Institute of Terrestrial Ecology

Centre for Ecology & Hydrology

Natural Environment Research Council

ITE has six Research Stations throughout Britain, which allows the efficient use of resources for regional studies and provides an understanding of local ecological and land use characteristics. The Institute's administrative headquarters is at Monks Wood.

ITE sites

Monks Wood
(Admin HQ)
Abbots Ripton
HUNTINGDON PE17 2LS
Telephone 01487 773381-8
Fax 01487 773467
Email MONKSWOOD@ITE.AC.UK

Merlewood Research Station
GRANGE-OVER-SANDS
Cumbria LA11 6JU
Telephone 015395 32264
Fax 015395 34705
Email MERLEWOOD@ITE.AC.UK

Edinburgh Research Station
Bush Estate
PENICUIK
Midlothian EH26 0QB
Telephone 0131 445 4343
Fax 0131 445 3943
Email BUSH@ITE.AC.UK

Furzebrook Research Station
WAREHAM
Dorset BH20 5AS
Telephone 01929 551518-9, 551491
Fax 01929 551087
Email FURZEBROOK@ITE.AC.UK

Banchory Research Station
Hill of Brathens
Glassel, BANCHORY
Kincardineshire AB31 4BY
Telephone 01330 823434
Fax 01330 823303
Email BANCHORY@ITE.AC.UK

Bangor Research Unit
University of Wales, Bangor
Deiniol Road
BANGOR, Gwynedd LL57 2UP
Telephone 01248 370045
Fax 01248 355365
Email BANGOR@ITE.AC.UK

Details about the Institute are available on the Internet via the World Wide Web (http:/www.nmw.ac.uk/ite)

COUNTRYSIDE SURVEY2000

Module 13
Scientific Support and Information Management

Countryside Survey 2000 Integrated Data System (CIDS)

Data Model

Interim Report 1
Prepared by:

Dr. Timothy Moffat
NERC Environmental Information Centre
Institute of Terrestrial Ecology, Monks Wood

Mr. John Watkins
Institute of Terrestrial Ecology, Merlewood

Mrs. Kay Symes
Institute of Freshwater Ecology, Wareham

October 1998

Contents

List of Figures

1. Introduction

A key objective of Module 13 of the Countryside Survey 2000 (CS2000) is to design, implement and manage a single database located at ITE Merlewood that will integrate and link the data collected in each of the CS2000 Modules, and earlier surveys (e.g. Countryside Survey 1990 and ITE's National Survey 1978). The report describes progress to-date in developing and implementing a data model of the Countryside Survey 2000 Integrated Data System (CIDS).

Some of the diagrammatic techniques of the Structured Systems Analysis and Design Method (SSADM) have been used in the analysis and design of CIDS:

- Data Flow Modelling - to understand the flows of data around the system (Chapter 2).
- Logical Data Modelling - to understand the data that underlies the system and its interrelationships (Chapter 3).
- Physical Design - to convert the logical data models into a design and structure suitable for the physical environment (Chapter 4).

Meta-data describing the data collected by all the surveys, data products and the data stored within the integrated database are integral parts of the CIDS. Emphasis has been placed on providing a meta-data system that allows users to explore the structure of the CIDS data sets. As the meta-data is physically linked to the raw data, users will be able to follow this meta-data exploration into the data sets themselves without needing intimate knowledge of the underlying physical data structures. The meta-data will also provide the basis for a WWW-searchable catalogue of data and data products.

Work will continue in developing and implementing the data model until all the data to be integrated into the CIDS is fully mapped in a physical model. A complete data model will be published at a later date.

2. Data Flow Model

Data flow modelling was used in the analysis and design of the Countryside Survey 2000 Integrated Data System (CIDS) in order to:

- Understand clearly the flows of data around the system.
- Define the processes that transform or manipulate the data.
- Identify the sources and recipients of data outside the system.
- Show where data are held within the system.
- Form the basis of function definition (e.g. data input, validation, enquiries and access).

2.1 CIDS Top Level Context

The context diagram presents the entire system as a single process (shown as a square) with all flows to and from external entities. External entities are sources or recipients of data immediately outside the system boundary.

Figure 1 shows the flows to and from four external entities (shown as ovals and labelled a to d) associated with CIDS and Module 13. These are the supply of data from various sources (e.g. data from the CS2000 modules and previous surveys) and the supply of data in various forms to external users (e.g. as Level 1 and 2 data outputs). Other external data sources will provide supporting data such as details of ownership and licensing arrangements.

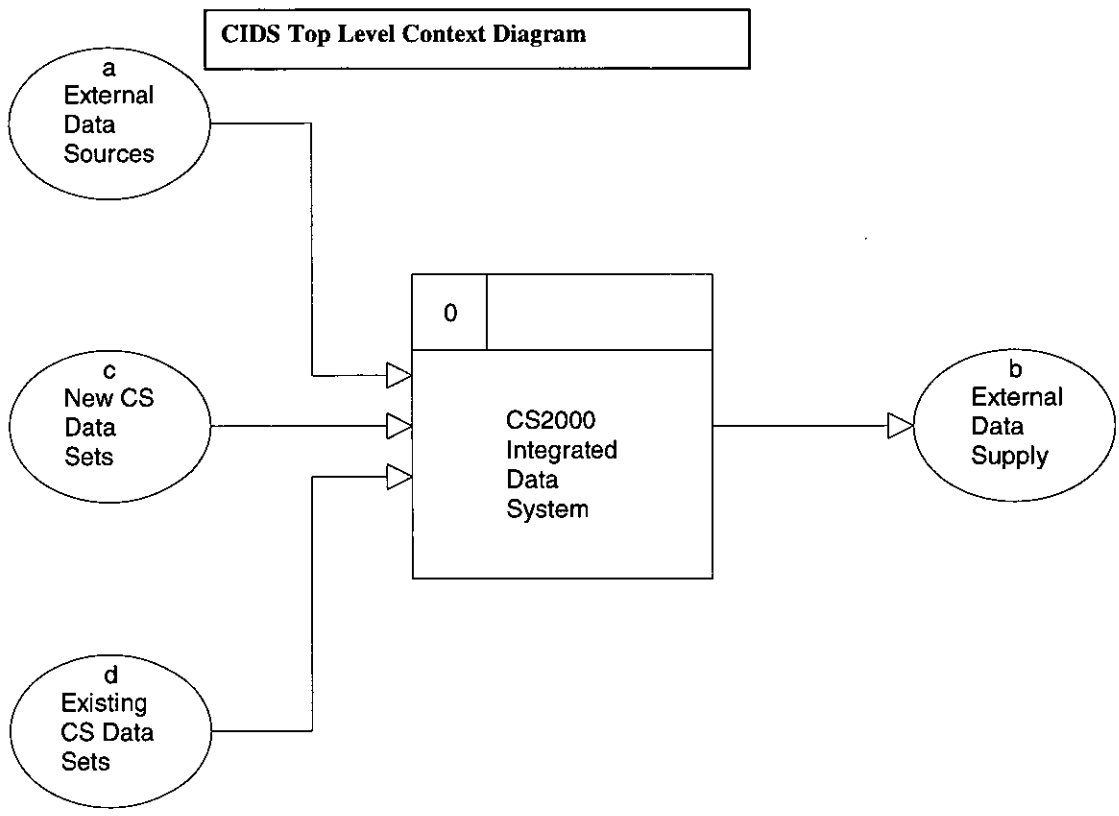

Figure 1: CIDS Context Diagram

2.2 CIDS Main Processes

Figure 2 shows the main processes (shown as squares) and data flows (shown as arrows) associated with CIDS. Processes represent the transformation or manipulation of data, data flows are "channels" down which predefined sets of data may flow, and the "Integrated Database" is a data store (labelled D1) which represents data at rest.

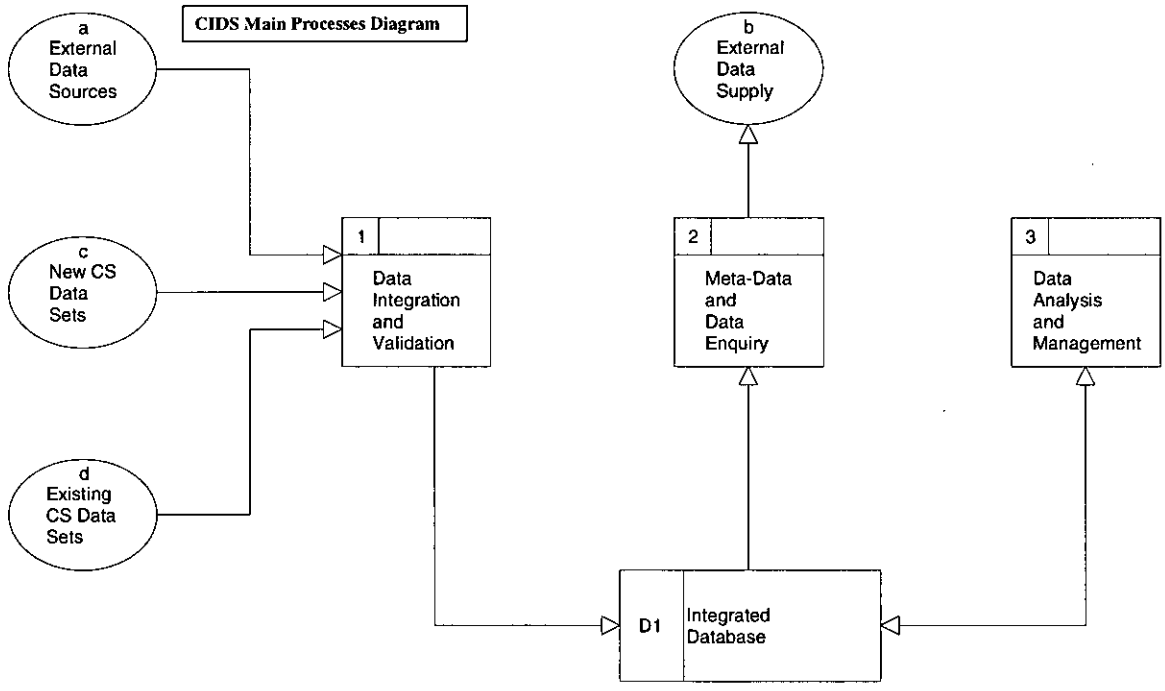

Figure 2: CIDS Main Processes Diagram

For CIDS, there are three main processes:

- Process 1 (Data Integration and Validation) takes data from all outside entities and integrates them into the main database (D1). Depending on where these data have come from, some of the sub-processes within Process 1 may take place in other CS2000 Modules. For example, survey data from Module 1 and 2 will be entered and validated as part of those modules. The processing of data from previous surveys will be carried wholly in Module 13. Integration of data from different sources will create an additional validation requirement.

