The ECN Information System

Susannah Rennie

The ECN Database was initially developed by Mandy Lane.

The development of the ECN Database is a team effort. The other members of the team are Lynne Irvine and Lorna Sherrin.

The ECN Information System

1. Introduction

The emergence of environmental change as an issue on the global agenda has generated a demand for access to objective, reliable and up-to-date environmental information. This requires information resources which provide long-term runs of data at a range of spatial scales both nationally and globally, and which are essentially:

- Reliable (stable, holding data of good and known quality)

- Secure (maintained in perpetuity with suitable access controls and backup systems)

- Integrated (enabling integration of multidisciplinary data across spatial and temporal scales)

- Comprehensive (incorporating integral data, metadata and knowledge)

- Flexible (designed to respond to changing requirements)

- Accessible (timely access to data for a wide range of users)

- Analytical (support for spatio-temporal analysis)

The Environmental Change Network (ECN) is the UK's long-term environmental monitoring and research network. ECN collects information on a broad baseline of integrated environmental information, enabling the analysis of relationships between environmental variables and across ecosystem components. The ECN programme is sponsored by a consortium of UK government departments and agencies with an interest in the environment, who contribute to the programme through funding either site monitoring or network co-ordination activities.

There are currently 12 terrestrial and 45 freshwater sites in the network, selected to cover as far as possible the main range of environmental conditions present in the UK. The monitoring programme includes a wide range of physical, chemical and biological 'driving' and 'response' variables, identified as being important for the assessment of environmental change. This suite of variables is measured at the same position within each site at the same time, using standard protocols incorporating standard quality control procedures.

ECN's robust and integrated system of information management forms the core facility for its programme. ECN's informatics approach views data management as an integrated component of a system of transforming raw environmental data into higher-grade knowledge. It aims to achieve a seamless flow of data from capture in the field through to access and interpretation tools for the detection of environmental change. A close liaison between database staff and scientists is considered important for ECN. It is important that there is a functional partnership between data managers and scientists so that the information can be tailored towards research objectives.

ECN is co-ordinated, and its database is managed and developed, by the ECN Central Co-ordinating Unit (CCU). Figure 1 gives an overview of the ECN information management system; its main components are:

- Data input: data capture at ECN sites, transfer by e-mail and validation;

- Data storage: database, meta-database and GIS;
- Data access: remote data access systems.

ECN may be extended in the future by the establishment of a Targeted Monitoring Network. This will concentrate on the effects of air pollution and climate change on biodiversity. If this new network is set up it will, where possible, follow the information management principles established by ECN. For full details, see the proposal (Morecroft *et al*., 2006).

2. ECN Data

At the start of ECN, working groups comprising scientists and statisticians representing a range of different environmental disciplines agreed a list of variables to be monitored. ECN has adopted the term 'core measurement' to mean an aspect of the environment on which a set of measurements will be made, e.g. Macro-invertebrates, Meteorology, etc. Standardised methods for sample collection and processing are set out for each core measurement in the protocols. These have been published (Sykes, J.M. and Lane, A.M.J., 1996; Sykes, J.M., Lane, A.M.J. and George, D.G., 1999) and are available online (http://www.ecn.ac.uk/protocols/index.asp).

The data requirements are an integral part of the protocols and include specifications of variables, units, reporting precisions, dimensions, resolutions, reference systems and quality assurance procedures. These specifications, together with as much information as possible about user requirements, have been used to design the database, construct standard formats for data transfer and standard field forms for each dataset.

The ECN database is complex due to the heterogeneous nature of the data being collected, in terms of the core measurement areas covered (from vertebrate populations to water quality), the formats involved (from field data to satellite imagery) and the range of temporal and spatial scales covered by the protocols. The spatial dimension for ECN sampling can relate to a point (e.g. automatic weather station), line (e.g. butterfly transect), area (e.g. land cover) or distributions (irregular or regular locations at varying distances). The time dimension can relate to an instant in time (e.g. temperature at 9:00) or may be summary or cumulative data for time periods (e.g. average wind speed, number of invertebrates trapped) which may or may not be contiguous.   ECN sampling intervals range from 15min to 20 years. Consequently an extensive meta-database, describing the data and their proper use, is also required.

