# RESEARCH SUB-PROGRAM

## MANAGEMENT AND DISSEMINATION OF DIGITAL DATA AND RELEVANT INFORMATION COLLECTED THROUGH THE GREEN PLAN

**November 1997**

**COESA Report No.: RES/MON-009/97**

**Prepared by:** Ken Denholm, Bruce MacDonald, Fenghui Wang and Gary Watson
Data Management Project Team, Ontario Land Resource Unit
Greenhouse and Processing Crop Research Centre
Agriculture and Agri-Food Canada, Guelph, Ontario

**On behalf of:** Research Branch, Agriculture and Agri-Food Canada,
Pest Management Research Centre (London)
1391 Sandford St.
London, Ontario N5V 4T3

# *FORWARD*

This report is one of a series of **COESA** (Canada-Ontario Environmental Sustainability Accord) reports from the Research Sub-Program of the Canada-Ontario Green Plan. The **GREEN PLAN** agreement, signed Sept. 21, 1992, is an equally-shared Canada-Ontario program totalling $64.2 M, to be delivered over a five-year period starting April 1, 1992 and ending March 31, 1997. It is designed to encourage and assist farmers with the implementation of appropriate farm management practices within the framework of environmentally sustainable agriculture. The Federal component will be delivered by Agriculture and Agrifood Canada and the Ontario component will be delivered by the Ontario Ministry of Agriculture and Food and Rural Assistance.

From the 30 recommendations crafted at the Kempenfelt Stakeholders conference (Barrie, October 1991), the Agreement Management Committee (AMC) identified nine program areas for Green Plan activities of which the three comprising research activities are (with Team Leaders):

1. **Manure/Nutrient Management and Utilization of Biodegradable Organic Wastes** through land application, with emphasis on water quality implications
   - **A.** Animal Manure Management (nutrients and bacteria)
   - **B.** Biodegradable organic urban waste application on agricultural lands (closed loop recycling) (Dr. Bruce T. Bowman, Pest Management Research Centre, London, ONT)

2. **On-Farm Research**: Tillage and crop management in a sustainable agriculture system. (Dr. Al Hamill, Harrow Research Station, Harrow, ONT)

3. **Development of an integrated monitoring capability** to track and diagnose aspects of resource quality and sustainability. (Dr. Bruce MacDonald, Centre for Land and Biological Resource Research, Guelph, C

The original level of funding for the research component was $9,700,000 through Mar. 31, 1997. Projects will be carried out by Agriculture and Agri-Food Canada, universities, colleges or private sector agencies including farm gro

This Research Sub-Program is being managed by the Pest Management Research Centre, Agriculture and Agri-Food Canada, 1391 Sandford St., London, ONT. N5V 4T3.

_____
Dr. Bruce T. Bowman, Scientific Authority
E-Mail:

**Green Plan Web URL:** http://gis.lrs.uoguelph.ca/AgriEnvArchives/gp/gphompag.html

***The following report, approved by the Research Management Team, is reproduced in its entirety as received from the contractor, designated on the previous page.***

August, 1997

# Executive Summary

Existing GIS databases and spatial analysis tools have been extensively used in studies of Ontario agricultural systems and their environmental impacts. However, limited data and information have been collected to support on-farm decision-making and provide linkages between broad-level analysis and research results at the site or farm-level. The amount of information converted to computerized databases for multi-source data integrations is even more limited. One of the objectives of the research activities under the Canada-Ontario Agriculture Green Plan, particularly the resource monitoring sub-program, is to fill some of these data gaps and to develop capabilities for monitoring and assessment of agricultural resources. Development of a computerized data management and analysis system using geographic information system (GIS) technology is one of the projects of the Green Plan Research Program.

The objectives of this project are to:

1) develop a data management and analysis system with the capability to correlate, interpret and document linkages between the data collected through Green Plan research projects;
2) carry out data and spatial analysis to verify the content and quality of digital data received from the Green Plan Monitoring component;
3) provide standardized, documented copies of data for assessment and use by other projects and agencies and for comparison with future measurements.

To achieve the objectives, the following activities were conducted by this project during 1994-1996:

- The review of related methodological and technical issues and development of a framework, including: i) the system architecture and components, ii) database model and structure; iii) structure of data organization in relation to existing data sets; iv) data validation and quality control procedures; v) procedures for developing metadata (data about data) and a metadata interface.

- The upgrade and re-configuration of the computer hardware and software at the Ontario Land Resource Unit (OLRU), which was necessary for implementing the data management system.

- The design and prototyping of an Entity-Relationship (ER) model database structure.

- The design and prototyping of a web-based metadata system.

- A routine data check, validation, conversion and documentation of each data set received from related Green Plan research projects.

A data management and analysis system has been developed and used for managing the data collected through the Green Plan Research Program, especially data collected through the Resource Monitoring

Sub-program. The system was built on the existing GIS system at OLRU and currently maintained by OLRU technical staff. The system consists of: 1) GIS and DBMS hardware and software; 2) database sub-system; 3) data management and quality control procedures; 4) metadata and meta-information sub-system; 5) GIS analytical tools and procedures (built-in functionality of commercial software, in-house routines, application procedures, etc.)

The entity-relationship data model was used as the logic tool to design database tables and structures. Three sets of entities were identified: project entities, sampling entities and property entities. A limited number of data tables (structures/templates) were designed based on these entities and their relationships. The data table structures are relatively standard but flexible enough to accommodate data from a wide range of sources. New data can be added to the database system without having to create new tables or columns. This can ensure: 1) the efficiency of data entry, update and maintenance; 2) the reduction in redundancy and increased consistency and integrity of data within the database system. With the standard but flexible database structure, data collected through other agricultural resource and agri-environmental research programs or projects can also be stored and managed in the system.

Metadata and user interface, integral parts of the system, have been developed and demonstrated. In addition to metadata tables attached to each data set, such as the project entity table and the sampling entity table, the document type metadata in hypertext format were created for each data set. The Canadian National Metadata Standards were used as a basis in defining metadada components and contents. Other meta-like documents containing explanatory information related to specific data sets, including reports from private contractors and related publications are also considered as part of the metadata and meta-information system. The web-based metadata interface with query and search capability has shown good potential in improving data and information service for related research and decision-making.

# Sommaire

On a fait grand usage de bases de données de SIG (systèmes d'information géographique) et d'outils d'analyse spatiale dans des études de systèmes agricoles et de leurs incidences environnementales menées en Ontario. Toutefois, peu de données et d'information ont été recueillies pour appuyer la prise de décisions à la ferme et établir des liens entre les analyses d'ensemble et les résultats des recherches effectuées sur les lieux des études ou à la ferme. Et la quantité d'information versée dans des bases de données informatisées provenant de sources multiples est encore plus limitée. Un des objectifs des activités de recherche poursuivies dans le cadre du Plan vert Canada-Ontario en agriculture, en particulier le sous-programme de surveillance des ressources, consistait à combler ces lacunes et à développer des moyens de surveillance et d'évaluation des ressources agricoles. Un des projets du programme de recherche du Plan vert visait la mise au point d'un système informatisé de gestion et d'analyse de données faisant appel à la technologie des SIG.

Voici quels étaient les objectifs de ce projet :

1) Mettre au point un système de gestion et d'analyse de données permettant la corrélation, l'interprétation et la documentation des liens entre les données recueillies pendant l'exécution des projets de recherche du Plan vert.
2) Effectuer une analyse de données et une analyse spatiale pour vérifier la nature et la qualité des données numériques amassées au cours des activités de surveillance du Plan vert.
3) Fournir des copies uniformisées et documentées d'ensembles de données en vue de leur évaluation et de leur utilisation par les responsables d'autres projets et divers organismes et de leur comparaison avec les résultats de futures mesures.

Pour atteindre ces objectifs, on a poursuivi les activités suivantes de 1994 à 1996 :

- Examen des aspects méthodologiques et techniques et établissement d'un cadre, à savoir :

  i) l'architecture et les composantes du système ainsi que le modèle et la structure de la base de données;
  ii) la structure de l'organisation des données par rapport aux ensembles de données existants;
  iii) les méthodes de validation des données et de contrôle de qualité;
  iv) les méthodes de production de métadonnées (données sur les données) et d'une interface de métadonnées.

- Mise à niveau et reconfiguration du matériel informatique et des logiciels utilisés par l'Équipe pédologique de l'Ontario, opérations nécessaires à la mise en oeuvre du système de gestion de données.

- Conception d'une structure de base de données fondée sur un modèle entité-relation et production d'un prototype de cette structure.
- Conception d'un système de métadonnées relié au réseau Internet et production d'un prototype de ce système.

- Vérification, validation, conversion et documentation systématiques de chaque ensemble de données produits au cours de recherches menées dans le cadre du Plan vert.

On a mis au point un système de gestion et d'analyse de données qu'on a utilisé pour la gestion des données recueillies dans le cadre du programme de recherche du Plan vert, en particulier le sous-programme de surveillance des ressources. Le contrôle de ce système, basé sur le SIG utilisé par l'Équipe pédologique de l'Ontario, est assuré par le personnel technique de ce groupe. Le système comprend les éléments suivants :

1) matériel informatique et logiciels du SIG et du système de gestion de base de données;sous-système de base de données;
2) procédures de gestion des données et de contrôle de la qualité;
3) sous-système de métadonnées et de méta-information;
4) outils et procédures d'analyse de SIG (fonction intégrée d'un logiciel commercial,
5) programmes écrits sur place, procédures d'application, etc.).

Le modèle de données entité-relation a servi d'outil logique pour la conception des tables et des structures de la base de données. Trois ensembles d'entités ont été identifiés : entités de projets, entités d'échantillonnage et entités de propriétés. Un nombre limité de tables de données (structures/modèles) ont été conçues à partir de ces entités et de leurs relations. Les structures des tables de données sont relativement standard, mais assez souples pour accepter des données de nombreuses sources différentes. On peut ajouter des données à la base sans avoir à créer d'autres tables ou colonnes, ce qui permet de :

1) garantir l'efficacité de la saisie, de la mise à jour et du contrôle des données;
2) réduire la redondance et d'accroître la cohérence et l'intégrité des données dans la base.