- Process 2 (Meta-Data and Data Enquiry) handles all types of external data enquiries. This process will implement the three level access model required for Module 13. The data requests are broken down into meta-data and data product enquiries (i.e. Level 1 and 2 enquiries that can be serviced via the Internet) and exploratory enquiries of raw data (i.e. Level 3 enquiries that will be restricted)

- Process 3 (Data Analysis and Management) covers the internal processes necessary for managing the database and the analysis tools that will be used for creating specific Level 2 data products (e.g. CIS data files).

2.3 CIDS Processes Breakdown

Figure 3 shows a breakdown of the CIDS processes and provides further details of the data flows. An "Intermediate Data Store" (D2) will be used for validation purposes during the entry of new survey data. This will give greater control over the validation processes before the data are entered into the "Integrated Database" (D1), from which data analyses will take place. Data may go through several validation sets in the intermediate store before being past as "fit-for-analysis".

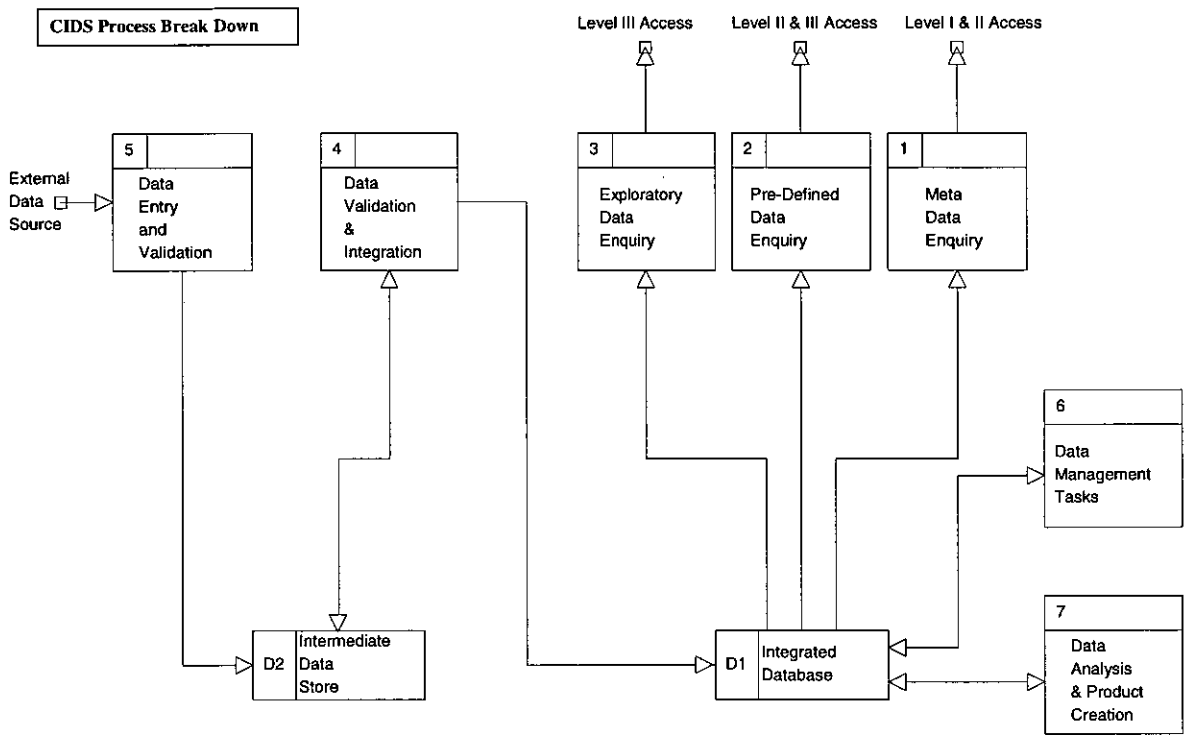

Figure 3: CIDS Processes Breakdown

3 Logical Data Model

The data were modelled in the analysis and design of the Countryside Survey 2000 Integrated Data System (CIS) in order to:

- Understand clearly the information (data) that will underlie the system and its interrelationships.
- Build the requirements for data onto this basic model.
- Incorporate the data items and access paths required by the system's processing.
- Act as a basis for the physical database design of the implemented system.

3.1 CIDS Integrated Database -Top Level Logical Data Structure

Figure 4 shows a breakdown of the Integrated Database (D1) into four entities which represent the main data modules in the CIDS. These modules (entities) are logical groupings of data, and the relationships between them are drawn as lines with "crow's feet" to represent the degree of relationship (e.g. one-to-many or many-to-many).

Module 1 (Ownership and Access Information), Module 2 (Data Set Structure Information) and Module 3 (Subject Search Information) form the CIDS Meta-Data System that describes the data sets. Module 4 (Countryside Survey Data Sets) contains the raw data and supporting data from the Countryside Survey 2000 and previous surveys.

The data structure breakdowns of each of the D1 Modules is given in the following sections.

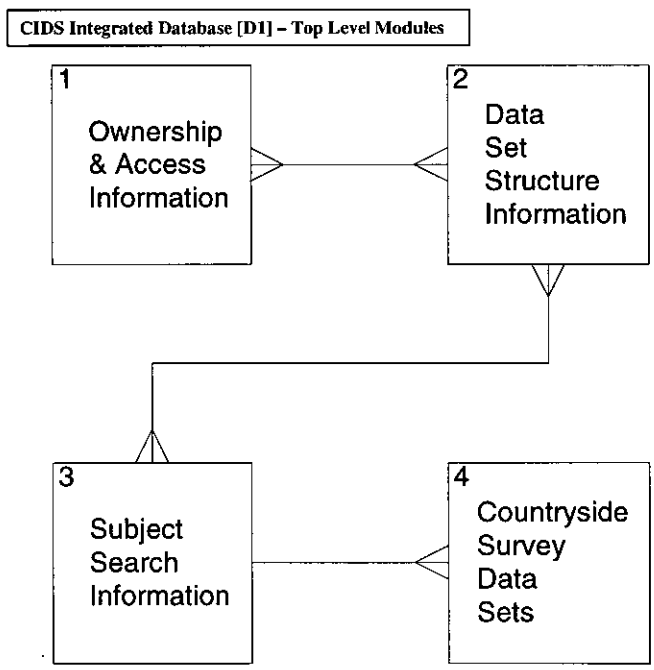

Figure 4: Top Level Modules of the CIDS Integrated Database

3.2 CIDS Meta-Data System – Logical Data Structures

The Meta-Data System forms a key component of the CIDS because it will provide users with comprehensive and invaluable information about all the survey data, any supporting data and the structure and contents of the CIDS Integrated Database. In particular, meta-data will be generated that describe individual data groups (termed "Data Sets") and how they are stored (termed "Storage Sets").

Data Sets represent groupings of data identified by their respective surveys or sources. Information on each Data Set can be split into a number of Data Items, and individual Data Items can be broken down into further items to form a *Data Hierarchy*. For example, the Countryside Survey 2000 represents a Data Set, and information about the individual surveys associated with the CS2000 form a number of Data Items (e.g. the Field Survey, the Freshwater Survey, the Satellite Survey and the Soil Survey etc.). Information about the individual surveys can also be broken down into a number of Data Items (e.g. information on the Field Survey can be split into Random Plots, Targeted Plots, Boundary Plots, Hedgerow Plots etc.).

Storage Sets represent groupings of data identified by their storage format and/or medium (e.g. Oracle Database, Arc/Info GIS Database, Field Assessment Books, Freshwater Samples, Soil Samples, CIS data files etc.) Information on each Storage Set can be split into a number of *Storage Items*, and individual Storage Items can be broken down further to form a *Storage Hierarchy*. For example, the CIDS Oracle Database represents a Storage Set, and information about the different entities that comprise the database form a number of Storage Items (e.g. Ownership and Access Information, Data Set Structure Information, Subject Search Information and the Countryside Survey Data Sets). Information about the structure and contents of these entities can be further broken down into a number of Storage Items (e.g. information on the Countryside Survey Data Sets can be split into Survey Square Details, Vegetation Survey Data, Freshwater Survey Data etc.). The Storage Hierarchy will provide very detailed information; for example, the definition of individual Oracle Tables.

The Meta-Data System will comprise of three modules and the data structure of each of these is presented below.

Ownership and Access Information

Figure 5 shows a breakdown of the Ownership and Access Information Module into a number of entities. These entities include the basic information on the ownership, access conditions and storage location of data.