Data Capture, Transfer & Validation

Data transfer via email using standard protocols

- *57 sites*
- *>250 variables*

Email transfer

QA feedback

Quality Assurance:
- *Control*
- *Validation*
- *Assessment*

ORACLE RDBMS

Meta-Database

Data Dictionaries, Quality Assurance, Research, Data Use, Publications

Database

Integrated database for all ECN Core Measurements

GIS (ArcGIS)
Spatial data handling and analysis

Data Access, Presentation & Analysis

Direct web-to-database interfaces

Summary data

Licensing system

Raw data

ArcGIS interface

Spatial data

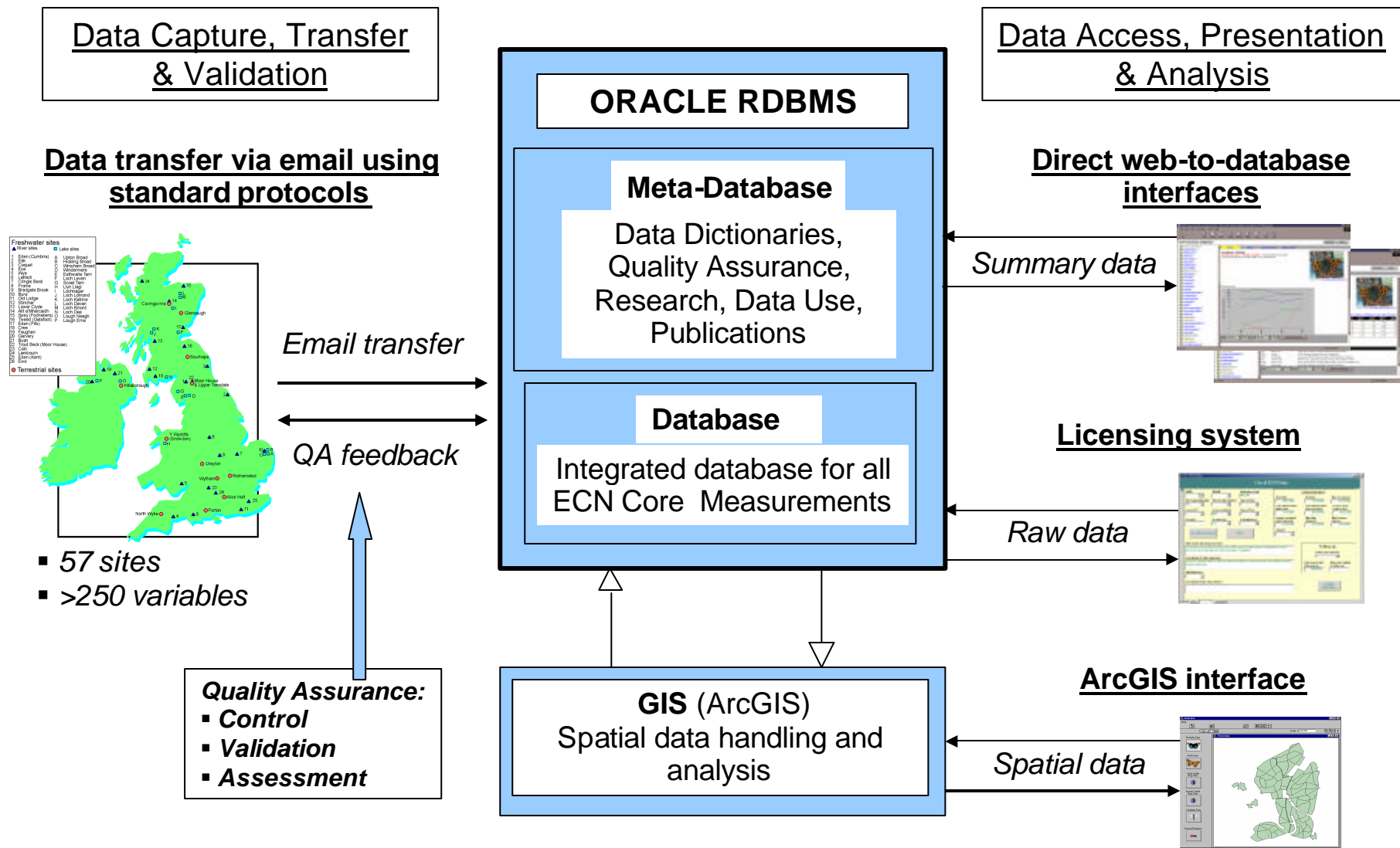**Figure 1: The ECN Information System**

3. Data Handling

ECN has adopted an integrated information management strategy. This considers all aspects of information handling as part of a unified system, identifying the necessary flows of data at each stage of the data handling process.

3.1 Data capture

ECN uses standardised and integrated data capture protocols to ensure that the data are comparable. ECN data capture requires mainly manual methods, recording on to standard field forms or alternatively on to maps where the measurement is concerned with spatial patterns over the site. Automatic data capture methods are also used to record measurements (e.g. hourly meteorological variables through automatic weather stations, 15-minute river flows at gauging stations) where appropriate. Wherever possible and appropriate, existing data capture techniques and common coding schemes have been adopted to maintain ECN's comparability with other sectoral networks.

Global positioning systems (GPS) are used for recording spatial information (e.g. vegetation plot locations). Aerial photography and remotely sensed imagery (satellite or air) are important sources of information about changing spatial patterns over time across ECN sites and their catchments; opportunities for using existing and specially commissioned remotely sensed imagery are exploited wherever possible. All ECN terrestrial sites were flown in 1994 to give colour air photography at 1:10 000 scale, these have been incorporated in the ECN GIS.

Standard recording forms for use in the field have been designed where required for selected protocols. The design of field forms takes into account ease of use in the field and ease of keying-in data in the laboratory, with the main emphasis being on minimising error. The use of computerised forms and mapping in the field is encouraged where sites have access to field computers, provided that output can be translated into ECN standard format.

