis; and assistance with the technical aspects of documentation and product

development. Toward the end of this phase, project staff members work to develop and finalize the products that were identified in the project planning documents (*i.e.*, protocol, study plan, contract, agreement, or permit).

4) Product Integration. During this phase, data and other products are integrated into national databases (if appropriate), metadata records are finalized, and products are distributed or otherwise made available to their intended audience. Certain projects may also have additional integration needs, such as when working jointly with other agencies for a common database.

5) Evaluation and Closure. For long-term monitoring and other cyclic projects, this phase occurs at the end of each field season and leads to an annual review of the project. For non-cyclic projects, this phase represents the completion of the project. After products are cataloged and made available, staff should work together to assess how well the project met its objectives and to determine what might be done to improve various aspects of the methodology, implementation, and formats of the resulting information. For monitoring protocols, careful documentation of all changes is required. Changes to methods, SOPs, and other procedures are maintained in a tracking table associated with each document. Major revisions may require additional peer review.

3.2 Data Life Cycle

During various phases of a project, data take on different forms and are maintained in different places as they are acquired, processed, documented, and archived. This data life cycle is characterized by a sequence of events that we can model to facilitate communication. These events involve interactions with the following objects:

- Raw data: analog data recorded by hand on hard-copy forms and digital files from handheld computers, GPS receivers, automated data loggers, etc.
- Working database: a project-specific database for entering and processing data for the current season (or other logical period of time). This might be the only database for short-

term projects where there is no need to distinguish working data for the current season from the full set of validated project data.

- Certified data and metadata: completed data and documentation for short-term projects, or one season of completed data for long-term monitoring projects. Certification is a confirmation by the project leader that the data have passed all quality assurance requirements and are complete and ready for distribution. Metadata records include the detailed information about project data needed for their proper use and interpretation.
- Master database: project-specific database for storing the full project data set, used for viewing, summarizing, and analysis; only used to store data that have passed all quality assurance steps.
- Reports and data products: information that is derived from certified project data.
- Edit log: a means of tracking changes to certified data.
- Outside databases and repositories: applications and repositories maintained by other entities. Also for sharing information with cooperators and the public.
- Local archives and digital library: local storage of copies of data, metadata, and other products generated by projects. Archives are for hard-copy items and off-line storage media, whereas the digital library is maintained live on a server.

Although the data life cycle may vary depending on specific project needs and objectives, the typical life cycle for Network projects proceeds as follows:

- Acquire data: for data recorded by hand in the field, data forms should be reviewed regularly (at least daily) for completeness, legibility, and validity in order to capture errors as close to their origin as possible.
- Archive raw data: copies of all raw data files are archived intact. Digital files are copied to the raw data folder for the project; hard copy forms are either scanned and placed in the active projects folder or are copied and placed in the archives. Archival or scanning of hard copy data forms may occur at the end of a season as a means of retaining all marks and edits made during the verification and validation steps.

- Data entry/import: analog data are entered manually; digital data files are uploaded to the working database.
- Verification, processing, and validation: verify accurate transcription of raw data; process data to correct data entry errors and remove missing values and other data flaws; validate data using database queries and other methods to capture missing data, out-of-range values, and logic errors.
- Documentation and certification: develop or update project metadata and certify the data set. Certification is a confirmation that the data have passed all quality assurance requirements and are complete and documented. It also means that data and metadata are ready to be posted and delivered.
- Archive versioned data set: copies of the certified data and metadata are placed in the project archive folder. This can be accomplished by storing a compressed copy of the working database or by exporting data to a more software-independent format (*e.g.*, ASCII text).
- Post data and update national databases: to make data available to others, certified data and metadata are posted to PCEMP repository. Note that data and data products may not be posted on public sites if they contain protected information about the nature or location of rare, commercially valuable, threatened or endangered species, or other natural resources of management concern (see Section 8.2.3).
- Upload data: certified data are uploaded from the working database to the master project database. This step might be skipped for short-term projects where there is no need to distinguish working data for the current season from the full set of certified project data.
- Reporting and analysis: certified data are used to generate data products, analyses, and reports, including semi-automated annual summary reports for monitoring projects. Depending on project needs, data might be exported for analysis or summarized within the database.
- Store products: reports and other data products are stored according to format and likely demand – either in the digital library, on off-line media, or in the document archives.