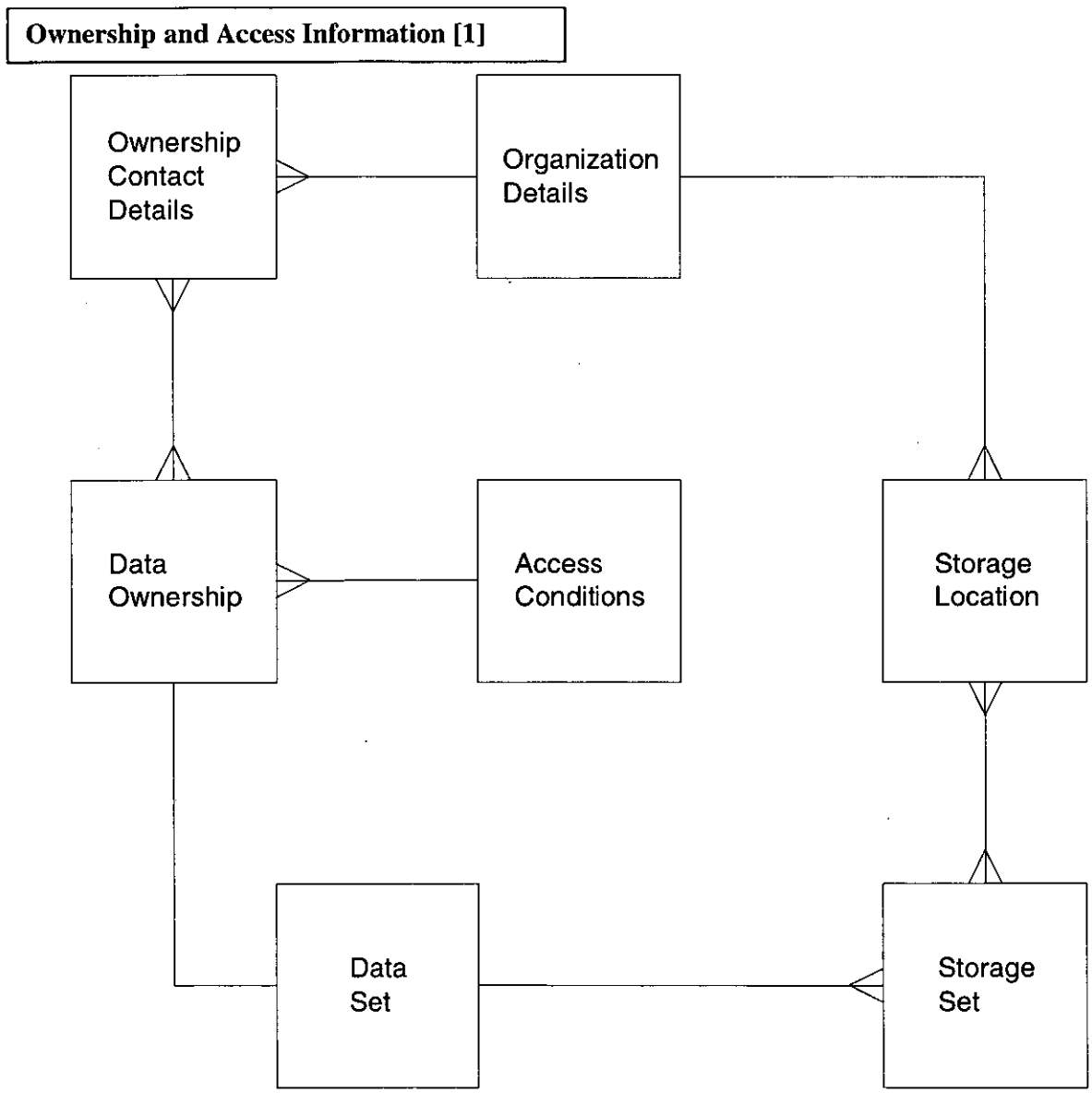

Figure 5: CIDS Integrated Database - Ownership & Access Information Module

Data Set Structure Information

Figure 6 shows a breakdown of the Data Set Structure Information Module into a number of entities. The entities, including ones for Data Set and Storage Set (as defined above), describe the basic structure, storage and links of all the data collected by the CS2000 and earlier national surveys. Information on the quality (e.g. accuracy, resolution, integrity etc.) of individual data and storage items will be stored within this module.

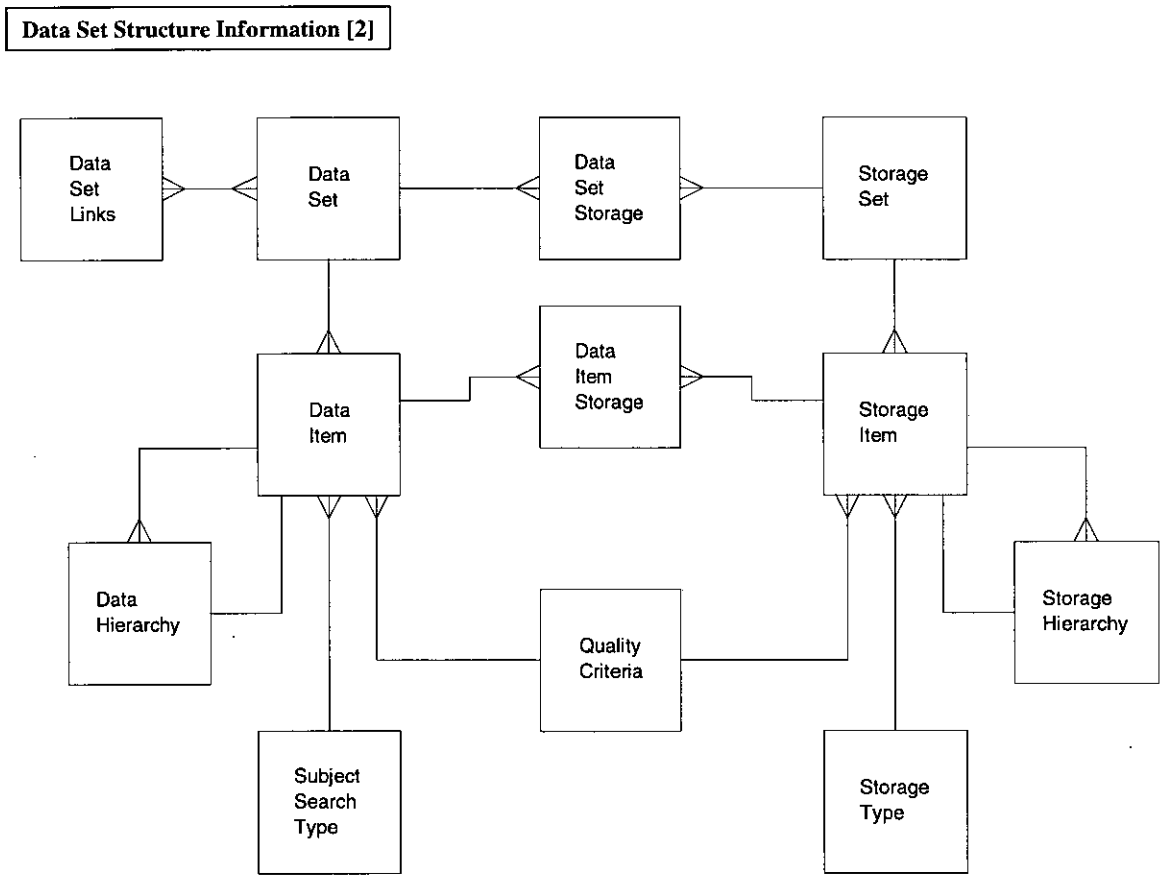

Figure 6: CIDS Integrated Database - Data Set Structure Information

Subject Search Information

Figure 7 shows a breakdown of the Subject Search Information Module into a number of entities. The entities provide a comprehensive glossary of all the various codes, definitions and descriptions used in the Countryside and National Surveys.

This module will provide a direct link between the meta-data and the actual data, and will allow users to search for data by subject and data type.