Standard data entry forms with built-in validation procedures have been developed for manually recorded data where in-house data entry systems are not already in use. Data entry software is particularly useful where numeric coding systems for species are in use; numbers are less memorable and mistakes in one digit of a code can produce serious errors. This approach has proved popular for the ECN vegetation data and will be extended to cover other ECN protocols. This development could be linked to the introduction of robust computers for data entry in the field. The use of these are being monitored for future use when resources allow.

3.2 Data transfer

The ECN Site Managers are responsible for sending data to the CCU in machine-readable form using detailed data transfer documentation, which include rules for handling missing values and data quality information. The data transfer formats reflect the sampling requirements of the protocol, data

entry from field forms and to some extent ease of restructuring for database input. The wide variety of software systems in use by ECN organisations has made it difficult to standardise on specific software for data transfer. The most straightforward solution has been to standardise on ASCII text files in comma-separated format, which most software products are able to support and which are easily transferred by e-mail. For example:

BF,T05,01,20-JAN-1994,22-JAN-1994,12-FEB-1994,1-JUN-1994,Q,000

The data transfer documents detail the order that the CCU expects the data to be in the data record. A 'Q' separator is used to indicate where the quality codes begin in the record. The 'Q' separator provides a useful check on the accuracy of the data input since they occur in predictable position in the data record.

Rules for handling missing values, missing or non-standard sampling occasions and problems during sampling are also defined within the data transfer documentation. Pre-defined quality codes (which describe common sampling problems affecting ECN measurements) and free text (where these quality codes are insufficient or inappropriate) may accompany the data. These are included on the end of data records (i.e. they apply to a given sampling date at a single location). Further details on ECN quality assurance are provided in section 4.

3.2.1   Missing data

Data values of zero are equally as important as non-zero values, and it is important that the distinction between zero values and missing data is understood. For example, in the context of beetle trapping:

- 'no catch' when an insect trap was not set = missing data

- 'no catch' when an insect trap was set but no individuals found = zero

Missing data values are recorded as 'null' fields by simply including the separating comma in the data record where otherwise the data value would be. This preserves the order of the separating commas, so it remains clear which data field refers to which variable. For example:

BF,T05,01,20-JAN-1994,,12-FEB-1994,1-JUN-1994,Q,000

Information about the reasons for missing data is given either through the quality codes attached to the data records or, if no code is suitable, as free-text information accompanied by dates or date ranges.

ECN samples or recordings are usually made on standard sampling dates and according to the sampling periods specified in the protocols. However, there may be occasions where sampling or recording is not possible, and all data fields are missing. As a general rule, data records for all dates on which sampling/recording is due, even if missing, should be included in the dataset, with null data fields, and appropriate quality codes and/or text supplied. If another date is substituted for the sampling date, for example if a surveyor is prevented from making a recording because of bad weather but successfully attempts to make a recording the following day, then the sampling date can be

omitted. The substitute date is included instead, with the quality code 222 "non-standard sampling date" attached to the record.

3.2.2 Frequency of data transfer

E-mail is the preferred medium for transfer. Data are emailed to a central email account (ecn@ceh.ac.uk). This is accessible by all members of the ECN data group. There is a set of 'rules' available for how this email account is managed which are designed to ensure that all incoming emails are properly logged.

How often data are transferred, and the time-lag between data capture and data transfer, will depend to some extent on the core measurement protocol. Datasets are sent in quarterly for frequent measurements taken throughout the year. Data for less frequently measured protocols are sent in at an appropriate frequency.

3.3 Data processing

The receipt and management of data for a diverse range of variables from ECN's sites require significant administration and processing effort. On receipt, datasets are stored by the CCU in a network drive. Receipt of datasets is acknowledged immediately.

Datasets are logged by site, core measurement code and receipt date in the ECN meta-database, prior to processing. The meta-database logs data coming into the CCU and tracks it through the various stages of the data handling process (Rennie, 2002). It also links quality information to datasets and logs any changes made to the data or meta-data; thereby providing an audit trail for additions and changes to the ECN data. An MS Access Interface has been developed to allow for the easy entry of data into the meta-data tables, and to allow users to query this information.

Information on the data handling process is also logged on Excel Charts to provide a quick visual aid for data management staff and ECN participants to view where data are in the data handling process.

Data loading, transformation and validation programs have been developed and documented for each core measurement. The procedures perform numeric range checks, categorical checks, formatting and logical integrity checks, (e.g. on dates, number of samples, and links between datasets). Appropriate range settings for ECN variables have been selected following discussion with specialists in each field. They also incorporate routines to create the summary data as soon as data are added to the database, ensuring that the most up-to-date data are available for users (see section 6.2).

The time taken to process datasets is strongly affected by the type and number of errors they contain, and may vary between ten minutes and several days. Where possible, queries and problems are solved through communicating over e-mail; although if errors are numerous or particularly difficult to decipher a request is made for a repeat dataset to be sent.