- Post products and update national databases (if appropriate): to make data available to others, reports and other products are posted to appropriate repositories.
- Distribute data and information: data, metadata, reports, and other products can be shared and distributed in a variety of ways – via the web-based repositories, by FTP or mailing in response to specific requests, or by providing direct access to project records to cooperators. Some limitations will be established to protect information about sensitive resources.
- Track changes: all subsequent changes to certified data are documented in an edit log, which accompanies project data and metadata upon distribution. Significant edits will trigger reposting of the data and products to national databases and repositories.

This sequence of events occurs in an iterative fashion for long-term monitoring projects, whereas the sequence is followed only once for short-term projects. For projects spanning multiple years, decision points include whether or not a separate working database is desirable and the extent to which product development and delivery is repeated year after year.

3.3 Integrating and Sharing Data Products

Once project data and derived products have been finalized, they need to be secured in long-term storage and made available to others. In future versions of this plan we will develop an appropriate system to accomplish this function. In addition, a process will be developed for the distribution of the data.

4 Data Acquisition and Processing

The PCEMP is being developed to help reach the PCEMP and SDCP goals. To accomplish this task, the program plans to both collect our own data as well as acquire data from other sources. This chapter describes steps to accomplish both tasks.

4.1 Program Activities

Biological monitoring projects will be the focus of PCEMP activities, but other endeavors will also be used including shorter-term research and inventory projects. Regardless of the type of activity, there are a range of tools that might be used for collecting data. For each there are opportunities and constraints to their use.

- Paper field data forms are the most common method of recording field data. Although inexpensive, more opportunities for errors exist during the data collection/data entry process. They also require neat, legible handwriting and very rigorous QA/QC.
- Field computers increase data collection and data entry efficiency. Data can be downloaded directly from field computers to office desktops, thereby eliminating manual data entry. Fewer chances for error exist as QA/QC checks can be built into the database, but these devices may not be the optimal choice if copious amounts of notes or comments must be recorded in the field. In addition, these portable units are subject to environmental constraints such as heat, dust, and moisture. When handheld computers are used for data entry in the field, the data should be downloaded daily to avoid potential loss of information. The use of a memory card that will store the data in case of damage to the unit or battery failure can prevent accidental loss of data.
- Handheld computers or Personal Digital Assistants (PDAs). The small size and relative low cost of these devices make them attractive options for collecting data in the field. Although they work well for small field projects, they are not powerful enough for large, data intensive field projects. PDAs can be customized to withstand a range of adverse environmental conditions fairly easily and inexpensively. Most run either Windows CE or Palm operating systems, which may require additional processing/programming to transfer/create the database structure in the field units.
- Tablet PCs have the same properties as most laptops and provide the user with the convenience of a touch screen interface. They are bulkier, more expensive, and harder to customize for fieldwork than the PDAs but are more powerful. They work well for field projects that are very data intensive. Because these units run Windows XP (Tablet

Edition), the project database can be directly transferred from desktop units to field units without additional programming steps.

- Automated data loggers are mainly used to collect ambient information such as weather data or water-quality information. Data loggers are an efficient method for recording continuous sensor data, but routine inspections are necessary, and environmental constraints, as well as power (*i.e.*, sufficient battery charge) and maintenance requirements, are potential pitfalls when using these instruments. Regular downloads should be required since physical memory is usually limited. Proper calibration is important.
- Permanently deployed devices are often very expensive, and data must be retrieved and batteries changed on a regular basis.
- Portable hand-held devices are deployed for sampling only during site visits. They are generally less expensive than permanently-deployed field units.
- GPS receivers are often used during fieldwork in network parks to collect location information. There are two main types of hand-held units and the choice of the two will depend on the accuracy of information required:
- Handheld recreation-grade GPS units are relatively inexpensive and are good for collecting general position information, but they are not recommended for obtaining high-accuracy location information.
- Mapping-grade GPS receivers are good for collecting highly accurate (sub-meter) location information, but they are more expensive than recreation-grade units, and more training is required to use these units correctly.