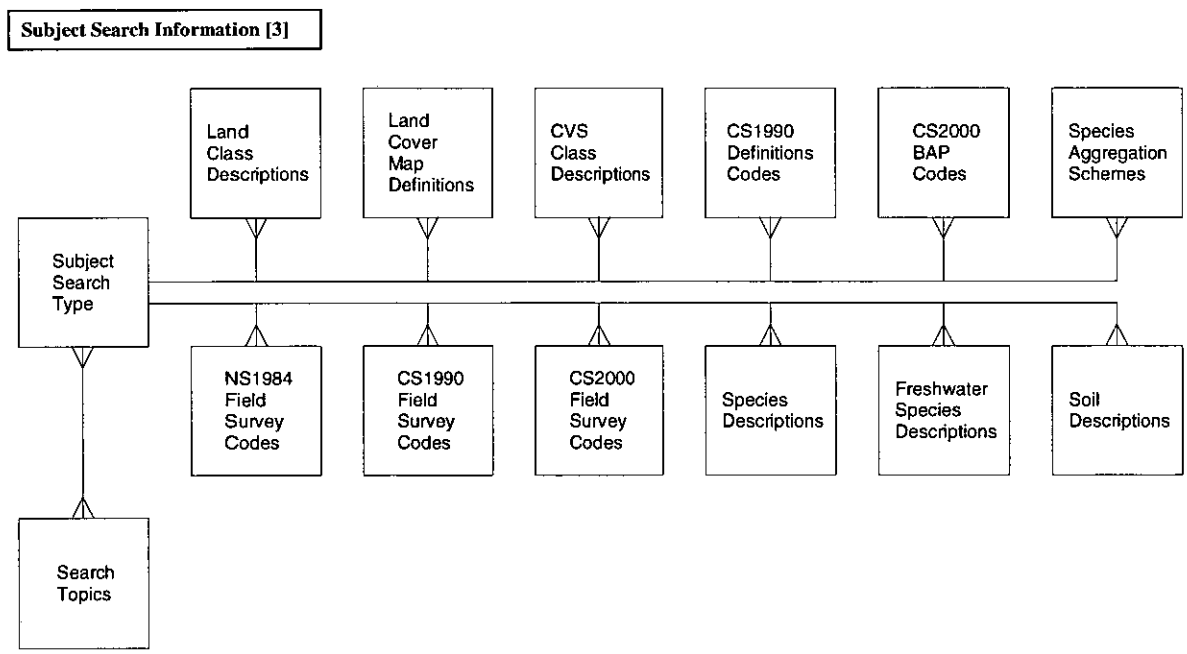

Figure 7: CIDS Integrated Database - Subject Search Information Module

3.3 CIDS Integrated Database – Countryside Survey Data Sets

The fourth module of the CIDS Integrated Database is the Countryside Survey Data Sets. This module will contain the raw data from the Countryside Survey 2000 and previous surveys. Figure 8 shows a breakdown of this module into a number of entities, including the main groupings of data: Satellite Land Cover Data, Field Survey Land Cover Data, Vegetation Survey Data, Freshwater Survey Data, Soil Survey Data and Additional Survey Data.

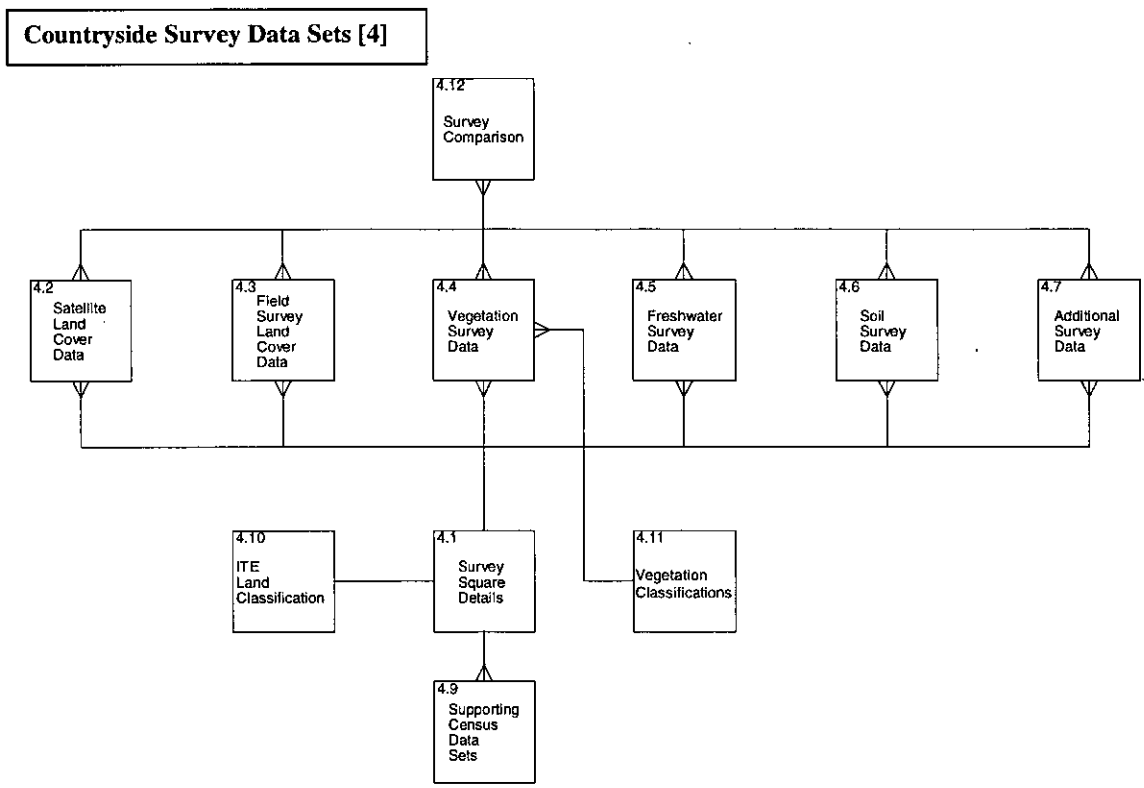

Figure 8: CIDS Integrated Database - Countryside Survey Data Sets Module

The individual entities shown in Figure 8 can be further broken down. Some examples are given in Figures 9 to 12.