ECN aims to have no more than a six-month time-lag between data collection and availability in the database, for measurements sampled through out the year. Seasonal measurements should be available in the database by the end of June in the following year. The delay reflects the degree of quality assurance necessary to establish a long-term environmental research database. However, where a rapid response is required, instant access to recent data can be provided with qualifications about the degree of validation performed and consequent data reliability. Regular reminders are emailed to site contacts to encourage data submission.

4. Data Quality

Quality assurance is an essential part of any long-term programme. Data of poor or unknown quality are unreliable. It is important to set quality standards at the outset of a data-gathering exercise, but it is equally important to monitor how far those standards are met, and to ensure that this information accompanies the data for future use. The ECN approach to quality assurance is shown in figure 2.

4.1 Quality control

The ECN Protocols are documented standard operating procedures (SOPs) designed to ensure consistency in measurement methods and data handling over time and across all of ECN's sites. They incorporate target specifications for quality criteria such as accuracy and recording resolution, where appropriate. Quality control procedures go hand-in-hand with SOPs and have been included in the protocols, e.g. correct handling of equipment and samples, maintenance schedules and calibration specifications, and unambiguous instructions for measurement and data handling.

Details of any deviations from these procedures, faulty instrumentation and common problems occurring during the sampling period can be reported as part of the standard data transfer formats either as pre-defined quality codes or as free-text if the standard quality codes are insufficient. These are attached to data records, (i.e. they apply to a given sampling date at a single location). The list of quality codes uses a 3-figure numeric sequence; codes are added to the list if problems continue to recur. Any number of codes can be appended to a single data record. This information is stored in the ECN Meta-database.

4.2 Quality Assessment

Quality assessment can be regarded as a monitoring exercise to keep measurements 'on course' by feeding information back to the data capture stage. They provide quantitative data on, for example, observer and sampling effects in biological measurements and variations in chemical analyses at different laboratories. In ECN, where the measured feature can be kept (e.g. archived invertebrate samples), or re-visited (e.g. vegetation plots), the accuracy of identification has been assessed at a later date through sub-sampling by an independent expert (Scott, W.A. and Hallam, C.J., 2003).

The quality of more ephemeral measurements such as meteorology or water quality can only be similarly assessed by running duplicate or parallel systems. Duplicate systems are expensive, and in practice assessment normally involves regular checks for instrument drift and recorder error. As part of its terrestrial site network, ECN runs manual weather stations concurrently with automatic stations to provide some parallel records, and has regular maintenance schedules for equipment which help to maintain standards across the network and over time. Where possible, when new instrumentation or methods need to be introduced, new and old systems are run in parallel to assess the relationship between the two.

ECN sites send water samples to their own associated laboratories for analysis. The cost of standardising methods of analysis across all ECN laboratories is prohibitive; instead analytical guidelines list reference and approved techniques for each determinand with corresponding limits of detection. Organisations need to maintain their own continuity in methods for existing long-term runs of data. Each laboratory practises its own internal quality control, and most participate in national quality assurance schemes. In addition, ECN has conducted inter-laboratory trials using standard solutions. These accompany each batch of water samples from the field. Analysis of the spread of values across laboratories can help to highlight problem areas. Details of the existing methods used in each laboratory are incorporated into the ECN meta-database, and linked directly to each analytical record.

4.3 Data validation

Data validation can be regarded as part of quality assessment and involves screening data for 'unacceptable values' which may have occurred at any stage during data capture and handling. As described in section 3.3, standard validation procedures are applied to datasets before import to the database.

When 'unacceptable values' are identified, ECN adopts a cautious approach to discarding data on the principle that apparent errors may be valid outliers. Values which fall outside the ranges will only be discarded if there is a clear explanation, such as instrumentation error, and corrections made where possible. If the reason is unclear, then the values will be stored, but qualified in the meta-database. Data values identified by validation software as 'unacceptable' are treated in one of three ways:

- where values are clearly meaningless due to a known cause, (e.g. an instrumentation fault, and cannot be back-corrected), the data are discarded and database fields set to null (no data);

- where values are clearly in error, or out of range due to known calibration errors, and can be back-corrected, data are stored separately until the correction can be made;

- where there is no straightforward explanation for outliers, the data are stored in the database, accompanied by meta-data 'flags' and associated text.

Sites are strongly encouraged to check their data before sending them to the CCU, but not to alter them unless a reason for error is clear, and in any case to inform the CCU of their actions.

The data validation checks described above are important 'first-pass' procedures, but they are relatively coarse and may fail to identify erroneous data within the valid range. More subtle problems may only be revealed through multivariate or time-series checks, based on known processes or expected patterns in the data. Procedures for implementing these as a 'second-pass' validation stage are currently being developed.