4.2 Non-Programmatic PCEMP Data

PCEMP will leverage resources by collecting and integrating information collected by other entities. Data collected by others falls into two types: those collected by other within the current County reserve network and those collected in other areas of Pima County or the surrounding region. Emphasis will be placed on gathering data collected within the existing County reserve system, particularly natural resource inventory data (e.g., species lists) as

well as historic photographs, maps, and voucher specimens. The agencies or organizations that compile these data may not have the expertise to apply proper quality control procedures and the capability to function as a repository and clearinghouse for the validated data. When the data are not kept in-house, data may be acquired via downloads from online databases or requests for data on CD, DVD, or other media.

4.2.1 Data Processing and Data Mining

Unlike data from PCEMP sources, much of the data collected from external sources must undergo some degree of processing to meet the standards of the PCEMP; however some of the basic processing steps are very similar.

All GIS data obtained from other entities will be stored in the proper format and include accurate spatial reference information and FGDC-compliant metadata.

All biodiversity data received from other entities, such as Breeding Bird Survey data sets, should be entered into a species database (yet to be created). In addition, if the data were taken from a report or published document, the reference must be entered into the monitoring EndNote database (as of May 2010 it contains >2,800 entries).

Particular emphasis will be placed on collecting information on the species that will be covered under the Section 10 permit. Known as the Covered Species Information Database (CSID), each year Pima County will query researchers and other governmental entities and non-governmental organizations regarding any data collected on covered species in the preceding year (see chapter 8 of the report). These data would be entered into a separate database.

Certain data sets will require more than the basic processing steps than described above. The level of data processing and mining required for external data sets will vary based on the source. Specific protocols might need to be developed that outline the necessary data processing requirements.

Remote sensing data sets (*e.g.*, satellite imagery or aerial photography) may require geospatial or spectral processing, depending upon the formats in which they are received.

Ideally, all spatial data sets will be received in a geo-referenced format and may require only geographic transformations.

Data mining will be particularly important at the beginning of Phase III as protocols are being developed. Data sources that will be queried include bibliographic/literature searches, geographic data, and biological/natural resources data. Much of these data are accessible via the Internet, but some can only be accessed through visits to local libraries, academic institutions, museums, and other land management agencies. All information collected during the data discovery process is maintained either electronically or in hard copy format, depending on how it was collected. Any geographic data sets collected during this process should be accompanied by FGDC-compliant metadata.

5 Data Quality Assurance / Quality Control

Data collected by the PCEMP is an invaluable resource that must be preserved over the long-term, but the long-term utility can be compromised by poor data. In particular, data of inconsistent or poor quality can result in loss of sensitivity and lead to incorrect interpretations and conclusions. The potential for problems with data quality increases dramatically with the size and complexity of the data set (Chapal & Edwards 1994).

Palmer (2003) defines *Quality Assurance* (QA) as “an integrated system of management activities involving planning, implementation, documentation, assessment, reporting, and quality improvement to ensure that a process, item, or service is of the type and quality needed and expected by the consumer.” He defines *Quality Control* (QC) as “a system of technical activities to measure the attributes and performance of a process, item, or service relative to defined standards.” QA procedures maintain quality throughout all stages of data development; QC procedures monitor or evaluate the resulting data products.

This section presents the procedures the PCEMP will employ to ensure that projects are of the highest possible quality. In short, we will establish and document protocols for the

identification and reduction of error at all stages in the data life cycle, with the goal of attaining 95%-100% accuracy.

Not long ago, maintaining data meant filling filing cabinets full of notebooks and paper. Now we are more likely to use computer hardware and software – technology that changes rapidly. If we expect our current data to be useful to future users, the data must survive changes in technology. We can promote data longevity through high-quality documentation and maintenance during all phases of data management. Well-documented data sets are especially important when sharing data.

5.1 Quality Assurance and Quality Control Mechanisms

QA/QC mechanisms are designed to prevent data contamination, which occurs when a process or event other than the one of interest affects the value of a variable.

Contamination introduces two fundamental types of errors into a data set (1) *Errors of commission* include those caused by data entry and transcription errors or malfunctioning equipment. They are common, fairly easy to identify, and can be effectively reduced upfront with appropriate QA mechanisms built into the data acquisition process, as well as QC procedures applied after the data have been acquired. (2) *Errors of omission* often include insufficient documentation of legitimate data values, which may 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.