Satellite Land Cover Data

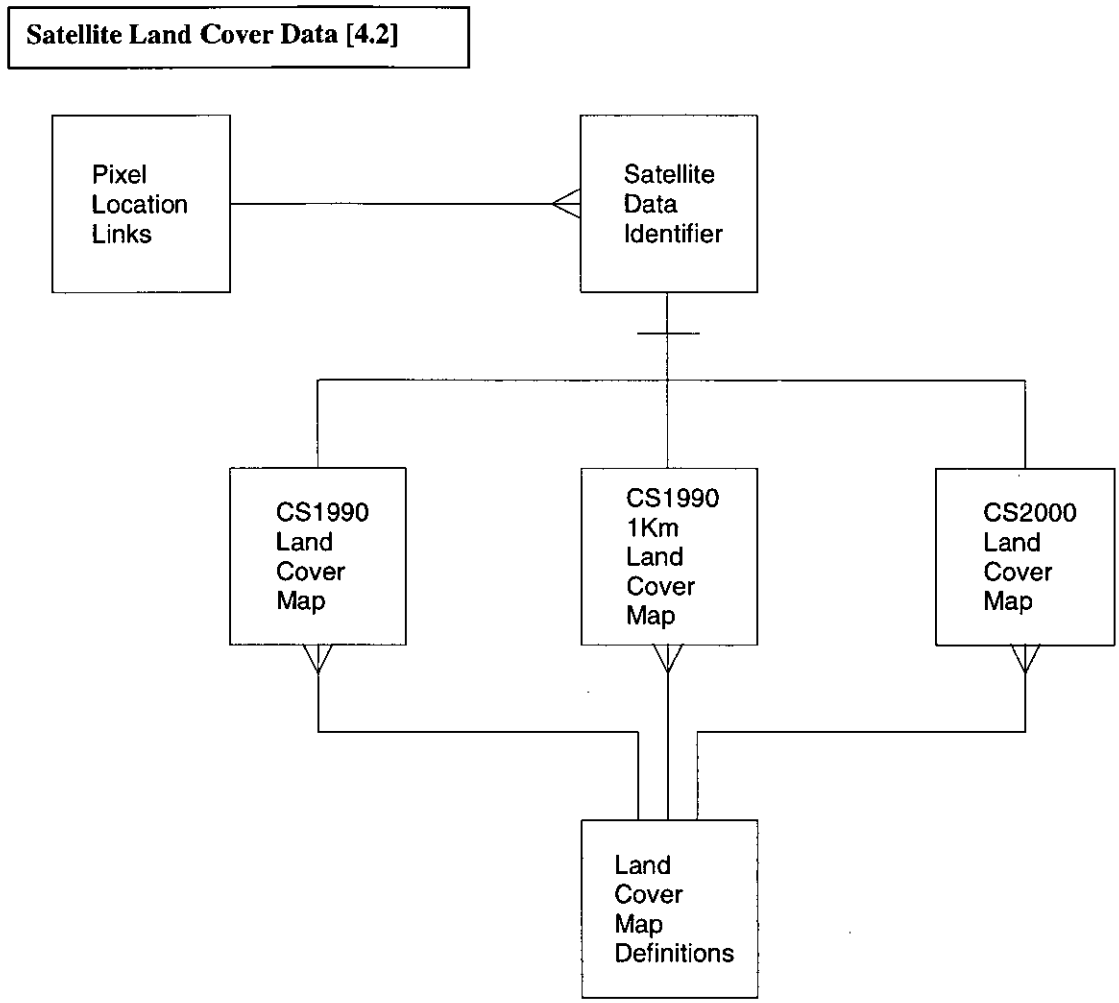

Figure 9: CIDS Integrated Database - Satellite Land Cover Data

Vegetation Plot Data

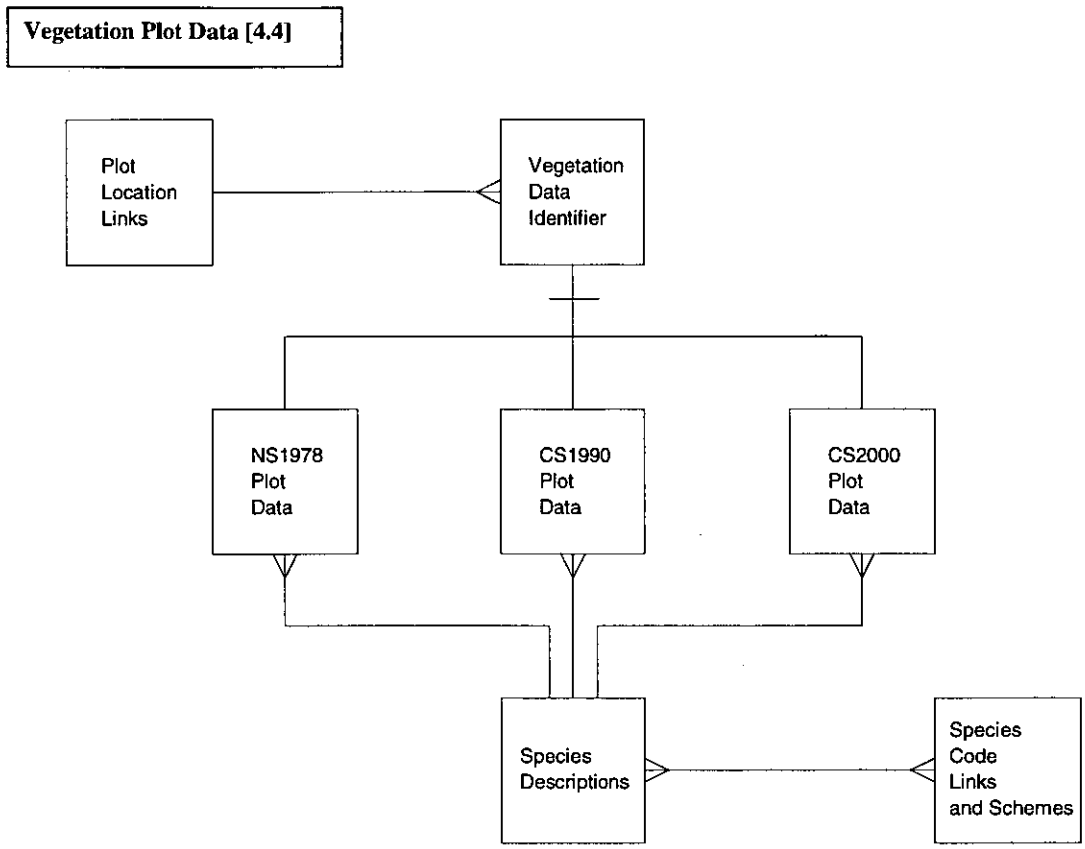

Figure 10: CIDS Integrated Database – Vegetation Plot Data

Survey Comparison Modules

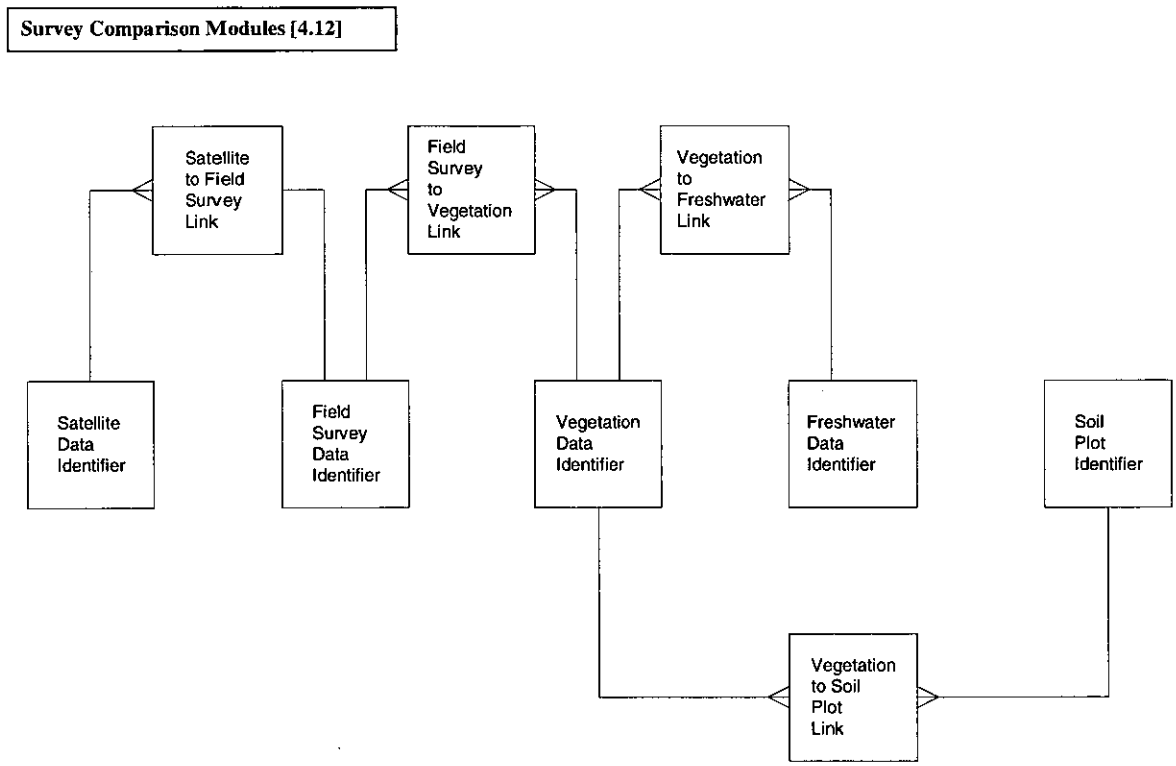

Figure 11: CIDS Integrated Database – Survey Comparison Modules

Field Survey Land Cover Data

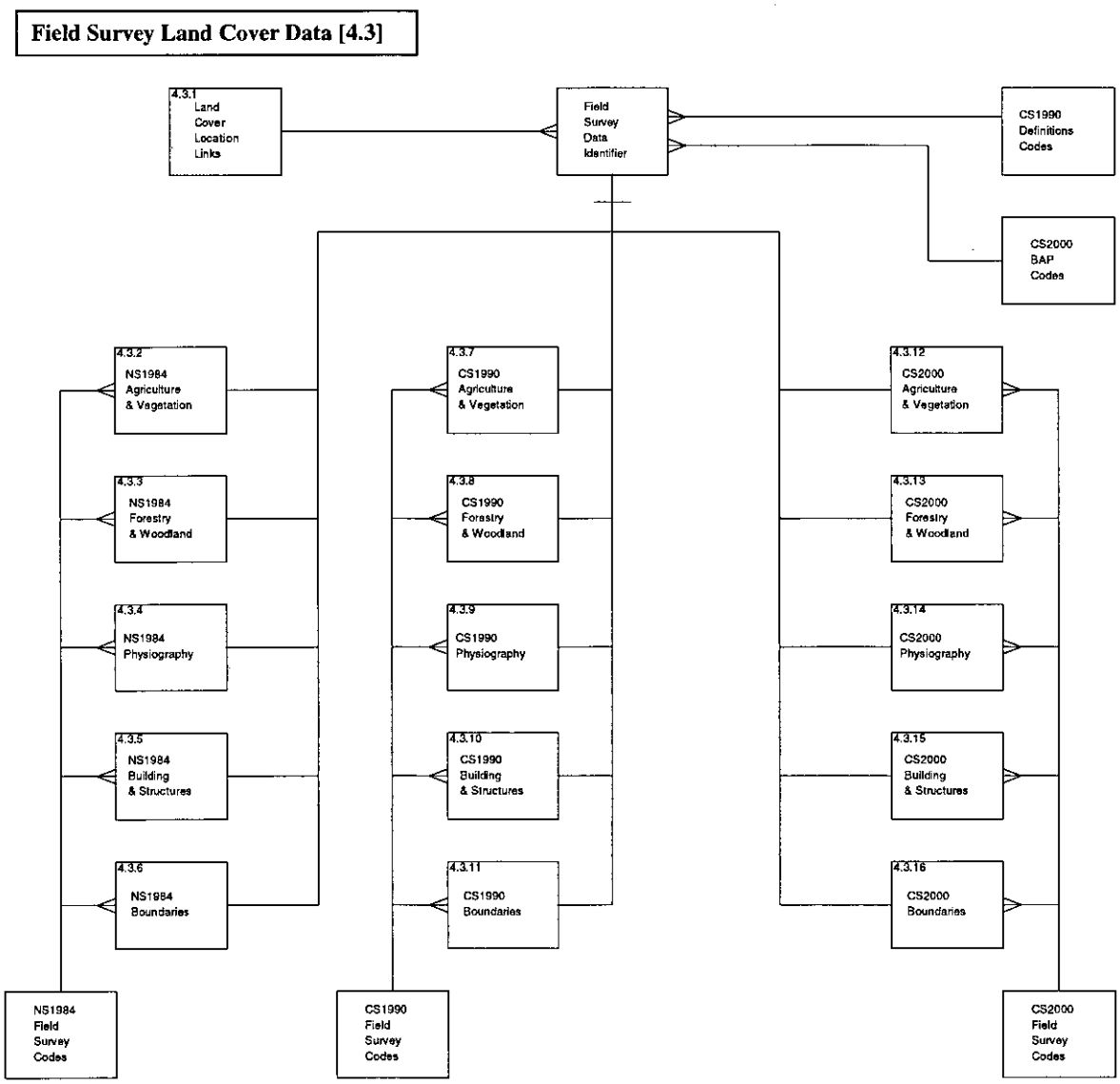

Figure 12: CIDS Integrated Database – Field Survey Land Cover Data

4 Physical Data Structures

The purpose of the physical design of the Countryside Survey 2000 Integrated Data System (CIDS) is to convert the logical data model into a design suitable for the physical environment. The physical environment will be provided by the ORACLE Relational Database Management System which is the NERC/ITE corporate standard.

Work is in progress to directly map the entities specified in the logical data model into ORACLE tables and files. This work will be iterative in that the complete physical data design will not be implemented until the attributes of all the data from the other Modules are known. However, the intention is to keep the overall data structures as close to the logical model as possible to support easy maintenance, flexibility and comprehension of the design. Some optimisation of the design will be necessary to ensure that CIDS performs as efficiently as possible.

Some examples of the physical data structures are presented in the following sections.