Grâce à cette structure de base de données standard mais souple, on peut stocker et gérer dans le système les données résultant d'autres programmes ou projets de recherche agro-environnementale et d'études des ressources agricoles.

Les métadonnées et l'interface utilisateur, qui font partie intégrante du système, ont fait l'objet d'une démonstration. Outre les tables de métadonnées reliées à chaque ensemble de données, comme la table d'entité de projets et la table d'entité d'échantillonnage, on a créé des métadonnées de type document en format hypertexte pour chaque ensemble. La définition des composantes et de la nature des métadonnées s'appuyait sur les normes canadiennes relatives aux métadonnées. D'autres documents s'apparentant à des métadonnées qui contiennent des informations explicatives sur des ensembles de données précis, dont des rapports d'entrepreneurs et des publications connexes, sont également considérés comme faisant partie des métadonnées et du système de méta-information. L'interface d'accès aux métadonnées est accessible sur Internet et comporte des fonctions d'interrogation; elle pourrait améliorer les services de données et d'information pour l'exécution de travaux de recherche connexes et la prise de décisions à cet égard.

# Table of Contents



## List of Tables

## List of Figures

# 1. Background and Objectives

Reliable information about agri-environment and agricultural resources is needed to provide a basis for wise decisions to ensure sustainability of agricultural production and health of agroecosystems. In today's environment the information is frequently obtained from geospatial data stored and managed in computerized information systems. The usefulness of GIS technology for data management, manipulation and analysis in support of resource management and planning decisions is well recognized. The recent development of internet technology has made it possible to develop web-based Agricultural Information Integration and Exchange System (MacDonald et al, 1997). In the past, the principal data applications have been carried out on a project by project basis and generalized to deal with problems at broad regional levels. There is, however, a developing recognition of the importance of consistent databases to derive resource management decisions on an ongoing basis. In this context, when questions arise which require data about agroecosystems and agricultural resources, the first recourse is the existing computerized databases. The condition of existing data, however, often limits or precludes reliable integration and analysis (Siderelis, 1991). Frequently, it will be necessary to supplement the existing data with additional information specific to the problem at hand. In an ongoing GIS and database system, the additional information will be documented and added to become a part of the data resource of geospatial information. As a result, a comprehensive database will be developed over time. The decisions based on these data will have the benefit of consistent information to support them and the overall resource management and planning process will have continuity.

Existing GIS databases and spatial analysis tools have been extensively used in studies of Ontario agricultural systems and their environmental impacts (for example, MacDonald et al, 1995; Jarvis et al; 1996, MacDonald and Gleig, 1996). However, limited data and information have been collected to support on-farm decision-making and provide linkages between broad-level analysis and research results at the site or farm-level. The amount of information converted to computerized databases for multi-source data integrations is even more limited. One of the objectives of the research activities under the Canada-Ontario Agriculture Green Plan, particularly the resource monitoring sub-program is to fill some of these data gaps and to develop capabilities for monitoring and assessment of agricultural resources, including a computerized data management and analysis system using geographic information system (GIS) technology. In contrast to many past studies, the current monitoring projects will deliver to Agriculture and Agri-food Canada not only a written report providing analyses and interpretations but also a substantial volume of both raw and processed data in digital formats. As a result, a federal government in-house research project was initiated. The objectives of the project were to: 1) develop a data management and analysis system with the capability to correlate, interpret and document linkages between the data collected through Green Plan research projects; 2) carry out data and spatial analysis to verify the content and quality of digital data received from the Green Plan Monitoring component; 3) provide standardized, documented copies of data for assessment and use by other projects and agencies and for comparison with future measurements.

# 2. Main Research and Development Activities

To achieve the objectives of the study, the following activities were conducted during 1994-1996:

- The review of the existing GIS system and data sets related to land resource and agriculture at OLRU.

- The review of related methodological and technical issues and development of a framework, including: i) the system architecture and components, ii) database model and structure; iii) structure of data organization in relation to existing data sets; iv) data validation and quality control procedures; v) procedures for developing metadata (data about data) and metadata interface (Wang,1995).

- The upgrade and re-configuration of the OLRU computer hardware and software, which is necessary for implementing the data management system (Denholm and Wang, 1995).

- The design and prototyping of an Entity-Relationship (ER) model based database structures (Watson,1996).

- The design and prototyping of a web-based metadata system (Wang et al, 1996; Denholm et al, 1997).

- A routine data check, validation, conversion and documentation of each data set received from related Green Plan research projects.

Because of the special nature of this project, namely, the data check and validation, data documentation and dissemination are dependent on the final deliverables of other research projects, some data checking, validation and documentation are therefore underway or will be conducted when the expected data are received by this project. The procedures and policy for data dissemination, especially through the internet will be finalized in conjunction with the Green Plan Technology Transfer Program(TT).

# 3. Methods and Procedures

The standard methods which are commonly used by system developers and database designers (Watson, 1996), such as requirement analysis, component and structure design and integration information, the database prototyping (e.g. ER diagraming), as well as the web technology etc. have been used in the development and implementation of the data management system. The major steps of the system design and implementation can be outlined as:

1) Overall design of system structure and components
2) Design and prototyping of database structure
3) Development of criteria and procedure of data validation and quality control
4) Development of metadata and user interface

## 3.1 Design of system structure and components

### 3.1.1 Identification of the key system components

The data management and data analysis system for Green Plan is designed as a standard database management system with considerations of current and potential user (internal and external) needs. As the first step of the design, the major components of the system were identified. They are:1) GIS and DBMS hardware and software; 2) database system; 3) data management and quality control procedures; 4) metadata system; 5) GIS analytical tools and procedures (built-in functionality of commercial software, in-house routines, application procedures, etc) and 6) GIS personnel.

### 3.1.2 Analysis of specifications

After the system components were identified, the specific requirements and technical details, particularly, the capacities and functionalities of hardware and software were analyzed. This analysis provided the basis for the system implementation. Table 1 gives an example of analysis of hardware and software requirements (Table 1).

### 3.1.3 Upgrade and re-configuration of existing GIS system

The data management system for the Green Plan was designed to be built on the existing GIS system at the Ontario Land Resource Unit at Guelph, Ontario. Some of the required hardware and software components which the existing system did not have were purchased and added into the existing system. The major system upgrade and re-configuration included migration from a DEC/VMS workstation to a UNIX workstation, addition of a 10 gigabyte disk storage unit and upgrade of the LAN network software.

## Table 1. Hardware and software requirements for the Green Plan data management system

| Capacity/functionality required | Hardware and software components required | |
|---|---|---|
| | **Basic** | **Additional** |
| Data acquisition and input | - A GPS receiver<br>- A digitizer<br>- A scanner<br>- Arc/Info capability to generate GIS data sets from ASCII data sources [1] | - ArcScan module of Arc/Info (for converting scanned raster data into Arc/Info format) |
| Large volume of data storage, data backup and restore.<br><br>Data security management | - 10 GB disk storage unit<br>- A tape drive<br>- A CD drive | - 20 - 30 GB disk storage unit<br>- One more tape drive ( for alternative data backup) |
| Database maintenance, data conversion, validation, correlation and re-compilation<br><br>Data analysis and modelling | - A UNIX workstation<br>- Arc/Info and Oracle<br>- AML, and INFO programming, Oracle-SQL programming<br><br>- TIN, GRID | - Spatial Database Engine (SDE) module of Arc/Info (for more data management functions) |
| On-line data and metadata query and display. | - Arcplot and AML<br>- ArcView | |
| Data/maps/report output | - A color inkjet/laser printer<br>- A black and white laser printer | |
| Data/metadata transfer/ distribution | - CD recordable<br>- Agri-Net and gateway internet access | - Internet Map Server<br>- Upgrade from intranet web server to internet web server |

[1]The mention of a trademark, proprietary product or vendor does not imply endorsement by Agriculture and Agri-Food Canada to the exclusion of other products or vendors

## 3.2 Design and prototyping of the database

This section draws extensively on work carried out by Watson (1996) as part of his M.Sc. research supported by this project. His research is described in detail in Watson (1996). Some adaptation of this work was required to meet the needs of the Green Plan data management system.

Database design is mainly the development of the structure of the database. The basic objectives of database design of this study are:

- to design and standardize the database structure to accommodate a much broader range of data sources related to agri-environment andagricultural resources;
- to improve the procedures of data input and conversion;
- to deliver a limited number of standard data tables for data collected through the Green Plan, especially the resource monitoring program under the research sub-program.

The processes involved in the design and prototyping included:

1) requirement analysis;
2) conceptual design, including review and selection of a data model and data model mapping;
3) physical design and implementation;
4) prototyping and evaluation of the performance and usability of the designed data tables and structure;
5) data conversion.

These procedures are illustrated in Figure 1.

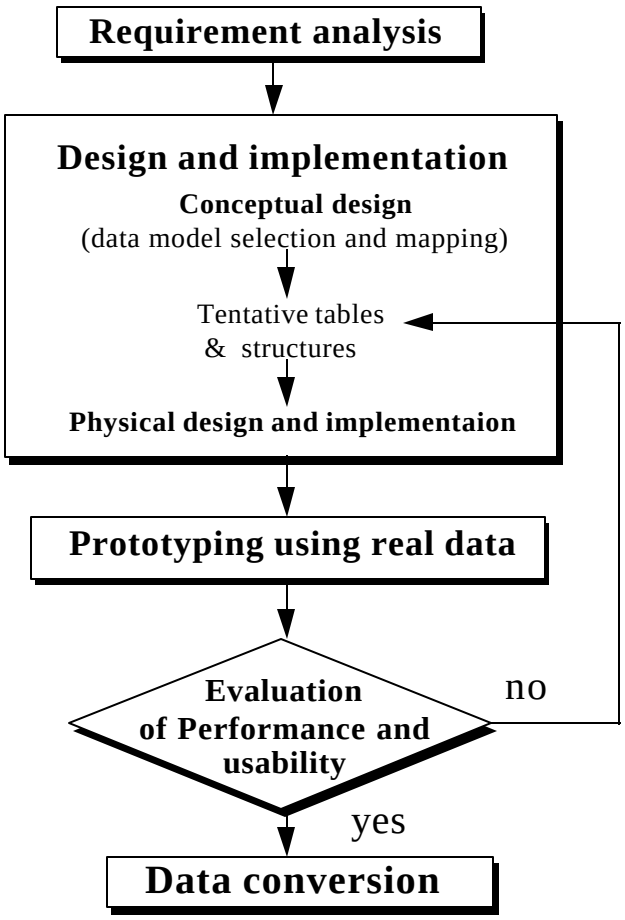


Figure 1. Procedures of database design and prototyping

### 3.2.1 Requirement analysis

As the first step of database design, the general and specific requirements for achieving the above mentioned objectives were analyzed.