QA/QC procedures applied to ecological data include four procedural areas (or activities), ranging from simple to sophisticated, inexpensive to costly:

- Defining and enforcing standards for electronic formats, locally defined codes, measurement units, and metadata
- Checking for unusual or unreasonable patterns in data
- Checking for comparability of values between data sets
- Assessing overall data quality

5.2 Data Collection

Careful, accurate recording of field observations in the data collection phase of a project will help reduce the incidence of invalid data in the resulting data set. All field sheets and field data recording procedures must be reviewed and documented in the protocol. Field crews understand the procedures and closely follow them in the field and this should be reinforced with proper training. Field crew members will be expected to understand the data collection forms, know how to take measurements, and follow the established procedures. Whatever the method of data collection, there must be procedures in place to reduce errors such as using project-specific data sheets (i.e., not field notebook). These will be highlighted in the specific protocols that are developed. After data have been collected, they must be entered into the appropriate database with the goal of 100% accuracy. This can be achieved by having the technician collecting data enter it into the database, but have a separate technician check all data against the original data sheet. Prior to data entry, the database should be structured so that the entry forms are efficient and easy-to-use (i.e., they look like the original datasheet) and that automatic error checking is built into the database, through use of auto-filled fields, range limits, pick lists, and constraints).

5.3 Verification and Validation Procedures

Data quality is ensured by applying verification and validation procedures as part of the quality control process. *Data verification* checks that the digitized data match the source data, while *data validation* checks that the data make sense. It is essential to validate all data as accurate and not misrepresent the circumstances and limitations of their collection. Although data entry and data verification can be handled by personnel who are less familiar with the data, validation requires in-depth knowledge about the data.

Data verification immediately follows data entry and involves checking the accuracy of the computerized records against the original source, usually hard copy field records, and identifying and correcting any errors. When data have been verified, the original data can be

archived. Among the tasks involved in data verification are: visual review at and after data entry, duplicate data entry, and simple summary statistics.

5.4 Data Quality Review and Communication

The PCEMP should require QA/QC review prior to communicating/disseminating data and information, and only data and information that adhere to the outlined quality standards should be released.

6 Data Documentation

Thorough, complete, and accurate documentation is critical during every stage of processing in the life cycle of a data set. At times, data sets appear to take on “lives of their own”; some are often found on multiple hard drives, servers, and other storage media. Others become hidden in outdated digital formats or in forgotten file drawers. In addition, once data are discovered, a potential user is often left with little or no information regarding the quality, completeness, or manipulations performed on a particular ‘copy’ of a data set. Such ambiguity results in lost productivity as the user must invest time in tracking down information or, worst case scenario, renders the data set useless because answers to these and other critical questions cannot be found. Therefore, data documentation must include an upfront investment in planning and organization.

Complete, thorough, and accurate documentation should be of the highest priority for long-term studies, and since long-term data sets are continually changing, this documentation must remain up-to-date. Data documentation involves the development of *metadata*, which at the most basic level can be defined as ‘data about data,’ or more specifically as information about the content, context, structure, quality, and other characteristics of a data set. Metadata provide the information necessary to relate the raw data to the underlying theoretical or conceptual model(s) for appropriate use and interpretation (Michener 2000). Additionally, standardized metadata provide a means to catalog data sets within Intranet and Internet systems, thus making these data sets available to a broad range of potential users.

Past efforts to standardize metadata content and format have focused primarily on geospatial data sets. Therefore, the term 'metadata' is often associated with documentation compliant with formal Federal Geographic Data Committee (FGDC) standards. However, in this plan, the term 'metadata' encompasses all forms of data documentation, including those for spatial and non-spatial tabular data, as well as project-level documentation.

The details of what products to use for compiling metadata will be more clearly articulated after the initiation of the PCEMP. However, a number of products and procedures are worth noting and will likely be adopted by the PCEMP. In particular, ArcCatalog is a management tool for GIS files contained within the ArcGIS Desktop suite of applications. With ArcCatalog, users can browse, manage, create, and organize tabular and GIS data. In addition, ArcCatalog comes with support for several popular metadata standards that allow one to create, edit, and view information about the data. There are editors to enter metadata, a storage schema, and property sheets to view the data. Users can view GIS data holdings, preview geographic information, view and edit metadata, work with tables, and define the schema structure for GIS data layers. Metadata within ArcCatalog are stored exclusively as Extensible Markup Language (XML) files.