4.1 CIDS Intermediate Data Store (D2)

As described in Section 2.3, the Intermediate Data Store will be used for validating the new survey data (from CS2000 Module 1). The structure of part of the Data Store for input of vegetation plot data is shown in Figure 13.

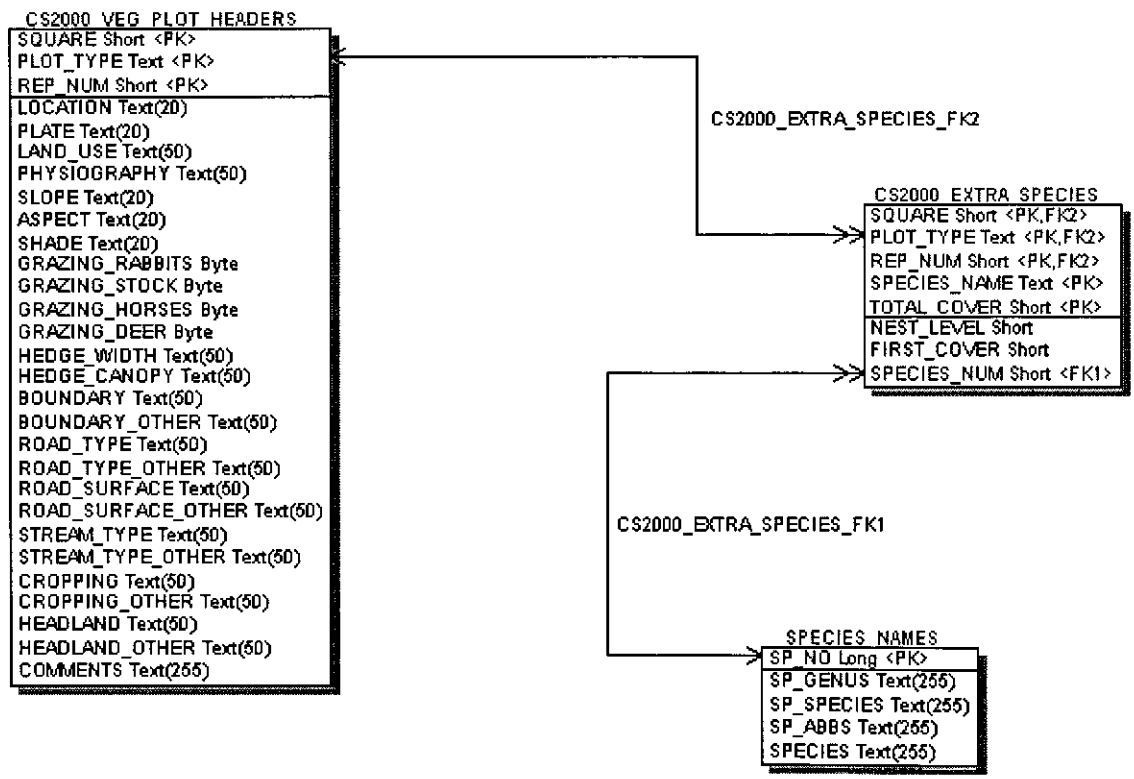

Figure 13: Intermediate Data Store (D2)

4.2 CIDS Integrated Database (D1) – Meta-Data System

As described in Section 4.2, the CIDS Meta-Data System comprises three modules that will provide comprehensive information and documentation of all the data collected by the Countryside and National Surveys.

The direct mapping of the logical data structure of "Data Set Structure Information" Module (Figure 6) into a physical data structure is given in Figure 14 below.

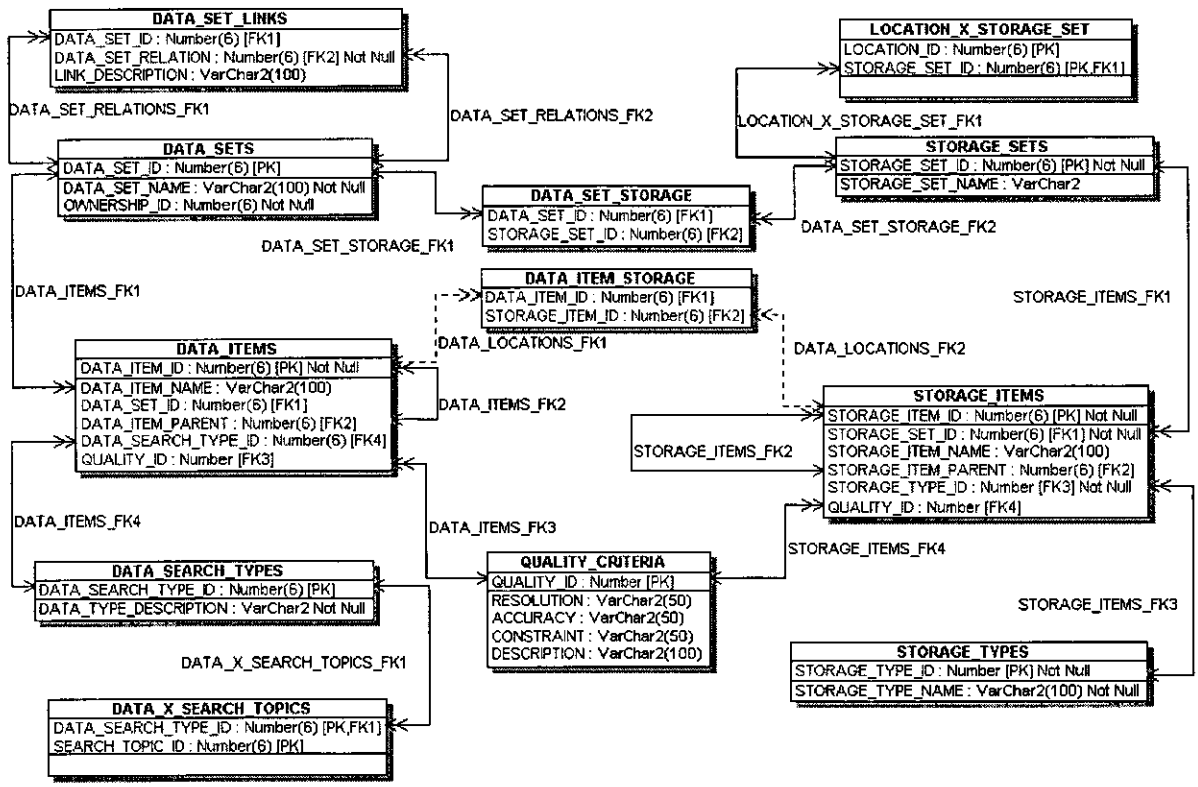

Figure 14: Data Set Structure Information

4.3 CIDS Integrated Database (D1) – Countryside Survey Data Sets

As described in Section 3.3, the raw data from the Countryside Survey 2000 and previous surveys will be stored within the "Countryside Survey Data Sets" Module.

The logical data model for this Module (Figure 8) shows the main data entities. For some of these, such as the Field Survey Land Cover Data, it is necessary to group the data by survey (e.g. CS2000, CS1990 and NS1978) because of changes in the coding systems adopted by each survey. In terms of data structures and tables, this will result in an extremely complex physical design. However, because this Module will be linked directly to the Meta-Data System, it will be possible for (authorised) users to navigate and query the raw data from the meta-data without having to comprehend the underlying data structures of CIDS.

4.4 CIDS Integrated Database (D1) – Field Survey Land Cover Data

Some examples of the mapping of the logical data structure of the Field Survey Land Cover Data entity (Figure 12) into physical data structures are given in Figures 15 to 17.

Field Survey Data Identifier

The Field Survey Data Identifier serves as an index to the data items stored within the Module and provides the links to the Meta-Data System. Figure 15 shows the physical data structure of this entity for CS1990. Similar structures will be constructed for the other surveys.