The following aspects were considered during the requirement analysis:

- Kinds and characteristics of data to be accommodated
- Data management requirements
- Data use requirements (usability and end-user functions requirements)
- Standardization requirements for data sharing, exchange and future use.

1) Requirements identified from data sources

As a mandatory requirement, this data management system had to accommodate the data collected through the 10 research projects of the Resource Monitoring program (GP-RES-3) under the Green Plan Research Sub-program (GP-RES).

A non-mandatory requirement was for this data management system to accommodate the data or metadata collected through:

- Green Plan research programs other than the Resource Monitoring, such as the On-farm Research (GP-RES-2) and the Manure Management (GP-RES-1) programs;

- other Green Plan programs/activities, such as the Rural Conservation Clubs (RCC) and the Wetlands, Woodlands and Wildlife (3W) programs;

- past agri-environmental research programs or projects, such as the Soil and Water Environmental Enhancement Program (SWEEP), and the Tillage 2000 (T2000) program;

- future agri-environmental research programs or projects in Ontario.

The data generated by these programs or projects consists of spatially referenced observations (primary survey or sampling data), derived data by interpretation and data manipulation, metadata as well as the relevant documents of findings.

In terms of the spatial-temporal features, the primary data collected through surveys, field experiments and sampling can be grouped into:

1) *point oriented sampling or experiment data* at site, plot or field level with specific slope position, soil layer, and specific time under specific treatments (crop, tillage, etc.);
2) *areal oriented survey data* at soil polygon, watershed, township, county or municipality level which do not have a specific time dimension in contrast to the

site sampling data.

Table 2 provides a summary about the basic characteristics of data to be accommodated by this data management system as mandatory or optional requirements.

**Table 2. The basic characteristics of data collected through the Resource Monitoring program and other related program under the Canada-Ontario Agriculture Green Plan**

| Programs | Projects | Kinds of data | Contents and format of data | Spatial dimension | Time dimension |
|---|---|---|---|---|---|
| **Resource Monitoring*** (GP-RES-3) | 3.4, 3.5, 3.6, 3.7 3.11 | areal oriented survey data | - Arc/Info coverage<br>- associated attribute data (landform, soil, hydrography and land use, etc.) in Arc/Info format | soil polygons watershed, township or county level | relatively time invariant |
| | 3.1, 3.2, 3.3, 3.7, 3.8, 3.10 | point oriented sampling or experiment data | - coordinates of sampling locations<br>- raw or manipulated attribute data ( basic soil physical, chemical properties, resident biomass and organic carbon, $^{137}Cs$, soil enzymes, contaminants in water, etc.) in ASCII, spreadsheet, database or other format | specific site/ plot, slope position, soil horizon. | at specific time point (date, month, year) |
| **On-Farm Research**** (GP-RES-2) **Manure Management**** (GP-RES-1) | part | point oriented sampling or experiment data | - coordinates of sampling locations<br>- raw or manipulated attribute data (soil and other biophysical conditions, land use and land management practices, carbon, nitrogen and other nutrient transform-ations and losses, etc.) in ASCII, spreadsheet, database or other format | specific site/ plot, slope position, soil horizon. | specific time point (date, month, year) |
| **Rural Conservation Club (RCC)**** (GP-RCC) | all | location and descriptive data | - coordinates of project locations<br>- descriptive data about the projects and activities in ASCII, spreadsheet, or other format | specific site/farm | projects/ activities conducted during specific time period |

* it is a requirement for this project to manage the data collected through the program
** it is optional for this project to manage the data collected through the program

Table 2. (continued)

| Programs | Projects | Kinds of data | Contents and format of data | Spatial dimension | Time dimension |
|---|---|---|---|---|---|
| **Wetlands, Woodlands and Wildlife (3W)**** (GP-3W) | all | location and descriptive data | - coordinates of project locations<br>- descriptive data about the projects and activities in ASCII, spreadsheet, or other format | specific site | projects/ activities conducted during specific time period |

** it is optional for this project to manage the data collected through the program

As a result, the database tables and structures must have high flexibility with the capability to accommodate data with different:

1) spatial dimension (from sampling site to broad region, not only 2 dimension location, but also 3 dimension position and soil profile);
2) time dimension (from hourly, daily sampling data to time invariant data);
3) kinds, units and levels of attributes (climate, terrain, soil, water, biota, etc.);
4) format (ASCII, spreadsheet, database, GIS , etc.).

2) Data management requirements

In terms of the requirements of data management, some basic guidelines were used in the database design. For example, the database should:

- minimize redundancy and maintain relative data independence;
- minimize fragmentation of data (data to be relatively consolidated );
- maintain high consistency and integrity of the data;
- be flexible to update or to add new data;
- be easily converted to optional formats.

Minimizing data redundancy is always a concern when developing a database. Redundant data can lead to inconsistencies in the database. Redundant data also increase data storage requirements. While advances in computer technology continue to increase data storage capacities, data generation abilities have also increased. As a result, data storage volumes continue to be a concern and there is an ongoing need to minimize data redundancy. Some redundancy, however, may be acceptable if it provides a

significant enhancement to the functioning of the database. Such redundancies should, however, only be included to meet a performance objective. Similarly, maintaining data independence is a common goal for most modern database systems. By maintaining data independence, the options for database use and expansion and the opportunities for linking to external applications are optimized (Watson, 1996).

It is also important, for efficient database maintenance and management, to design limited numbers of standard database tables and have consistent definitions for the same or similar data fields in different data tables and in different data sets. Technically, the data table structure and format should be flexible for updating, adding new data and for easy data conversion from one format to another.

3) User requirements

The potential users of Green Plan data are primarily the AAFC professional and technical staff, researchers from both public and private organizations and university students. In spite of their discipline backgrounds and agri-environmental knowledge it may be difficult for them to understand the procedures of data collection and field sampling or experimentation or, the basic terms used in describing the data. Detailed information about data collection, access and use is required. The database design must ensure the data are understandable to the users and maximize the data accessibility and usability.

From the point view of 'ease-of-use', the most important requirement for database design is to identify metadata components and design metadata tables for holding the information about where, how and under what conditions (e.g. land use or crop, tillage system) the data were collected, who collected the data and so on. This can ensure the correct use of data, and also assist in searching or evaluating the data to determine whether or not it is useful for the application of interest. The metadata tables in database format can have overlaps with separately developed metadata in document format (which will be discussed in a later section) but with different emphasis. For most of the data collected through the Green Plan research projects, the essential elements of the metadata include: the project information, factors to define the sampling location, sampling method and time, etc.

4) Standardization requirements

For the requirements of standardization, two aspects were considered. One is standardization within this data management system. Another is to follow some of the national and AAFC's information system ( e.g. CanSIS/NSDB) standards or conventions.

## 3.2.2 Selection of a data model

A data model is a set of concepts for the design, construction and use of a database. Based on the review by Watson (1996), data models can be grouped into two classes: classical and semantic. The classes are differentiated by the structural components and the amount of semantics inherent within those structures. Examples of classical data models are flat-files, hierarchical, network and relational data

models. Classical models are record oriented in that they conform to the record-based structures used by computers. They describe the concepts upon which a DBMS is founded.

Semantic models, on the other hand, convey meaning rather than describing the physical components of the database. Semantic models are more reflective of the real world and provide useful tools for database design. They are also used to describe the associations between the data elements and how they have been organized and categorized to the database users. The most widely recognized semantic model is the entity-relationship (ER) model (Chen, 1976). Entities are things or objects in the real world, with an independent existence. Relationships describe the association between entities. Attributes are the properties that describe the entities and relationships. There are basic ER models and extended ER (EER) models. The EER model removes some limitations of the basic ER model and has been used for designing soil and other GIS databases (Fernandez and Rusinkiewicz, 1993). Figure 2 illustrates the EER model and its components. To achieve the basic goals of this study, the EER model was selected as a basic logic tool for database design.
In addition to the EER model, the database design approach and logic of the soil performance and management (P/M) file of CanSIS (MacDonald and Strzelczyk, 1986), especially the approach of treatment definition (combination of experimental factors and factor levels), was adopted for defining the

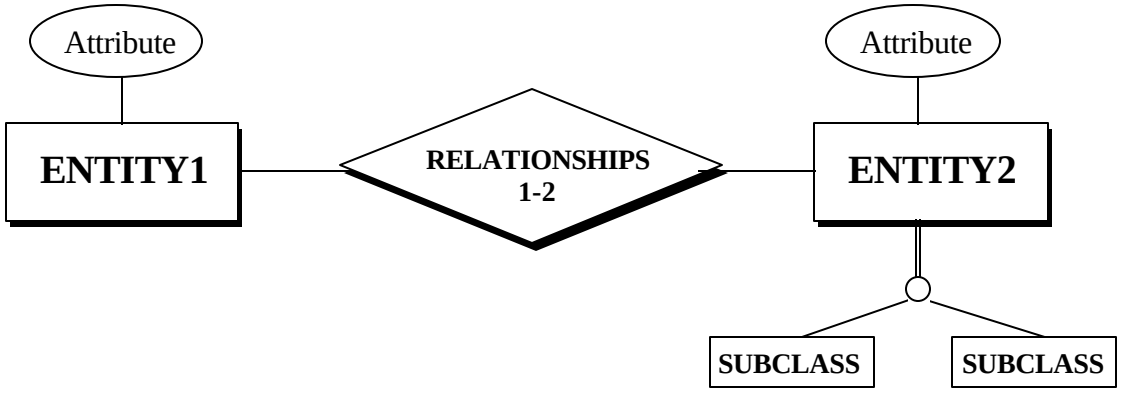


Figure 2. Entity-Relationship data model and its components (adapted from Fernandez and Rusinkiewicz, 1993)

sampling entity.