4.4 Storing quality information

Target specifications for quality criteria are stored as meta-information alongside instrument and sampling details, and units of measurement. Any deviations from these specifications or from the sampling methods given in the ECN Protocols are recorded with time-stamps. Details of laboratory methods and associated detection limits are stored similarly. Missing data and outliers which have been revealed by a quality assessment exercise but which cannot be corrected are qualified, using pre-defined quality codes or free-text descriptions. This information may be associated either with a particular sampling occasion or with an individual measurement variable on a sampling occasion. Site managers also use these codes or free text to describe factors affecting sampling outside their control, instrument damage or site management effects. Results of quality assessment exercises, (e.g. laboratory trials or vegetation re-survey), are also incorporated. All meta-information is linked directly with the data to which it relates.
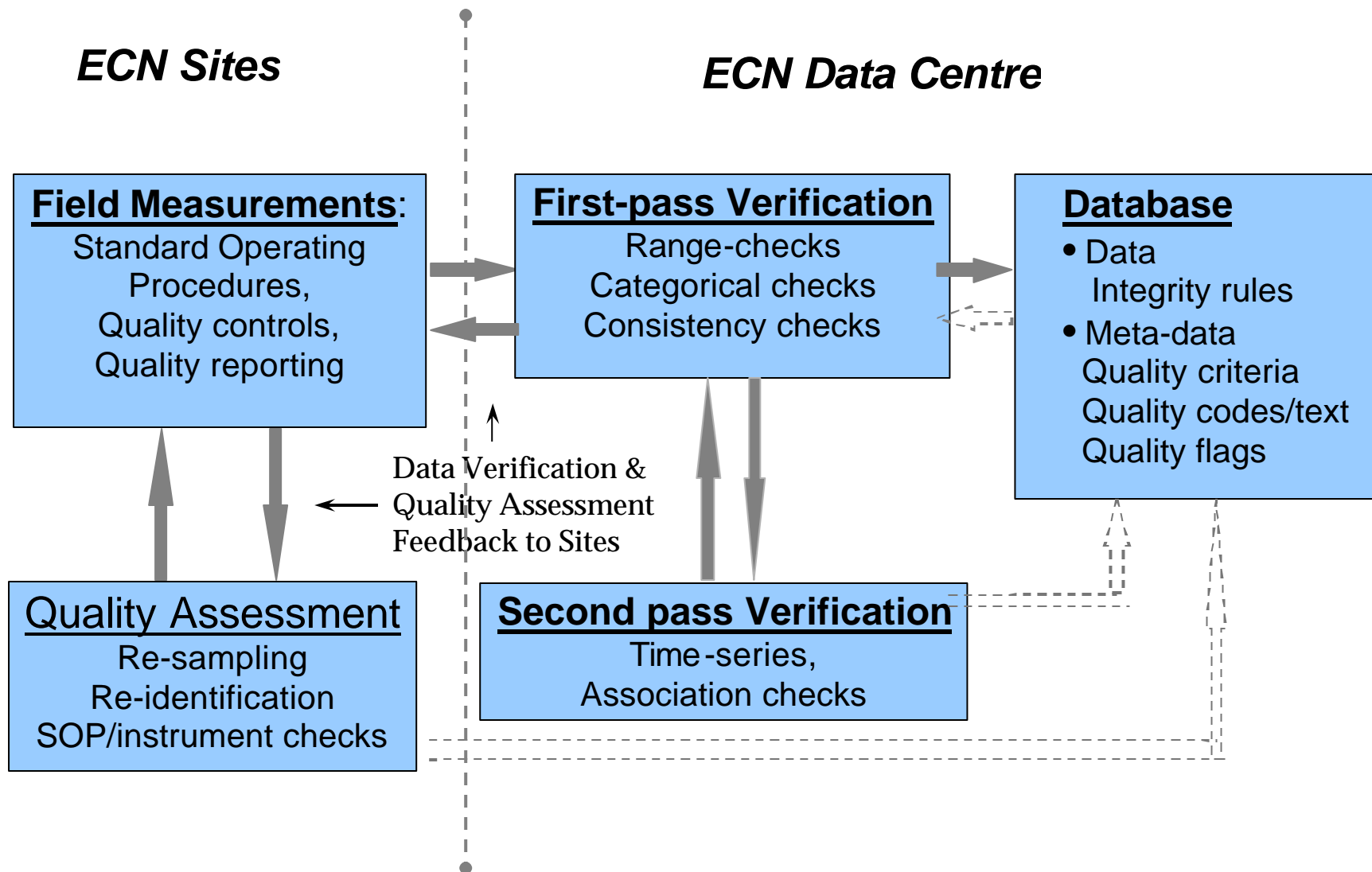
ECN Sites

ECN Data Centre

Field Measurements:
Standard Operating
Procedures,
Quality controls,
Quality reporting

First-pass Verification
Range-checks
Categorical checks
Consistency checks

Database
• Data
   Integrity rules
• Meta-data
   Quality criteria
   Quality codes/text
   Quality flags

Data Verification &
Quality Assessment
Feedback to Sites

Quality Assessment
Re-sampling
Re-identification
SOP/instrument checks

Second pass Verification
Time-series,
Association checks

Figure 2: ECN Quality Assurance

5. Database Design

The ECN database is an integrated information resource, which stores all data and associated meta-data from ECN's core measurements collected at its network of sites (see figure 1). These data, along with other historic data from ECN sites, are held in standardised structures within the ECN database, to support the cross-disciplinary analyses necessary for environmental change research.