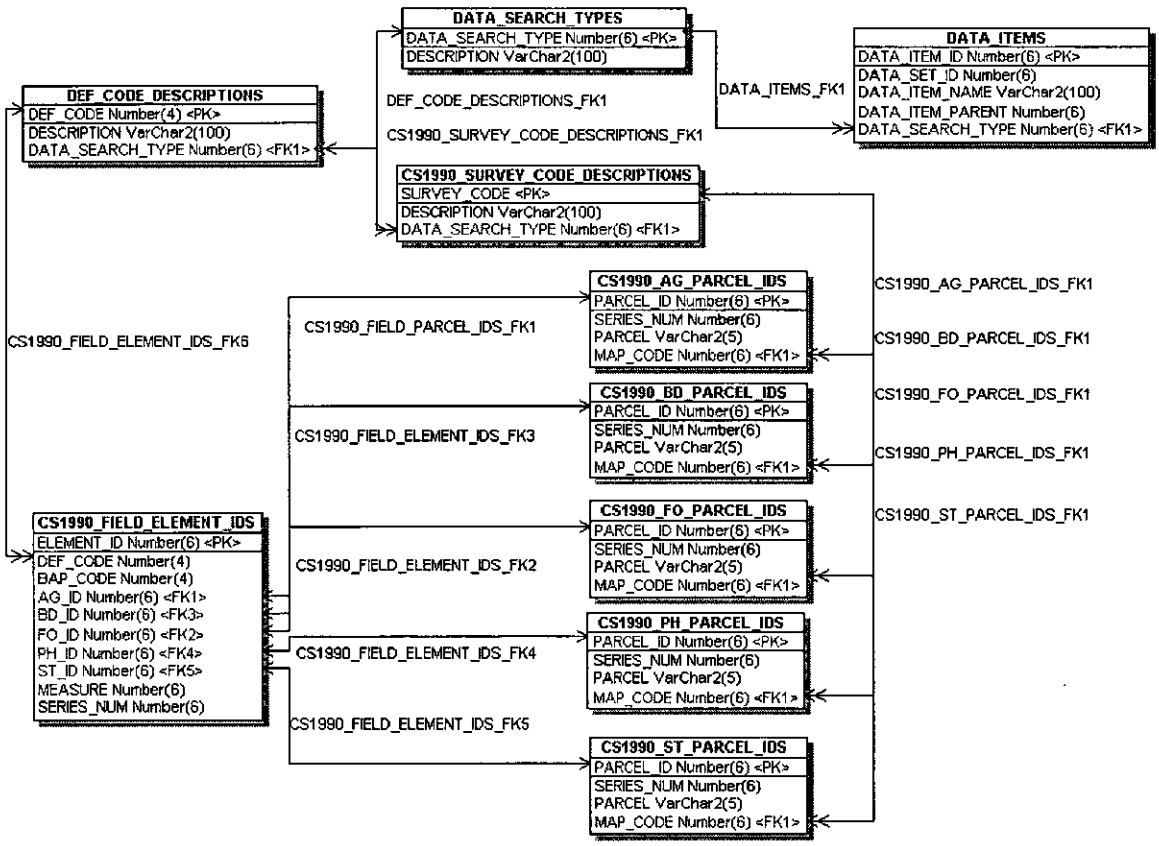

Figure 15: Field Survey Data Identifier

The identifiers (IDS) relate to particular themes: Agriculture & Vegetation (AG), Forestry & Woodland (FO), Physiography (PH), Building & Structures (ST), and Boundaries (BO). The "Data_Search_Types" Table provides the link to the Meta-Data System (see Figure 14).

Agriculture and Vegetation Data Sets

The breakdown of the physical data structures required for storing the agriculture and vegetation data (e.g. cover, use and species) for CS1990 are shown in Figure 16. Similar structures will be constructed for the other surveys.

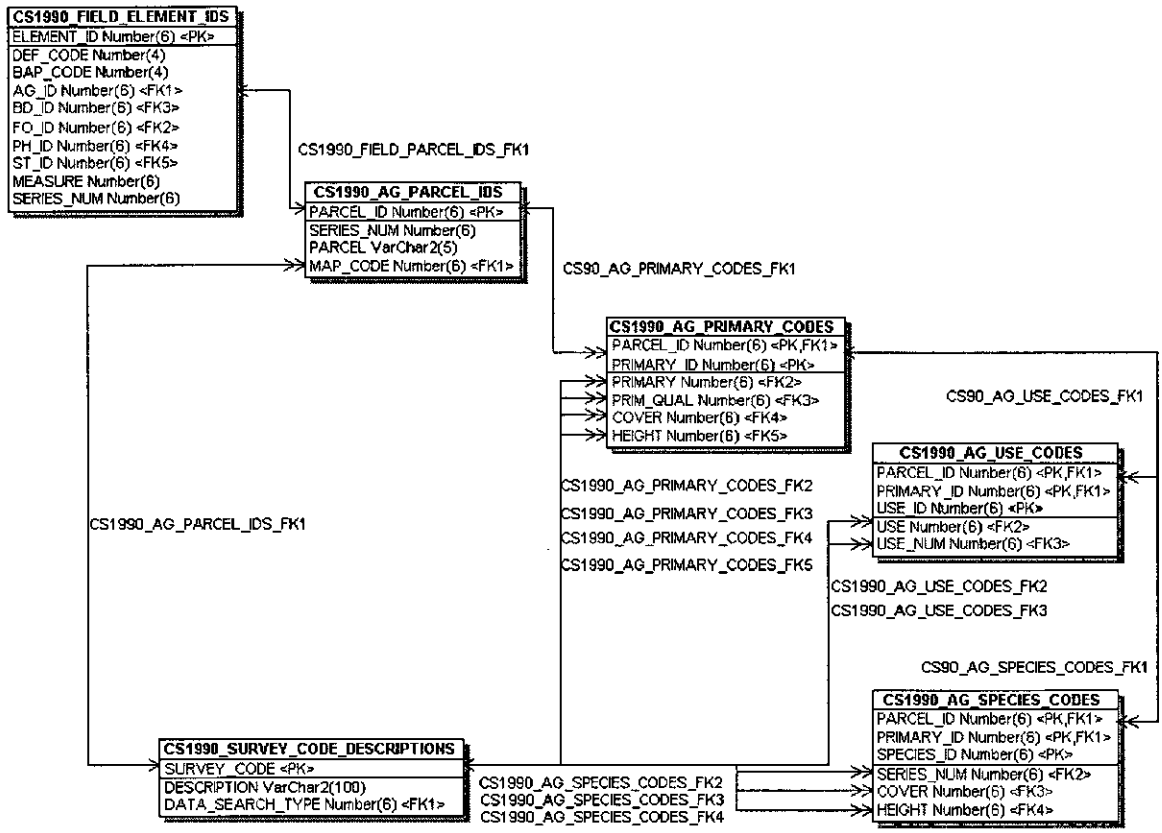

Figure 16: Agriculture and Vegetation Data Sets

Forestry and Woodland Data Sets

The breakdown of the physical data structures required for storing the forestry and woodland data (e.g. cover, use, features and species) for CS2000 are shown in Figure 17. Similar structures will be constructed for the other surveys.

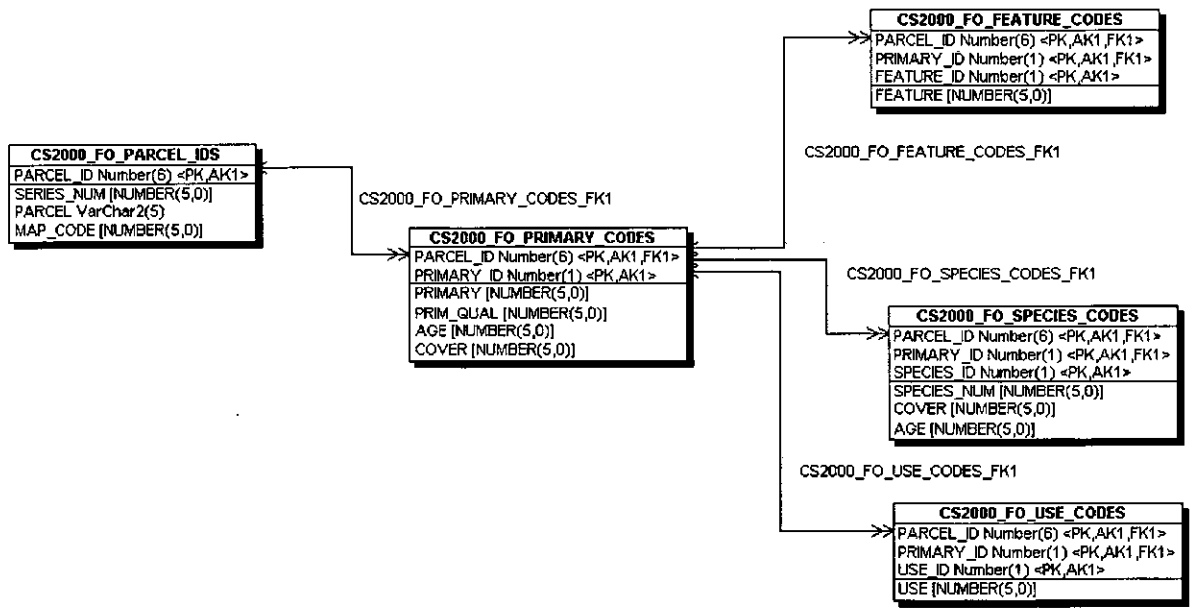

Figure 17: Forestry and Woodland Data Sets

4.5 CIDS Integrated Database (D1) – Freshwater Survey Data

The Freshwater Survey Data will be provided by the Institute of Freshwater Ecology (IFE), and will include aquatic macro-invertebrate data from CS1990 and CS2000, and River Habitat Survey data for CS2000. IFE will maintain the master copies of these data within their own database management systems; copies of the data will also be incorporated within CIDS.

For example, the Invertebrate Survey Data component of the CID Integrated Database will be an adaptation of the IFE's National Invertebrate Database System. This system was designed to store, retrieve and manipulate biological, environmental and chemical data collected for a variety of projects from sampling sites on rivers throughout the UK. The database was designed around the River Communities 'RIVPACS-style' invertebrate sampling methodology and so the physical data structures reflect the types of data usually collected (Figure 18).