### 3.2.3 Entity-Relationship representation of the major Green Plan data components

In order to provide a logic framework for designing data tables and structures an Entity-Relationship analysis of the major Green Plan components was conducted.

As described in previous sections, the data generated by the Green Plan and other related programs or projects consists of primary survey or sampling data, data derived by interpretation and data manipulation, metadata as well as the relevant documents of findings. These data can be translated into either entities or entity attributes and their relationships can also be defined. To meet the different management requirements for actual data and the metadata, the entities, relationships and attributes were divided into two levels, a data-level and a metadata-level (Figure 3).

As shown in Figure 3, there are three kinds of entities: project entities (PE), sampling entities (SE) for point, plot or field oriented sampling data (or survey entities for areal oriented survey data) and property entities (PPE). Project entities and sampling entities, to a large extent, are metadata-level entities, while

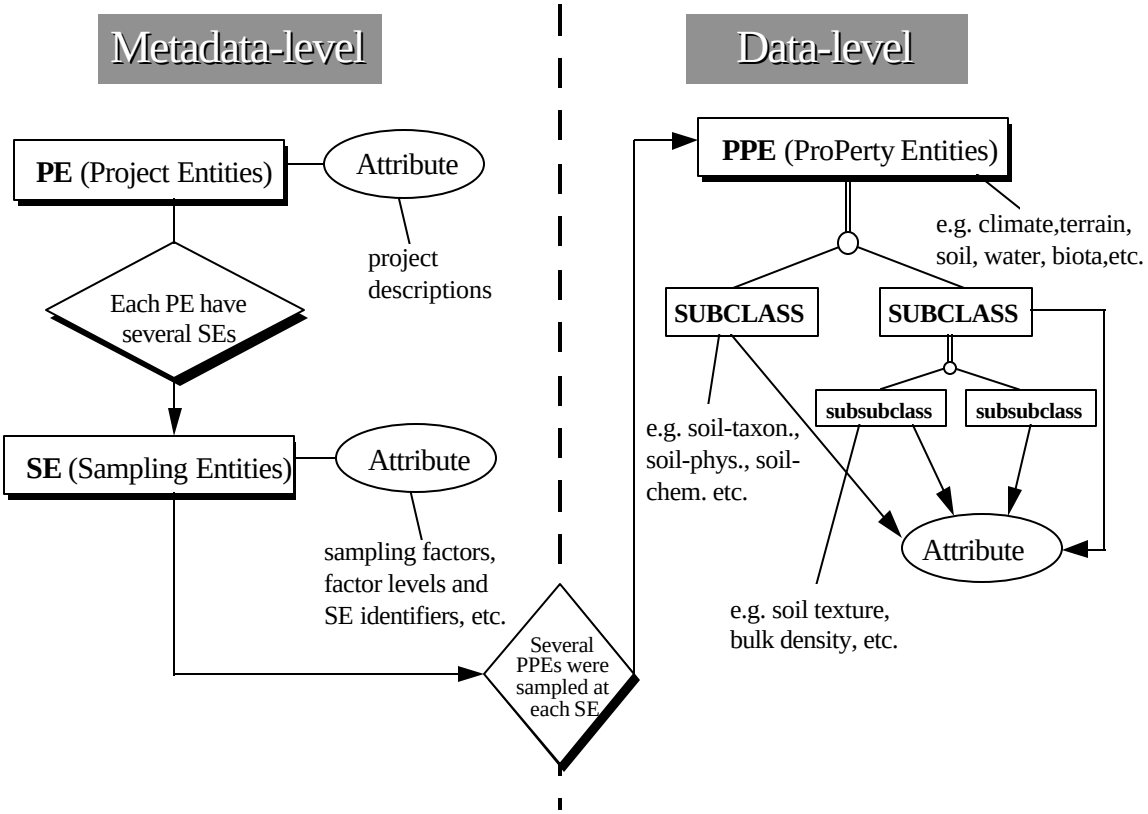


Figure 3. Entity-Relationship representation of the Green Plan data components
(entity sets indicated with rectangles, relationships with diamonds and attributes with ovals)

property entities are data-level entities. Each kind of entity is defined or described by a set of attributes which form one or more data tables or metadata tables.

### 3.2.4 Design of data tables

In order to have flexibility for future modifications of data table design, we proposed data table based entity sets as shown in Figure 3 and left the number of tables in each table group open. Three groups of tables at metadata and data levels were designed (for details, see Table 7 in Section 4.3 and Appendix 2)

### 3.2.5 Prototyping and evaluation of performance and usability

Prototyping was conducted with Tillage 2000 data (Watson, 1996) during the early stages of this project. It provided useful experience for re-designing the database structure. The performance and usability were tested and evaluated with the prototype data tables against the user requirements and data management requirements as documented in section 3.2.1

### 3.2.6 Data conversion

This is the final step of database design and implementation. The real data sets which have been received from separate research projects were converted into new standard data tables and structures.

## 3.3 Development of procedures for data validation and quality control

### 3.3.1 Defining the criteria of data validation and quality control

Traditionally, data validation, quality assurance and quality control are key components and procedures in resource survey, mapping and cartographic production. While there are some differences in specifications and standards, the basic aspects and content are very similar (Aronoff, 1989; Banting, 1993). Based on the commonly used data quality elements, the following criteria (Table 3) were adapted for use for Green Plan data validation and quality control:

**Table 3 Major aspects and preliminary criteria for Green Plan data validation and quality control.**

| Data validation and quality control aspects | Positional | Attribute | Other |
|---|---|---|---|
| **Completeness** | - complete areal coverage | - complete data files, items as they were designed.<br>- no missing records. | - a data set must be a logically complete group. |
| **Consistency** | - same geographic features (boundaries, points, etc.) in different data layers/sets should be logically and physically consistent. | - the name, structure, item names and definition of same data file for different areas should be consistent.<br>- attributes representing same features or themes should have same name and definition. | - data format should be consistent.<br>- conditional inconsistency of data format is allowable when conversion between data formats is possible. |
| **Precision/Resolution**<br><br>Note: precision used here refers to **fineness** of data converted to GIS data format or afterward conversion; resolution refers to **minimum polygon** (vector) **or grid-cell** (raster) **size.** | - to use Arc/Info precision control procedure to set proper precision (e.g. single or double precision for coverage, tolerance settings, etc.).<br><br>- to use **minimum map unit/area** for each map scale (compilation) to control resolution in feature integration and generalization (Table 8). | - to keep the precision of data capture and/or measurement in the field as possible<br><br>- consistent and reasonable precision should be kept in rounding off digits, quantitative conversion, reclassifying of attributes in data file ( logical judgement needed).<br><br>- proper temporal resolution should be taken into consideration in multi--temporal data files. | - this component should be documented in metadata documents of each data layer/ set. |
| **Accuracy**<br><br>Note: accuracy used here refers to **closeness** of measurement or estimation to true values or the values accepted as being true (NCDCDS, 1988). | - provisionally, CanSIS criteria (i.e. on manuscript, the point on coverage should within 0.5 mm distance to standard reference point)<br>- to be judged by experts opinion.<br>- accuracy of GPS field measurement needs to be tested and defined. | It is defined for different measurement and estimation. | |

Table 3. (continued)

| Data validation and quality control aspects | Positional | Attribute | Other |
|---|---|---|---|
| **Reducibility (least redundancy)** | n/a | - avoiding unnecessary redundancy in attributes, records, etc. | (note: it is also an issue in designing data file structure, see section 3.2) |
| **Timeliness** | - any change to features should be updated in a timely fashion. | - changes to associated attributes should also be updated promptly. | - metadata and meta-documents should be updated accordingly. |
| **Clearness** | n/a (Arc/Info feature classifications are very clear) | - each data file name, item name and definition (especially for derived attributes) should be kept as clear as possible) | - high clearness is required for metadata and meta-documents. |

The criteria for the minimum map unit/area (MacDonald et al, 1992) were used for checking and re-formatting the areal oriented survey data in Arc/Info format (Table 4)

**Table 4. Minimum map unit/area used by CanSIS (after MacDonald et al, 1992)**

| Map Scale | Approximate Dimensions for Map of Canada (cm) | Minimum Area (ha) Displayed in 0.25 $cm^2$ |
|---|---|---|
| 1:25 million | 21 x 18 | 1,562,500 |
| 1:5 million | 107 x 91 | 62,500 |
| 1:1 million | 535 x 455 | 2,500 |
| 1:500 thousand | 1070 x 910 | 625 |
| 1:100 thousand | 5350 x 4550 | 25 |

One important issue in data validation and quality control is the elimination of errors or their control to within acceptable levels. Muller (1987) and Aronoff (1989) have given in-depth discussions on error issues. Aronoff has described the common sources of errors in GIS data input, manipulation, output and use of results (Table 5). These descriptions were used as a reference when this data check and validation work were carried out as part of this project..

**Table 5. Common sources of error encountered in using GIS (adapted from Aronoff, 1989)**

| Operations | Sources of Errors |
|---|---|
| Data collection | errors in field data collection<br>errors in existing maps used as source data<br>errors in the raw remote sensing data |
| Data input | inaccuracies in digitizing or scanning caused by operator and equipment<br>inaccuracies inherent in geographic feature |
| Data storage & conversion | insufficient numeric precision<br>insufficient spatial precision |
| Data manipulation | inappropriate class intervals<br>boundary errors<br>error propagation as multiple overlays are combined<br>slivers caused by problems in overlay operation and/or<br>unavoidable slivers not properly eliminated |
| Data output | scaling inaccuracies<br>error cause by inaccuracies of output device or conversion procedure<br>error cause by instability of the medium |
| Use of results | the information may be incorrectly understood<br>the information may be inappropriately used. |

## 3.3.2 Procedures of data validation and quality control

Based on the above criteria, data validation and quality control were conducted in conjunction with data analysis. The following steps were used in this project (Figure 4):

1) Systematically check of raw data by data providers before the data were delivered.