The design of any database must focus not only on the data and their interrelationships but on the purpose and requirements of the activity it is to support. ECN delivers data to a diverse community of scientific, policy and public users (see section 6). Where, as in the case of ECN, the styles of data are diverse and the range of potential uses difficult to define, the database system needs to be integrated (to aid data accessibility), but as versatile as possible to allow new structures to be generated, new datasets to be incorporated when required and ultimately give users free but guided access to data.

The database uses Oracle relational database management software, with links to Arc GIS for spatial data handling; this runs on a local area network with high-speed links to the Internet for remote access and incoming data. The database is logically divided into data and meta-data tables. As described in sections 3 and 4, data are regularly sent in from sites and are quality assured before being lodged in the database; a system of quality flags and codes ensures that the data are properly qualified in the meta-database.

5.1 Central *vs.* Distributed Databases

ECN's current strategy is to maintain a centrally managed core database with good remote access provision, and to establish links to other site-based and sectoral network databases using a distributed approach.

Networking developments over the past decade have meant that the physical location of a database is less significant for the user; the degree to which information systems are centralised or distributed begins to depend more on functional requirements and management resources. A more distributed model has clear benefits for environmental programmes concerned with co-ordinating existing databases in that it avoids duplication of data and effort. Also, for multi-agency programmes, distributed models allow data originators to maintain control of their data resources, while still integrating their data with their partners in the programme. But for ECN, which was founded on the collection of new environmental data, a centrally located database with remote network access was considered the most appropriate model. (Also when the database was set up (in 1992) a central database was the only viable model available). This has made it easier to ensure a fully integrated system with the required data quality and maintenance standards within the resource constraints of the programme, whilst meeting the requirement for direct and rapid access to the data.

The Grid is an emerging infrastructure that enables computers and networks to be linked together into a seamless common resource (see section 7). This raises the possibility that the data could be held by the owners of the data and

made available over a Grid-enabled system. This would require suitable computing resources and database staff to be based at each of the sites and/or partner organisations. However, this technology is still in its infancy and may not yet be well-developed enough to form the basis of an information management system for ECN. A watching brief will be kept on this technology because it could be adopted later on.

5.2 Core Database

The core database stores raw data at the resolutions specified in the ECN protocols. An associated summary database consists of monthly, quarterly, and/or annual summaries of these data using summary statistics appropriate to each measurement, as advised through ECN's expert committees. Summary data are generated as soon as new data are lodged in the database. Access to the summary data is freely available, whereas access to the raw data is more closely controlled (see section 6).

Each ECN Site has been assigned a 'Site Identification Code', and each ECN Core Measurement (or sub-category of a core measurement) a 'Core Measurement Code' which is referenced in the database and in datasets transferred to the CCU. A list of the site and measurement identification codes can be found at:

- Freshwater - http://www.ecn.ac.uk/protocols/Freshwater/codes.pdf
- Terrestrial - http://www.ecn.ac.uk/protocols/Terrestrial/codes.pdf

Data for each sampling location for a given core measurement within an ECN site are regarded as a logical 'dataset'. Each sample or recording occasion and associated measurements are uniquely identified in space and time by:

- Core Measurement Code
- Site Identification Code
- Location Code
- Sampling Date

This combination of information is allocated a unique identifier, which forms the dataset key.

Datasets may be spread over several physical tables, or several datasets may occupy one physical table, depending on their structure within the relational model. Constraints on fields and tables and referential rules between tables are applied wherever possible to automate data management and maintain the integrity of the database.

5.3 Meta-data

Meta-information is an essential part of the database: it is not sufficient simply to provide the data values themselves, but it is also necessary to include their description, derivation, measurement parameters, and quality criteria. At its simplest terms meta-data are 'data about data'. They can be regarded as a continuum of characteristics from general information about a dataset to

specific details about a data item or object within that dataset, with data quality as an important component. The value of meta-data lies in providing information about data availability, appropriate use, access and transfer methods essential for the integration of datasets and for efficient maintenance and use of data resources. The boundary between data and meta-data is not necessarily clear-cut; meta-data for one purpose might be regarded as data for another. A meta-data system should ideally be sufficiently flexible to allow for different views of databases by different users.

In ECN, a central meta-data table (M_DESC) forms the principle link between data and meta-data through the dataset key (see section 5.2). This table provides the essential meta-information for the dataset, such as the ECN core measurement description and sampling protocol, space/time dimensions of sampling strategy, physical location of sampling (which may be a point reference or a link to a spatial definition held with the GIS, depending on spatial type), data ownership and responsibility etc.

Associated meta-data tables hold linked information on units of measurement, quality criteria, quality codes and text relating to sampling occasions, and data dictionaries (e.g. species coding lists). Any deviation from the protocol specification, together with quality codes and text describing factors affecting sampling outside control are stored.

Protocols may occasionally change over time, for example when new instrumentation is introduced. Data generated by running new and old systems in parallel are used to define the relationship between them so that long-term runs of data can be maintained; this information is also included in the meta-database. Results of quality assessment exercises, for example laboratory trials or vegetation resurvey, are also stored within the meta-database and linked with the original data.