Regardless of the product used, all products will be FGDC-compliant. Databases, in particular will have the following information:

- Description of the project
- Location of the project study plan and work plan
- Project leader's name and contact information
- Principal investigator's name and contact information
- Data set contact's name and contact information
- Description of the database model (entity-relationship diagram and data dictionary)
- Sensitive data issues, if appropriate
- Description of data verification/validation methods and results (data quality report)

- Certification of the data set
- Additional comments/documentation references, where appropriate
- Description of the database model
- Entity-relationship diagram
- Data dictionary
- Data quality report
- Sensitive data report
- Certification of the data set

7 Data Analysis, Reporting, and Dissemination

7.1 Data Analysis and Reporting

The success of the PCEMP depends upon providing information managers and other decision makers to empower them to make science-based decisions, as well as disseminating this information to a wider audience of other agency personnel, external scientists, and the general public. Data analyses are the means by which we transform data into this essential information. For long-term monitoring, data should be summarized at least annually and fully analyzed at three-to-five-year intervals (or as specified in the monitoring protocols) in order to detect trends in resource conditions. The information derived from data analyses will be conveyed through a variety of written reports and presentations. Project leaders are ultimately responsible for analyzing data and reporting the results, but this section discusses how data management activities can facilitate those activities through automated data summaries and reports.

Each project will have a schedule for data analysis and reporting requirements specified in the monitoring protocol, study plan, cooperative agreement, or contract. However, in general, the PCEMP will complete data analysis and reporting within one year of seasonal data collection or the end of the project. Yearly project reports will be required for all long-term projects. Annual reports should convey the past year's network monitoring activities with audience being agency personnel, USFWS and AZGF cooperators and other interested

scientists and member of the public. Relevant information may include numbers of samples and from what areas, data management activities, any changes made to the protocols, and the status of resources. Annual reports will be written as part of a yet-to-be created monitoring report series produced by Pima County. A summary of each annual report that highlights key points will also be produced in a 'brochure' format for distribution to a wider audience, including Board of Supervisors and the general public. Findings from project-specific reports will be "rolled up" into annual 'state of the County' reports describing the current trends and conditions of County resources

Comprehensive reports incorporating detailed data analyses, syntheses, and descriptions of trends in resource conditions for each parameter will be produced every 3-5 years or according to the individual monitoring protocol requirements. As with annual reports, these will be produced as part of the technical report series.

Technical reports are critical to providing periodic syntheses of relevant data and for ensuring that monitoring activities are accepted by the scientific community. Yet solely relying on producing technical reports would miss constituents that would be interested in program's findings, such as decision makers and the general public. To identify these key constituents and tailor products to them, the PCEMP will undertake a communications plan after permit issuance. By highlighting opportunities to disseminate data to non-technical audiences, it will help ensure the long-term relevance and success of the program.

7.2 Data Ownership

Pima County defines conditions for the ownership and sharing of collections, data, and results based on research funded by the County. All cooperative and interagency agreements, as well as contracts, should include clear provisions for data ownership and sharing as defined by the Pima County:

- All data and materials collected or generated using Pima County personnel and funds become the property of Pima County.

- Any important findings from research and educational activities should be promptly submitted for publication. Authorship must accurately reflect the contributions of those involved.
- Program personnel should share collections, data, results, and supporting materials with other researchers whenever possible. In exceptional cases, where collections or data are sensitive or fragile, access may be limited.

Guidelines will be developed for collaborative agreements regarding data ownership and timeframes and formats for submittal of data from outside cooperators. Products or deliverables would include, but are not limited to, field notebooks, photographs (hardcopy and digital), specimens, raw data, and reports. Details on formatting and media types that will be required for final submission will also be highlighted in the next phase of this report.

7.3 Data Distribution

One of the most important goals of the PCEMP is to integrate natural resource inventory and monitoring information into Pima County planning, management, and decision making. To accomplish this goal, procedures must be developed to ensure that relevant natural resource data collected by PCEMP staff, cooperators, researchers, and the public are entered, quality-checked, analyzed, documented, cataloged, archived, and made available for management decision-making, research, and education. Providing well-documented data in a timely manner to managers is especially important to the success of the program.