2) Screening level validation in conjunction with report review by the professional members of the project team provided suggestions about important aspects of data validation, e. g. valid ranges and limits of data values, etc.

3) Detailed data validation by the technical personnel (data manager and GIS analyst) of this project team (validation methods depend on kinds of data to be validated).

4) Re-check with data providers, to see whether the data they provided were properly converted and represented in a GIS environment.

5) Reporting of validation and quality control results (creating data validation report and metadata).

6) Review of validation report and results by the professional members of the project team.

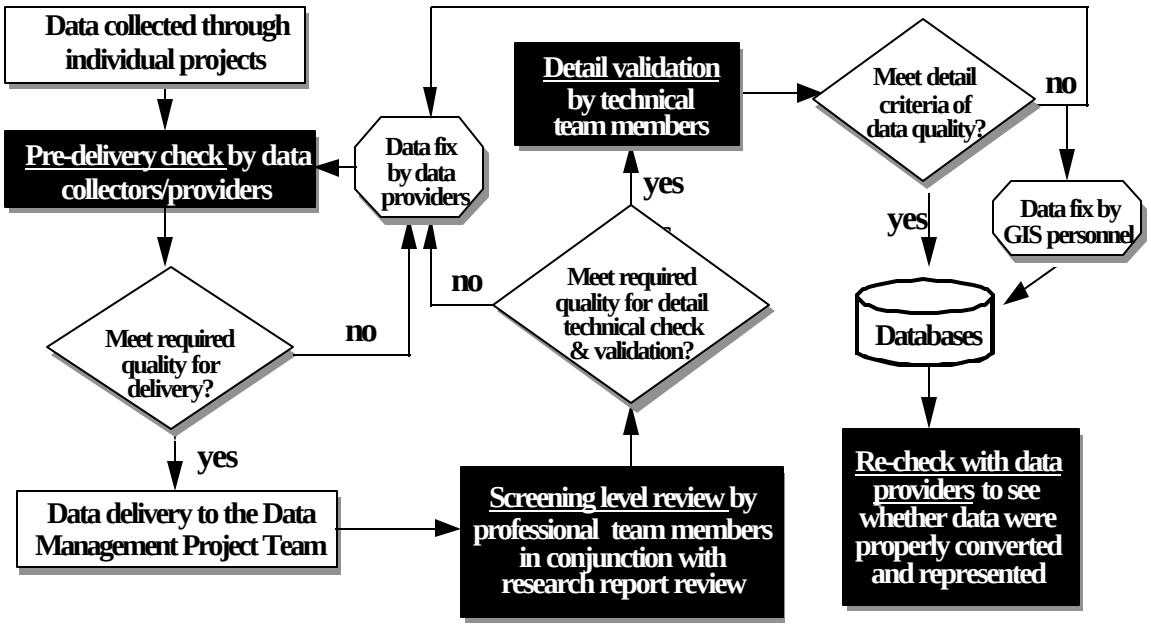


Figure 4. Procedures of data validation and quality control

## 3.4 Development of metadata and user interface

### 3.4.1 Identifying types and components of metadata

Metadata are data about the content, quality, condition, and other characteristics of data (FGDC, 1994; ESRI, 1995). In simple words, metadata is data about data.

To meet the requirements of this project, three types of metadata were identified:

1) **metadata as data tables** attached to each data set. This type of metadata has been documented in section 3.2., and includes the project table and sampling entity table (Table 3);
2) **standard metadata in document format** (hypertext documents, each data set has one);
3) **data catalog** (comprehensive list of all and related data).

Compared with standard metadata in document format, the metadata tables are project/ application oriented metadata which can have more customized components and contents. The components and contents of metadata tables (project table and sampling entity table) are listed and described in Table 5. Additional metadata tables can be designed and added to data sets which require more detailed meta-information to enable the user to understand and correctly use the data.

For the metadata in document format, there are national or international standards to follow (FGDC, 1994). Based on the Canadian National Standards (Canadian General Standards Board, 1995), the following basic components should be included in this kind of metadata:

- **Identification** (owner, scientific contact, data manager and contact for access);
- **Descriptive information** (content, areal coverage, vertical coverage and time coverage)
- **Data updating details**
- **Size**
- **Details of data collection** (base map used, source of data, circumstances and details associated with the data collection, map projection, coordinate system, geodetic datum, accuracy and comments on data quality)
- **Availability** (access policy, charges, etc.)
- **Information about the host computer** (OS, configuration)
- **Documentation and demonstration** to help the user .

The US Federal Geographic Data Committee's Content Standards for Digital Geospatial Metadata (FGDC, 1994) were also used as a reference (Table 6) in creating metadata by this project.

**Table 6. Metadata components defined by the US Federal Geographic Data Committee**

| Main components | | Support components | |
|---|---|---|---|
| 1) | identification information | 8) | citation information |
| 2) | data quality information | 9) | time period of content information |
| 3) | spatial data organization information | 10) | contact information |
| 4) | spatial reference information | | |
| 5) | entity and attribute information | | |
| 6) | distribution information | | |
| 7) | metadata reference information | | |

In general, a hypertext format is used for this kind of metadata. Tests by this study have also indicated that the HTML-based metadata format shows promise not only because it supports a convenient organization and structure for the information but also because it facilitates the search through a large and varied collection to select the portions which are relevant to a specific application. This latter aspect could be considered as part of the metadata interface.

## 3.4.2 Development of user interface

Metadata interface is the interface between data users and metadata or data. The basic function of the interface is to provide data users with an easy way to access the metadata documents with certain interactivity, such as interactive search and query. Additional functions could be developed to include interactive data display and mapping and intelligent consulting (expert system with a web interface), etc. All these could be implemented with GIS and RDBMS (Relational Data Base Management Systems) functionality or internet web technology. In this project, the web-based interface was chosen for experimental development.

# 4. Results and Deliverables

The results and deliverables of this study include four parts: 1) the prototype data management system, the metadata sub-system and the user interface; 2) data organization and data model; 3) the procedures of data management and analysis; 4) databases which have been checked, validated, converted into standard database tables and documented.

## 4.1 Systems and user interface

A GIS-based data management system was developed and used for managing the data collected through the Green Plan Research Program, especially data collected through the Resource Monitoring Sub-program. The system was built on the existing GIS system at the Ontario Land Resource Unit (OLRU), Greenhouse and Processing Crop Research Centre (GPCRC), Agriculture and Agri-Food Canada (AAFC) and is currently maintained by OLRU technical staff.

The major hardware and software components include: 1) a DEC Alpha UNIX server with 10 Gb disk storage capacity; 2) a commercial GIS - Arc/Info by Environment System Research Institute, Toronto; 3) a commercial DBMS by Oracle Corporation. The data are currently managed using the INFO sub-system of the Arc/Info GIS and will be converted and transferred into the Oracle system. The metadata and metadata-information sub-systems are also running on the DEC Alpha UNIX server with web-based interfaces which can be accessed within AAFC's wide area network - Agri-Net (full internet access in the future is suggested and under consideration).

With the further development and extension of the system, it coul become one of the core sources of agri-environmental data and information in Ontario and Canada. Figure 5 illustrates the major components of the system and the relationships with past and future agri-environmental related programs in Ontario.

A prototype interface was designed with an open structure of multiple selections in consideration of the multi-level needs of potential users of the Green Plan data. Figure 6 provides the optional interfaces of data dissemination and access, including the web-based interface. Figure 7 illustrates the basic options of the metadata user interface of the data management system at OLRU's web site (http://guelra.agr.ca/) within AAFC's intranet. Some of these options, such interactive mapping with Arc/Info are still under development.