Other sections of the meta-database include:

5.3.1 The PRU Meta-Database

The PRU Meta-database is designed to hold **P**ublication, **R**esearch and Data **U**se information about the ECN sites. This information is required to monitor the use of ECN data, maintain publication lists for the network and improve the co-ordination of research activities within the network (Rennie, 2000).

MS Access Interfaces have been developed to allow for the easy entry of data into the meta-data tables, and to allow users to query this information. The data are also made accessible through the web site (http://www.ecn.ac.uk/PRU/pru.asp).

5.3.2 The Contacts Meta-Database

The Contacts Meta-database holds information about the participants in and users of the ECN programme. Changes in the information held about these individuals can be tracked through time. Data are held in accordance with the Data Protection Act (see section 6.4.3). An MS Access Interface has been developed to allow for the easy entry of data into the meta-data tables.

5.4 Database Documentation

The ECN database structures are documented within the meta-database. This is accessible through a web interface:

http://www.ecn.ac.uk/database_documentation/index.asp

This documentation is used to automate how the data structures are described to users who request ECN data. It is also useful for describing the database structures to partners when working on integration projects.

For the majority of the core measurements there are tables for each individual site (e.g. for the NO_2 data - D1AN_DRA, D1AN_POR....etc). So to reduce redundancy in the database, types of database tables have been documented for the core measurements rather than every single table (i.e. a generic D1AN_sss table has been documented). For these tables, the table name is split into a prefix and suffix in the meta-data - so actual table names can be built if necessary (by substituting the suffix with the relevant site code/s).

Field names are stored separately from information about tables. This means that information about fields can be re-used if they appear in several tables. This can be used to check that similar information is stored consistently across the database and help to identify similar data that can be pulled together if appropriate.

5.5 Database Table Naming Conventions

The ECN data tables use the following naming conventions:

- Most ECN data tables have the form:

  Dnmm_sss

  where n is a series number for the table, mm is the measurement code, and sss is the 3-character site code.

  For example, D1MA_MOO (for automatic weather station data at Moor House).

- Some tables store data for a group of sites. The grouping type is used in the table name rather than the site code.

  For example, D1FIN_RIVER (the freshwater invertebrate data from the river sites).

- Some tables store data for all the sites so the site code is not required in the table name.

  For example, D1BF (for frog development event dates for all sites.)

- Where there is a known problem with data then they are stored in separate tables (indicated in the table name by an 'X' before the measurement code).

  For example D1XMA_sss (problem MA data from a particular site).

The ECN meta-data tables use the following naming conventions:

- Main meta-tables are given names of the form M_xxx, where xxx is a descriptive name of the information being stored.

  For example, M_SITE (for the meta-information about ECN sites).

- Look-up (reference) tables are given names of the form M_REFxxx, where xxx is the fieldname containing the code for which the description is required.

  For example, M_REFIG_SPEC (for the IG species codes)

- Link tables for many-to-many relationships are given names of the form M_LNKxTOy, where x and y are the names of the tables to be linked.

  For example, M_LNKNIDTOPARID (to link individuals personal details to their address).

- There are some administrative meta-data tables which are given names of the form A_xxx, where xxx is a descriptive name of the information being stored.

  For example, A_DATAPROCESSING (to store any changes made to the data).

The ECN summary tables use the following naming conventions:

- Data tables take the form:

  Sxn_mmm

  where x indicates the time period over which data have been summarised (A = annual, Q = quarterly, M = monthly), n is a sequence number for the table and mmm is the ECN measurement code

  For example, SA1_FIN (for annual freshwater invertebrate data)

- Associated meta-data tables (holding the number of samplings from which the data was derived) have the form:

  m_base_mmm

  where mmm is the measurement code

  For example, M_BASE_FIN (for freshwater invertebrate meta-data)

5.6 Database Security

The long-term nature of the programme has meant that the use of reliable, well-known and well-supported database software is of paramount importance for long-term security. Database security is an important consideration, to avoid corruption or loss of data through system faults and to protect against unauthorised access. Incremental back-ups of the database are made daily, a full back-up is made weekly, and monthly back-ups are kept for one year, off-site. Storage media are renewed regularly. Access controls and security monitoring software are in operation to prevent unauthorised use.

To help protect the integrity of the database a number of database accounts have been set up to perform various tasks. These accounts have only the permissions to do these tasks.

6. Database access

Availability of data is central to ECN's success as a resource for environmental research, policy purposes and public information. ECN data are used by a diverse community of science, policy and public users. Broadly, ECN anticipates three main types of data users:

- Scientific Researchers - who require access to raw data for detailed analysis of spatial and temporal patterns in the data. Scientific users should not be constrained by a set of predefined queries, they will require free rein to explore patterns in the data through ad hoc queries and flexible data exploration tools.

- Information Brokers (e.g. ECN sponsors or policy-makers) – who may not require access to the raw data but who need summaries and interpretations of scientific research on which to base policy decisions. They may require guided access to summary data for inclusion in reports or interpreted information about trends in the database, for example, whether a pattern indicates a long-term trend that is likely to continue.