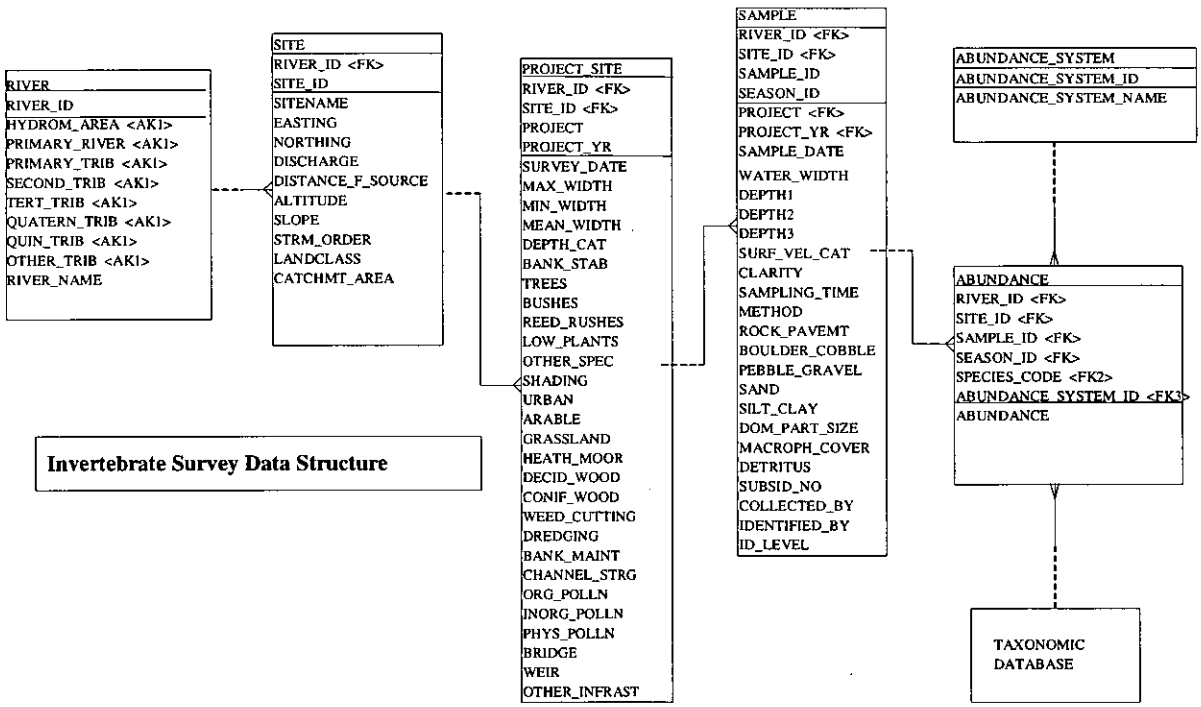

Figure 18: Invertebrate Survey Data

The principal entities are rivers, sites, project site visits (termed project_sites), samples and invertebrate taxa. The attribute data consist of geographical and physical attributes of sites as well as biological and chemical attributes of samples collected at each site. The physical site variables fall into two categories: invariant and variant data. Invariant data are collected once only during a project site visit in any year and are attributes of the entity project_site. Variant data are measured during each sampling event for any one project and so are attributes of the entity sample.

5 Data Management Tools

As part of the design and implementation of the Countryside Survey 2000 Integrated Data System (CIDS), some data management tools will be developed to capture and document the data of previous surveys and data produced by the CS2000 Modules. Two examples of tools in development are given below

5.1 Species Data Entry Form

Figure 19 shows the Species Data Entry Form which will be used to input new survey data into the Intermediate Data Store (D2) for vegetation plot data. In combination, they will allow data to be entered quickly by staff who have few IT skills. Staff with no botanical skills can use the form for entering validated, 'clean' data or by botanically skilled staff for validating data as it is entered.

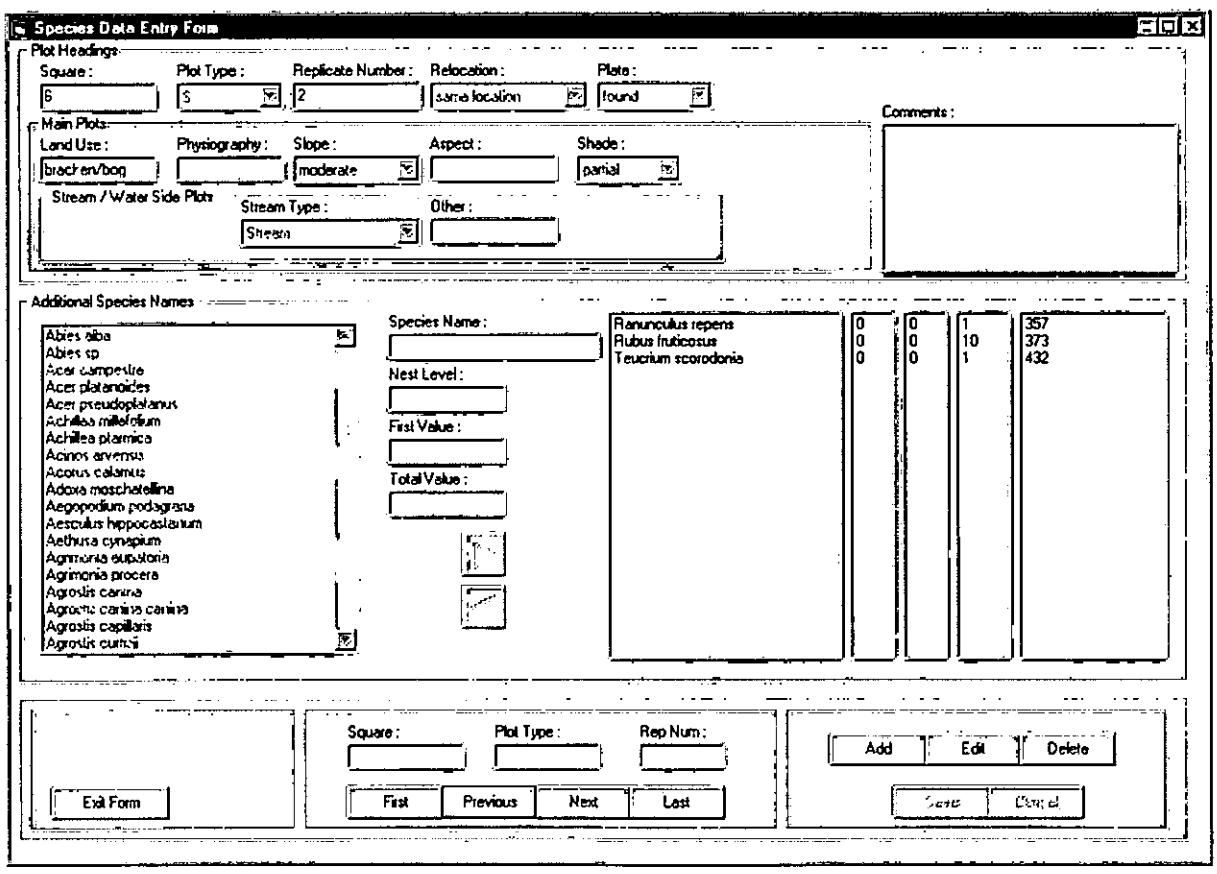

Figure 19: Species Data Entry Form

5.2 Meta-Data Management

Figure 20 shows the Meta-Data Management Form that will be used to document and explore all the data and storage sets.

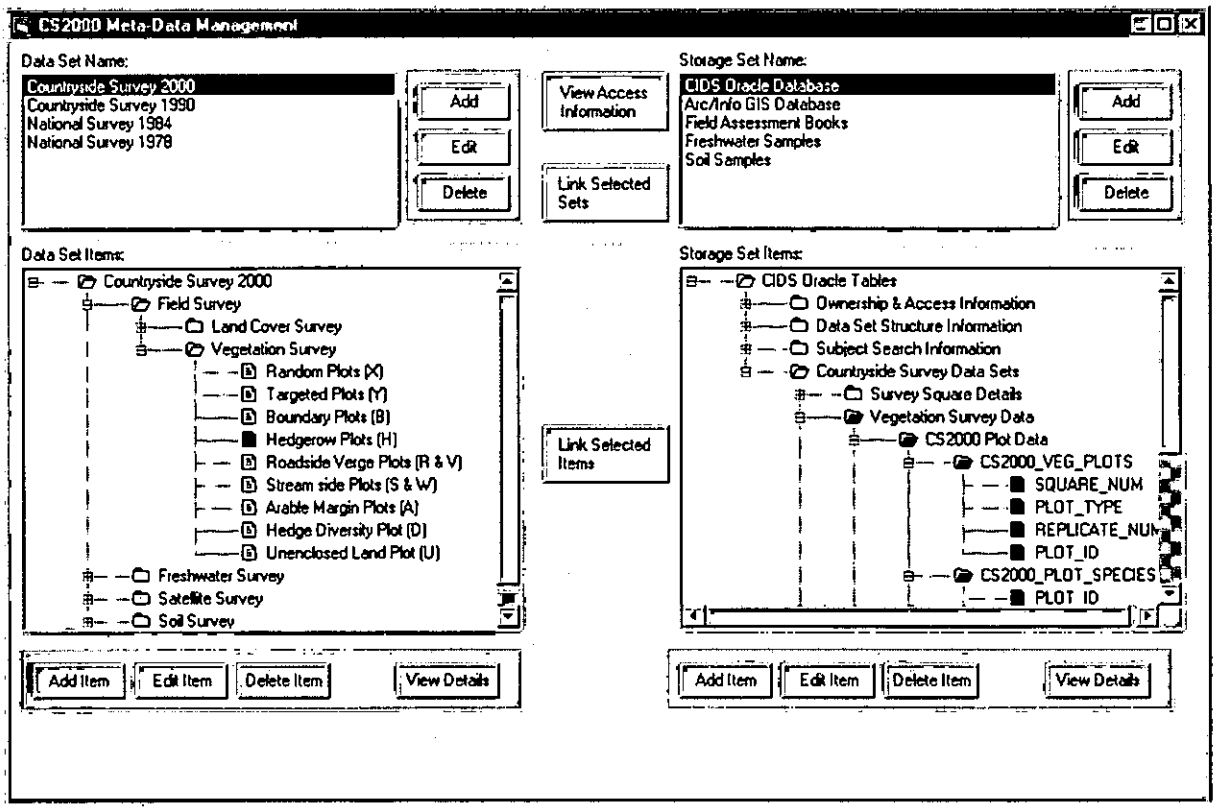

Figure 20: Meta-Data Management Form

Natural Environment Research Council