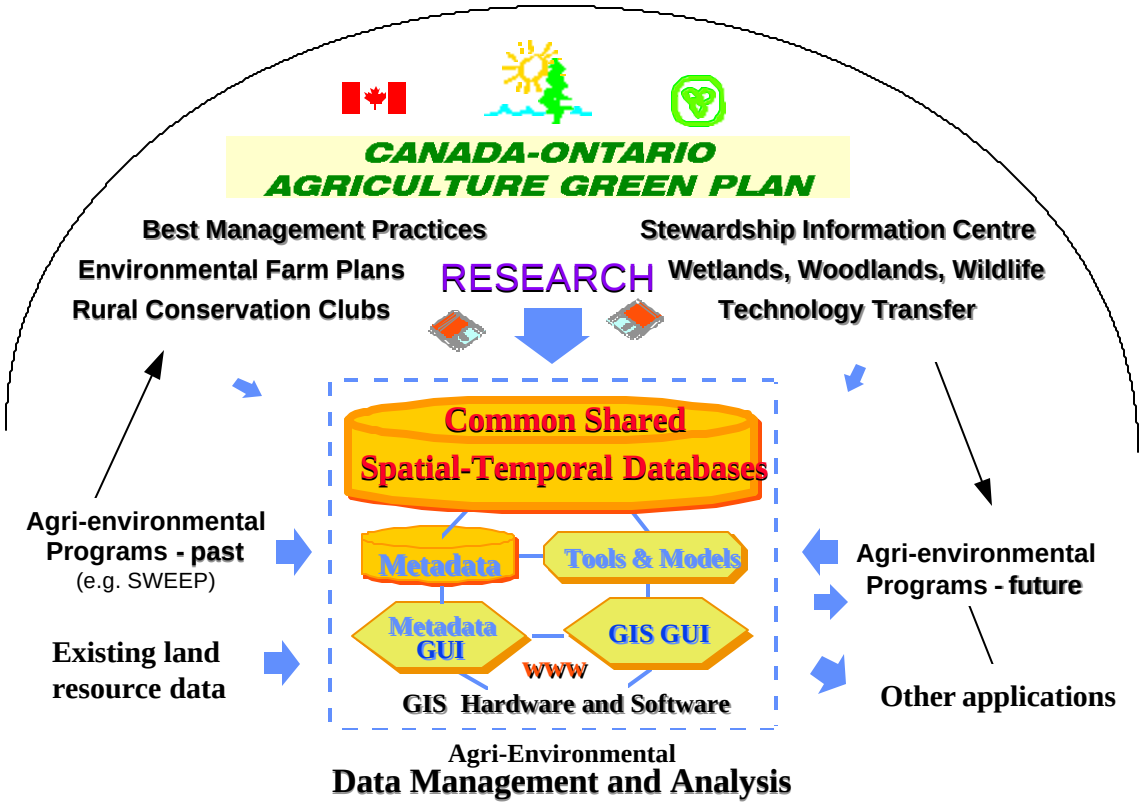


Figure 5 Major components of the Green Plan data management and analysis system

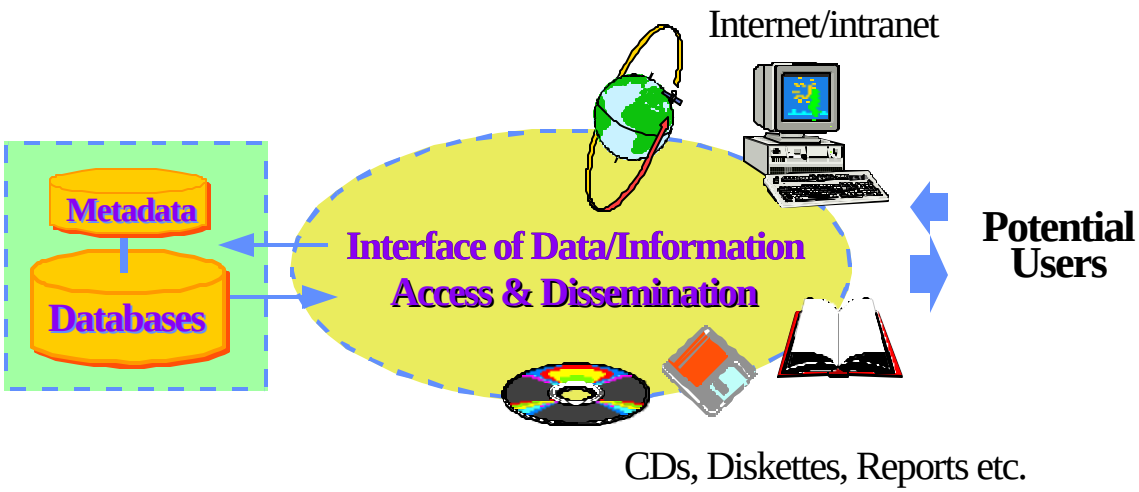


Figure 6. The optional interfaces of data access and dissemination

In addition to metadata for the main Green Plan data sets, the meta-documents in HTML format (i.e. hypertext-based metadocument) were also created to provide links to relevant information, especially published documents related to agricultural systems and management practices. The linkages between relevant topics are established by keywords and are searchable using a document search engine (Hardy et al, 1995). The hypertext documents provide pathways for navigating metadata and meta-documents and can also be considered as navigating metadata (Green Plan Technology Transfer Committee, 1997; MacDonald et al, 1997).


Figure 7 The metadata interface of the Green Plan data management system at OLRU's web site (http://guelra.agr.ca/) within AAFC's intranet.

## 4.2 Data organization and data model

A great deal of efforts of this project has been devoted to the development of data organization and a data model. The methodological framework and the model have been detailed in section 3.2 of this report and in the thesis of Watson (1996). In this section, we summarize the data tables designed to organize Green Plan data by using the Entity-Relationship data model (Table 9). Because both areal oriented survey data and point or site oriented sampling data have spatial entities associated with them, the native feature attribute tables of geo-spatial entities (point, line, area/polygon and surface) in Arc/Info were adopted. Specific details of table names as currently specified are given in Appendix 2. The data fields and definition of each table are provided in Appendix 3.

**Table 7. Entity sets, table groups and tables for the Green Plan data**

| Entity set | Table group | Table | Key attribute* |
|---|---|---|---|
| **Metadata Level** | | | |
| Project entities | Project descriptions | Project table (PJT) | proj-id (mandatory)<br>start-yr (optional) |
| Sampling entities | Sampling entity definition | - Sampling factor table (SFT)<br>- Sampling factor level table (FLT)<br>- Sampling entity identification table (SIT) | proj-id (mandatory)<br>start-yr (optional)<br>se-id (mandatory) |
| **Data Level** | | | |
| Property entities | Property observations/ measurements | Currently one generic table, property table (PPT) was proposed | proj-id (mandatory)<br>start-yr (optional)<br>se-id (mandatory)<br>horizon<br>udepth<br>ldepth<br>date/time<br>replicate# |
| Geo-spatial entities | geo-entity feature attributes (Arc/Info feature attribute table) | - Polygon attribute table (PAT)<br>- Point attribute table (PAT)<br>- Arc attribute table (AAT)<br>... | Arc/Info coverage<br>user-ID |

* used to link to other tables or attributes used to uniquely identify property value

## 4.3 Procedures

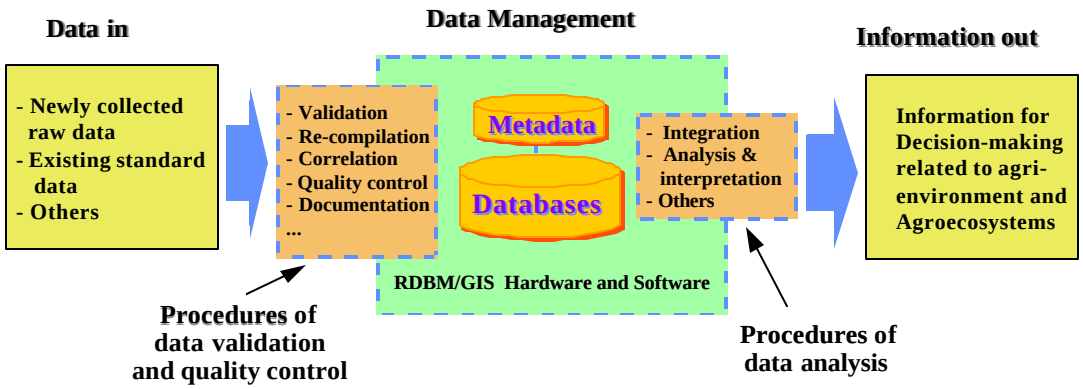


Figure 8 The use of data management and analysis procedures

The procedures developed by this project consist of three groups: 1) database design and prototyping procedures; 2) data validation and quality control procedures and 3) data analysis and modeling procedures.

Most of the procedures have been documented and illustrated in Section 3.2 and in the thesis by Watson (1996). The procedure of defining the data sampling entities (SE) and sampling entity identifiers ( SE-ID) and an example of data query procedure are provided in Appendix 4 and 5 .

The procedures of data validation and quality control and data analysis were developed based on the procedures used by CanSIS and OLRU GIS technical staff. Figure 8 illustrates the use of the data management and analysis procedures in a flow of 'data-in-and-information-out'.

## 4.4 Databases

The data sets, including those currently under validation and conversion, which can be delivered by this project are as follows:

1) Green Plan research site database
2) Green Plan 3W project/activity site database
3) Green Plan RES-3-A database for areal oriented survey data
4) Green Plan RES-3-B database for site oriented sampling data

**Table 8. Major data sets delivered by the Green Plan data management project**

| Data set name | Descriptions |
|---|---|
| GP-RES-SITES | **Green Plan** Research subprogram, **research site database** |
| GP-3W-SITES | **Green Plan** Wetlands, Woodlands and Wildlife(**3W**) subprogram, **project/activity site database** |
| GP-RCC-SITES | **Green Plan** Rural Conservation Clubs (**RCC**) subprogram, **project/activity site database** |
| GP-RES-3-A | **Green Plan** Research subprogram, Resource Monitoring subsubprogram (**RES-3**), A - **database for areal oriented survey data**<br><br>(3.4, 3.5, 3.6,3.7,3.11) |
| GP-RES-3-B | **Green Plan** Research subprogram, Resource Monitoring subsubprogram (**RES-3**), B - **database for point oriented sampling data**<br><br>(3.1,3.2,3.3,3.8,3.10) |

As an extension, projects under the other two research sub-programs (GP-RES-1 and GP-RES-2) have been selelected to assemble and deliver data to this project. The data sets will be created with the same database structure and conversion procedures.

# 5. Conclusions

A GIS-based data management and analysis system has been developed and used for managing the data collected through the Green Plan Research Program, especially data collected through the Resource Monitoring Sub-program. The system consists of: 1) GIS and DBMS hardware and software; 2) database sub-systems; 3) data management and quality control procedures; 3) metadata and meta-information sub-system; 4) GIS analytical tools and procedures (built-in functionality of commercial software, in-house routines, application procedures, etc.)

The entity-relationship(ER) data model was the logic tool used to design database tables and structures. Three sets of entities were identified, project entities, sampling entities and property entities. A limited number of data tables were designed based on these entities and their relationships. The data table structures are relatively standard but flexible enough to accommodate data from a wide range of sources. New data can be added to the database system without creating new tables or columns. This can ensure: 1) efficiency of data entry, update and maintenance; 2) reduced redundancy and increased consistency and integrity of data within the database system. Data collected through other agricultural resources and agri-environmental research programs or projects can also be stored and managed in the system.

Metadata and user interface have been developed and demonstrated as an integral part of the system. In addition to metadata tables attached to each data set, such as the project table and the sampling entity table, document type metadata in hypertext format were created for each data set. The Canadian National Metadata Standards were used as a basis in defining metadata components and contents. Other meta-like documents containing explanatory information related to specific data sets, including reports from private contractors and related publications, are also considered as part of the metadata and meta-information system. The web-based metadata interface with query and search capability has shown good potential in improving data and information service for related research and decision-making.