The PCEMP will make certain that:

- Data are easily discoverable and obtainable.
- Distributed data are accompanied by complete metadata that clearly establish the data as products of the PCEMP.
- Sensitive data are identified and protected from unauthorized access and inappropriate use (criteria will be developed later).
- A complete record of data distribution/dissemination is maintained.

Data distribution mechanisms will likely be the Internet, which will allow the data and information to reach a broad community of users. It is anticipated that the PCEMP will have our own website and this will be the primary outreach and dissemination portal for the monitoring information. This website should be linked to the Pima County Mapguide for the visual display of plot locations and access to data products. For this, the Santa Rita Experimental Range is a model for this approach (<http://ag.arizona.edu/srer/>).

7.4 Data Feedback Mechanisms

The PCEMP website should be developed to provide an opportunity for cooperators and the public to provide feedback on data and information gathered as part of the PCEMP. A ‘comments and questions’ link should be provided on the main page of the PCEMP site for this purpose.

8 Future Directions

The PCEMP has many of the basic design issues resolved, but many additional tasks lay ahead, including future development and implementation of the concept and guidelines that are introduced in this plan.

8.1 Staffing

Data management is about people and organizations as much as it is about information technology, database design, and applications. Therefore, an important aspect of the program will be for each member of the program team to have sets of roles and responsibilities. Because the PCEMP has not begun implementation, it is premature to assign such roles until funding is acquired. Once underway, the key to implementation of this data management plan will be hiring a data manager, who, as the title implies, will be the data stewardship leader and will be responsible for most of the data management activities. The role of the data manager and all other program staff will be articulated in the next phase of this data management plan.

8.2 Budgeting

The principles and guidelines outlined in this document demonstrate the level of detail that must be directed to data management. The level of detail may seem unreasonable or “overkill”, but it can not be stressed enough that the foundation of any long-term endeavor, such as being proposed by the PCEMP, rests on the data. With staff turnover, rapid advances in technology, and new methods, data quality can be jeopardized and therefore resources wasted. But even more important than resources, the greatest loss resulting from poor data management would include not identifying trends that occurred or for managers and decision makers to make haphazard or inappropriate decisions because of a lack of data.

Because of the importance of data management, approximately 25-30% of a program’s resources should be directed to this effort. Fortunately for the PCEMP, some of the elements of good data management are already part of the Pima County system including proper archiving of weather and GIS data, and system backup infrastructure. Ultimately, these elements will reduce program costs.

9 Literature Cited

- Brunt, J. W. 2000. Data management principles, implementation and administration. Pages 25-47 *in* W. K. Michener and J. W. Brunt, editors. Ecological data: Design, management and processing. Blackwell Science Inc., Malden, MA.
- Chapal, S. E., and D. Edwards. 1994. Automated smoothing techniques for visualization and quality control of long-term environmental data. Pages 141-158 *in* W. K. Michener, and J. W. Brunt, and Susan G. Stafford, editors. Environmental information management and analysis: Ecosystem to global scales. Taylor & Francis Ltd., London.
- Michener, W. K. 2000. Metadata. Pages 92-116 *in* W. K. Michener and J. W. Brunt, editors. Ecological data: Design, management and processing. Blackwell Science Inc., Malden, MA.
- Palmer, C. J. 2003. Approaches to quality assurance and information management for regional ecological monitoring programs. Pages 211-225 *in* D. E. Busch and J. C.

Supplement E: Data Management Plan for the PCEMP

Trexler, editors. Monitoring ecosystems: interdisciplinary approaches for evaluating ecoregional initiatives. Island Press, Washington, DC.

Pima County. 2000. Draft preliminary Sonoran Desert Conservation plan. Draft report to the Pima County Board of Supervisors for the Sonoran Desert Conservation Plan, Tucson, AZ.

RECON Environmental Inc. 2006. Draft Pima County multi-species conservation plan. Report to the Pima County Board of Supervisors for the Sonoran Desert Conservation Plan, Tucson, AZ.

RECON Environmental Inc. 2007. Ecological effectiveness monitoring plan for Pima County: Phase 1 final report. Report to the Pima County Board of Supervisors for the Sonoran Desert Conservation Plan, Tucson, AZ.