Because data from some projects have just recently been delivered to this project or will be delivered to this project by the end of the Green Plan Program, some data checking, validation, conversion and documentation are still underway or will be conducted when the expected data are received. The procedures and policies for data dissemination, especially through the internet, will be finalized in conjunction with the Green Plan Technology Transfer Program.

This system has great potential to become one of the core sources of agri-environmental data and information in Ontario and Canada, particularly if the system is further developed and data holdings are extended to include data collected through the agri-environmental research programs/projects (past and future) other than the Green Plan. It will provide strong database support for the analysis and assessment of agricultural resources and agroecosystem health in Ontario.

# 6. References

Aronoff, S., 1989, Geographic Information Systems: A management Perspective, WDL Publications, Ottawa, Canada, pp.133-293

Banting, D.,1993, Data quality management in GIS, The Operational Geographer. 10 (4), pp. 22-26.

Canadian General Standards Board, 1995, The Directory of Information Describing Geo-Referenced Data Sets (CAN-CDSB-171-171.3-95), Ottawa, Canada.

Chen, P., 1976, The entity-relationship model - Toward a unified view of data. ACM Trans. Database Syst. 1(1), pp. 9-36.

Coleman, D. J., and J. D. McLaughlin, 1994, Building a global spatial data infrastructure: usage paradigms and marketing influences, In: *Proceedings of 1994 Canadian Conference on G I S* , Ottawa, Canada, pp.31-45.

Denholm, K. A., and F. Wang, 1995, Development of a computerized data management system for the Green Plan Monitoring Program, Poster presentation at 1995 Green Plan Annual Meeting, March 28-29, 1995, London, ON. Canada.

Denholm, K.A., K.B. MacDonald, and F. Wang, 1997, A fork in the road to information acquisition, In *Proceedings of the 11th Annual Symposium on Geographic Information Systems (GIS'97)*, February 17-20, 1997, Vancouver, BC. Canada. pp. 521-523.

Ecological Stratification Working Group, 1995.,A National Ecological Framework for Canada, Agriculture and Agri-Food Canada, Research Branch, Centre for Land and Biological Resources Research and Environment Canada, State of The Environment Directorate, Ecozone Analysis Branch, Ottawa/Hull. Report and national map at 1:7,500,000.

ESRI. 1995. Metadata management in GIS. White Paper Series. Environmental Systems Research Institute, Inc. Redlands, California, USA.

Fernandez, R. N., and M. Rusinkiewicz, 1993, A conceptual design of a soil geographic database for geographical information systems. Int. J. Geographical Information Systems.7:525-539.

FGDC (Federal Geographic Data Committee), 1994, Content Standards for Digital Geospatial Metadata(June 8), Federal Geographic Data Committee. Washington, D.C. USA.

Green Plan Technology Transfer Committee, 1997, Final Report of the Technology Transfer Committee, Canada-Ontario Agriculture Green Plan Resaerch Report.

Hardy, D.R., M.F. Schwartz, and D. Wessels, 1995, Harvest User's Manual, Version 1.3, Technical Report Cu-CS-743-94 Department of Computer Science, University of Colorado, Boulder, USA.

Jarvis, I.E. and K.B. MacDonald, 1997, Characterization of the state and trends of agriculture in the Great Lakes Basin: A bi-national, ecological approach, In *Proceedings of the 11th Annual Symposium on Geographic Information Systems (GIS'97)*, February 17-20, 1997, Vancouver, BC. Canada. pp. 524-528.

MacDonald, B.K and K.W.G. Vlentine, 1992, CanSIS/NSDB: A general description, Centre for Land and Biological Resource Research, Agriculture Canada, Ottawa, Canada.

MacDonald, K.B. and Z.S. Strzelczyk. 1986. The Canada Soil Information System (CanSIS); Users' output manual for the soil performance file of CanSIS. Land Resource Research Institute. Research Branch. Agriculture Canada. Ottawa, Ontario, Canada. 83 pp.

MacDonald, K.B., D.A. Swayne, R. Denzer, A. Hess, and D.Jessberger, 1997, Development of an Agricultural Information Integration and Exchange System, Green Plan Technology Transfer Program, Canada-Ontario Agriculture Green Plan Research Report.

MacDonald, K.B., I. Jarvis, and F. Wang, 1995, Regional agricultural practices and their potential for land and water contamination, In: *Summary of Achievements - Great Lakes Water Quality Program (1989 - 1994)*, Agriculture and Agri-Food Canada, pp. 326-358.

MacDonald, K.B. and B. Gleig, 1996, Indicator of risk of water contamination: notrogen component, Progress Report No 7, Apri-Environmental Indicator Project, Agriculture and Agri-Food Canada, Guelph, Ontario, Canada. 39 pp.

Mapping Science Committee 1993, *Toward a Coordinated Spatial Data Infrastructure for the Nation*, US National Research Council, National Academy Press, Washington, D. C. USA, 169 pp.

McLaughlin, D., 1991, Towards national spatial data infrastructure, In: *Proceedings of 1991 Canadian Conference on GIS* , Ottawa, Canada, pp.1-5.

NCDCDS (National Committee for Digital Cartographic Data Standards), 1988, The proposed digital standard for digital cartographic data, *American Cartographer,* Vol. 15 pp. 28.

Siderelis, K. C., 1991, Land resource information systems. In:Geographical information systems: principles and applications. Vol 2: applications (ed Maguire, Goodchild, Rhind and Marlow) Longman Group UK Ltd pp. 261-273.

Wang, F., 1995, Implementation of a data management system for Green Plan data (DMSGP) - Methodolgical framework, Interim Report No. 2. Ontario Land Resource Unit, Agriculture and Agri-Food Canada, Guelph, Ontario, Canada.

Wang, F., K.B. MacDonald, and K.A. Denholm, 1996. Operationalizing the procedures of data management for assessing agroecosystem health. In *Proceedings of the 10th Annual Symposium on Geographic Information Systems (GIS '96)*, March 19-22, 1996, Vancouver, BC. Canada. (CD-ROM version)

Watson, G., 1996. The development of a standard approach for managing data generated through land resource field research, Msc. Thesis, Department of Land Resource Sciences, University of Guelph, Guelph, Ontario, Canada. 197 pp.

# Appendix 1. Key Terms and Acronyms

| | |
|---|---|
| 3W - | Green Plan Wetlands, Woodlands and Wildlife Sub-program |
| AAFC - | Agriculture and Agri-Food Canada |
| AML - | Arc Macro Language (Arc/Info) |
| Arc/Info - | A commercial GIS produced by Environment System Research Institute, USA |
| CanSIS - | Canadian Soil Information System |
| DBMS - | Data Base Management System |
| ER - | Entity-Relationship (data model) |
| GIS - | Geographic Information System |
| GP - | Green Plan (refer to The Canada-Ontario Agriculture Green Plan in this report) |
| GUI - | Graphic User Interface |
| RES - | Green Plan Research Sub-program |
| HTML - | Hypertext make-up language |
| INFO - | DBMS component of Arc/Info. |
| NSDB - | National Soil Data Base (Canada) |
| OLRU - | Ontario Land Resource Unit |
| Oracle - | A commercial DBMS produced by Oracle Corporation |
| RCC - | Green Plan Rural Conservation Club Sub-program |
| SWEEP - | Soil and Water Environmental Enhancement Program |
| T2000 - | Tillage 2000 Program |
| TT - | Green Plan Technolgy Transfer Sub-program |

# Appendix 2. Data tables for major data sets of the Green Plan

| **Data set name** (short name used in data management system) | **Program/projects and description** | **Tables*** | | |
|---|---|---|---|---|
| | | **Metadata tables** | **Geo-spatial entity tables** (in native GIS format) | **Property tables** |
| GP-RES-SITES | Green Plan Research sub-program, research site database | gp-res-sites.pjt<br>gp-res-sites.sft<br>gp-res-sites.flt<br>gp-res-sites.sit | gp-res-sites.pat | gp-res-sites.ppt (location class only) |
| GP-3W-SITES | Green Plan Wetlands, Woodlands and Wildlife(3W) sub-program, project/activity site database | gp-3w-sites.pjt | gp-3w-sites.pat | gp-3w-sites.ppt (location class only) |
| GP-RCC-SITES | Green Plan Rural Conservation Clubs (RCC) sub-program, project/activity site database | gp-rcc-sites.pjt | gp-rcc-sites.pat | gp-rcc-sites.ppt (location class only) |
| GP-RES-3A | Green Plan Research sub-program, Resource Monitoring subsubprogram (RES-3), A - database for areal survey oriented projects<br><br>(3.4, 3.5, 3.6,3.7,3.11) | gp-res-3a.pjt | separate .pat files for each sub-data set | property attributes will be associated in geo-spatial entity table |
| GP-RES-3B | Green Plan Research sub-program, Resource Monitoring subsubprogram (RES-3), B - database for site sampling oriented projects<br><br>(3.1,3.2,3.3,3.8,3.10) | gp-res-3b.pjt<br>gp-res-3b.sft<br>gp-res-3b.flt<br>gp-res-3b.sit | gp-res-sites.pat | gp-res-3b.ppt |

* for meanings of table extension, see Table 3

# Appendix 3. Data fields and definition for major Green Plan data tables

| Table group | Table | Table field | |
|---|---|---|---|
| | | **Name** | **Note** |
| Project descriptions | Project table (.PJT) | - proj_name<br>- program<br>- sub-program<br>- subject<br>- proj-type<br>- **proj-id**<br>- **start-yr**<br>- proj-leader<br>- agency<br>- cooperator<br>- duration<br>- report#<br>- o-ref<br>- citation<br>- sec-level<br>- access-level<br>- contact | - short title of project name<br>- Green Plan program, e.g. RES, RRC, 3W<br>- Green Plan sub-program, e.g. GP-RES-1, GP-RES-2, ...<br>- research subject (keywords), e.g. Organic Carbon<br>- indicated what kind type of project, e.g. res, activity<br>- unique identifier of each project (using agreement#)<br>- year the project started<br>- name of project leader/supervisor<br>- agency conducting the project<br>- farmer who cooperated in conducting the project<br>- project duration<br>- report number<br>- other related reference, if any<br>- citation<br>- data security level (low, medium and high)<br>- data access level (internal, external)<br>- contact person of the project or data manager |
| Sampling entity definition | Sampling factor table (.SFT) | - **proj-id**<br>- **start-yr**<br>- factor<br><br>- factor# | - unique identifier of each project (using agreement#)<br>- year the project started<br>- factors defining sampling conditions, e.g. primary location, secondary location, till system, crop, etc.<br>- numeric code for each factor |
| | Sampling factor level table (.FLT) | - **proj-id**<br>- **start-yr**<br>- factor#<br>- level<br>- level# | -unique identifier of each project (using agreement#)<br>- year the project started<br>- numeric code for each factor<br>- level of factor<br>- numeric code for each factor level |
| | Sampling entity identification table (.SIT) | - **proj-id**<br>- **start_yr**<br>- **se-id**<br><br>- factor1<br>- factor2<br>... | - unique identifier of each project (using agreement#)<br>- year the project started<br>- unique identifier of each sampling entity ( the combination of factor# and level#)<br>- place level # within columns of appropriate factor |

| | | | |
|---|---|---|---|
| Property observations/ measurements | Property table (.PPT) | - **proj-id**<br>- **start-yr**<br>- **se-id**<br>- horizon<br>- udepth<br>- ldepth<br>- portion<br>- repl#<br>- date<br>- time<br>- class<br>- subclass<br>- component<br>- kind<br>- unit<br>- value-qual<br>- value-quan | - unique identifier of each project (using agreement#)<br>- year the project started<br>- unique identifier of each sampling entity<br>- horizon of sampling (apply to soil)<br>- upper depth of the sampling horizon (apply to soil)<br>- lower depth of the sampling horizon (apply to soil)<br>- sampling portion of entity (apply to biomass, etc)<br>- numeric code for spatial replicate of the sampling<br>- sampling date (yy/mm/dd)<br>- sampling time (hh/mm)<br>- property class<br>- property subclass<br>- property components<br>- kind of measurements/observation<br>- property unit<br>- property value- qualitative (symbolic/character)<br>- property value- quantitative (numeric) |

Note: the basic key in each table are in bold text, for example, the basic key to the attribute records in the property table consists of a proj-id, start-yr, se-id

# Appendix 4. Procedures of describing sampling data and defining sampling entity identification

Agri-environmental research projects can take a wide variety of forms ranging from projects where several farms are involved to projects where treatment plots are contained within replicate blocks within one field. Both may result insamples which are taken and transported to the laboratory for further treatment and analysis. As described in Appendix 3, the general nature of the project is defined within the project table which has as its unique key a project identification number (proj-id) and a starting year for the project (start-yr). In theory, proj-id would be sufficient but the incorporation of start-yr facilitates the naming convention by requiring only that the proj-ids be unique within a year.

Within any project, data and attributes are collected for one or generally a number of entities which are to be sampled. The actual sampling entity can take a wide variety of forms such as fields or plots or benchmark sites within plots with very little consistency in what the sampling entity is from one project to another. Frequently, within a project attributes will be sampled at more than one level (e.g. it may be desirable to sample benchmark sites within a field and also to sample attributes of the entire field). Sampling entities represent the spatial unit within an agri-environmental research project for which data are collected. Sampling entities are defined by a series of factors which provide the basis for their selection. Location, in absolute and relative terms, to define the X and Y coordinates, and also the size of the sampling entity, represent factors common to the definition of most sampling entities. Other factors which characterize the sampling entity tend to be related to specific biophysical or anthropogenic conditions which in effect impose a treatment on the sampling entity. The objective of most projects is to determine how the treatment affects the attributes or observations.

Three sampling entity tables, consisting of a Sampling Factor Table (SFT), sampling Factor Level Table (FLT) and Sampling entity Identification Table (SIT) (Appendix 3) provide a unique definition of the sampling entities associated with a project. These tables are highly flexible in content and as such are unique to an individual project. Unique sampling entity identification numbers (SE-ID) are then assigned arbitrarily based on a matrix of factor-levels (see the table in the example box). This approach parallels the treatment definition procedure developed within the Soil Performance and Management file of CanSIS (MacDonald and Strzelczyk 1986).

Following is an example of defining a set of factors and an unique sampling entity identification:

Consider the Green Plan Contract by the Ecological Service for Planning, Ltd. (ESP) which sampled various farms. It included sampling of various parts of the Ontario soil benchmark site at Rockwood. The specification of factors could be done as follows:

Factor 1: Primary Location - location of the farm where the sampling entities are located.

Levels: there are 4 levels for this factor; actual coordinates are recorded in the attribute table
Level 1: Lobb Farm
Level 2: McRae Farm
Level 3: Davis Farm
Level 4: Nelson Farm

Factor 2: Slope Position

Levels: there are 4 levels in this factor; relative locations are recorded in the attribute table
Level 1: Upper slope
Level 2: Mid slope

Level 3: Lower slope
Level 4: Depression

Factor 3: Tillage practice

Levels: there are 4 levels in this factor; additional details are recorded as land use in the attribute table
Level 1: Conventional tillage
Level 2: Conservation tillage
Level 3: No-till
Level 4: Native vegetation

Factor 4: Crop

Levels: there are 5 levels for this factor; specifics of crop varieties, seeding dates, etc. are recorded in the attribute table
Level 1: Winter Wheat
Level 2: Soybean
Level 3: Corn
Level 4: Barley
Level 5: Woodland

The unique identifier of each sampling entity could then be defined using the following matrix (partial list):

| Sample Entity ID | Factor 1 | Factor 2 | Factor 3 | Factor 4 | Sample Entity ID | Factor 1 | Factor 2 | Factor 3 | Factor 4 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 3 | 1 | 1 | 1 | 12 | 4 | 2 | 4 | 5 |
| 2 | 3 | 1 | 2 | 1 | 13 | 4 | 3 | 4 | 5 |
| 3 | 3 | 1 | 2 | 2 | 14 | 4 | 1 | 1 | 1 |
| 4 | 3 | 1 | 2 | 3 | 15 | 4 | 1 | 1 | 2 |
| 5 | 3 | 2 | 2 | 1 | 16 | 4 | 1 | 1 | 3 |
| 6 | 3 | 2 | 2 | 2 | 17 | 4 | 2 | 1 | 1 |
| 7 | 3 | 2 | 2 | 3 | 18 | 4 | 2 | 1 | 2 |
| 8 | 3 | 3 | 2 | 1 | 19 | 4 | 2 | 1 | 3 |
| 9 | 3 | 3 | 2 | 2 | 20 | 4 | 3 | 1 | 1 |
| 10 | 3 | 3 | 2 | 3 | 21 | 4 | 3 | 1 | 2 |
| 11 | 4 | 1 | 4 | 5 | 22 | 4 | 3 | 1 | 3 |

# Appendix 5. Procedures to conduct a user specified data query (example):

If a researcher is interested in soil organic carbon data, the query steps would be as follows:

**Step 1:** Look up the property classification and code tables which will be part of the metadata (see Appendix 6, the up-to-date version can be generated from the current version of database) to determine the keywords (descriptive terms) to be used for the query, for example:

| **Property class** | **Property subclass** | **Property component** (selected list) | **Kind of measurements /observations** (selected list) | |
|---|---|---|---|---|
| | | | **Abbreviation** | **Full name/explanation** |
| Soil | Soil Chem (soil-chem) | Soil organic carbon content (soil-orgcarb) | - orgmat<br>- totcarb<br>- calccarb<br>- orgcarb<br>- solcarb<br>- micro-bio-c<br>- mac-oc-sndfr<br>- mac-oc-whs<br>- mac-oc-spm<br>... | - soil organic matter<br>- total carbon<br>- calcium carbonate equivalent<br>- organic carbon<br>- soluble carbon<br>- microbial biomass carbon<br>- macro-organic carbon sand fraction<br>- macro-organic carbon in whole soil<br>- macro-organic carbon specific mass |

**Step 2:** Select the subset of data from the property table which can be made at a general 'property component' level, e.g.

*>resselect component = 'soil-orgcarb'*
*2657 record (s) selected* /* which are sub-set of property table records containing measurements of one table or another related to organic carbon content

**Step 3:** Narrow down the query, select a specific kind of measurement, e.g. organic carbon:
*>reselect kind = 'orgcarb'*
*332 record(s) selected*

**Step 4:** List the data fields: 'horizon', 'udepth', 'ldepth', 'replicate#' and 'date' within the same table, to find out the sampling depth, replicate number and date and list the data value, and unit, of selected records of 'orgcarb' measurement.

Data manipulation can be conducted at this stage, for example, to calculate commonly used statistic parameters - mean, max, min, sd within the selected records. Before doing the calculation, make sure the data are compatible, e.g. to verify the unit, used for the measurement, by different projects are the same, etc.

**Step 5:** Use the project ID (PROJ-ID) to query the project table (PJT) to find out the projects which conducted organic carbon sampling and the detailed information of the research projects.

**Step 6:** Use the PROJ-ID + the sampling entity ID (SE-ID) to query the location property withing the property data table (ppt) or the sampling factor table (SFT) and the sampling factor level table (FLT) to determine the locations of the sampling.

**Step 7:** Using the PROJ-ID + the SE-ID to query the sampling factor table (SFT) and sampling factor level table (FLT) to show the treatment conditions which affect the level of organic carbon.

# Appendix 6: Data catalog and metadata document of all data sets (as separate document)

# Appendix 7. Verified data in Arc/Info export file, dBase etc. format (on separate CD)

# Appendix 8. Research reports corresponding to each data set (on separate CD or on Green Plan web site

# http://res.agr.ca/ lond/gp/gphompag.html