- General public and school students – who are likely to want easy-to-understand interpreted information of environmental change, such as environmental indicators which enable easy assessment of how aspects of the environment are changing and how these changes could affect their lives. They may also want guided access to summary data if they wish to explore the data further.

A particular challenge for ECN is to provide data access methods to suit these different styles of use, which can give sufficient guidance and information to users unfamiliar with the structure of the data whilst at the same time providing users with flexibility in data query and presentation.

ECN data are owned jointly by their originating sponsoring organisation and NERC; the ECN sponsors have agreed a system of user licensing and authorisation for access to high-resolution (raw) and summary data:

- Raw data are available under licence with charges levied according to the proposed use.
- Summary data are freely available through the ECN web site and published data digests.

ECN promotes the use of its data through a number of different data access methods to suit a range of users. Data are made accessible according to the ECN data policy (http://www.ecn.ac.uk/datapolicy.htm).

6.1 Raw Data Access

The data licensing procedures are administered by the ECN CCU. An application form for access to the raw data is available on the ECN Web site

(http://www.ecn.ac.uk/request_form.asp). This is automatically e-mailed to the CCU on submission. The CCU checks the requests and the action taken depends on how the user intends to use the data:

- **Data to be used for non-commercial purposes** - In 2007, the Steering Committee agreed to allow the CCU to have delegated authority to license all reasonable requests for non-commercial uses related to environmental change research without consulting the data owners. (Sponsors were given the option to opt-out of this if they wanted). An annual report on data requests is sent to the sponsors.

- **Data to be used for commercial purposes -** Requests from commercial organisations are liable for a data use fee under Government (DTI) rules. Agreement could not be reached by the Steering Committee (2007) on a common simplified charging policy for commercial use of ECN data. Therefore the Committee agreed that the CCU would no longer handle these requests on behalf of ECN as a whole. Any such requests are forwarded to the relevant sponsors to handle directly according to their individual policy.

Before being given access to the data, users are asked to sign a licence agreement which defines the terms under which ECN data may be used.

Data are normally emailed to requestors as fixed-format files. To increase the efficiency of releasing data to users a data extraction library has been set up. This contains extracted data files (in year chunks) which can be quickly zipped up in whatever combination is necessary when new requests come in. This allows extracted datasets to be reused many times - making the management of the system more efficient.

Direct database access can also be provided for licensed scientific users who are familiar with SQL and the ECN database structures. Users can access the database *via* 'Telnet' and use SQL directly, or use a PC client front-end which constructs the SQL from a Windows-style interface.

6.2 Summary Data Access

For internal ECN data management and computer-literate scientific users, general-purpose access methods provide the most flexibility in querying data and extracting it for analysis. However, use of these means having knowledge of specific systems and extensive understanding of the structures and relationships within the database, as well as the time to work with the raw data, and so these methods may not be suitable for all users. Therefore, ECN has developed 'tailored' interfaces to the ECN summary database which require little or no initial learning process for the user, primarily through the web site but also published as data digests.

6.2.1 Web-to-Database Interfaces

The web is an obvious development platform for data discovery, query and development systems and it is a highly accessible broadcast medium where up-to-date information can be explored at low cost to the users. It is highly

versatile and the addition of database links to standard Web pages can generate a powerful information interface; enabling text, images and data to be presented together, allowing users to progress from browsing information to a guided database query, display and retrieval system.

All of ECN's web-to-database interfaces incorporate dynamic links to the database ensuring that the most up-to-date data are always available to users. These dynamic interfaces are 'costly', in development time, to set up. However, once they are up-and-running there is very little ongoing maintenance since the output for users are generated on-the-fly. These dynamic links are achieved by using Active Server Pages (ASP) with an ADO/ODBC connection to the Oracle database. ASP allows programming scripts to be embedded within web pages (Rennie *et al*., 2000).

Since users are accessing an automated system they do not have the opportunity for dialogue with the data originators or database staff therefore the systems described below have been designed to provide meta-information. Where download facilities have been provided these meta-data are automatically included in the download.

For ease of maintenance, where possible the systems described below have been written in a generic fashion allowing new measurements, types of graph and images to be added through meta-database table updates rather than through programming code.

There are a number of web-to-database interfaces available from the ECN web site:

- **Summary Database Interface**: http://www.ecn.ac.uk/Database/index.html

  This interface enables users to build their own database query by selecting any combination of ECN sites, core measurement variables and date ranges for instant on-the-fly generation of tables, cross-tabulations and graphs. Data (and associated meta-data) may also be downloaded *via* e-mail in 'column' format, for import into local software (Rennie *et al*., 2000).

  Access to data over the ECN summary data web interface is monitored by both IP and email address, and users are asked to provide information about how they intend to use the data before they download any data. Monthly usage statistics are generated automatically (see section 8).

- **Indicators**: ECN data have been used to develop indicators of climate change, water quality and biodiversity. These are available on the ECN website and update automatically as new data are added to the database.

  - Climate Change - http://www.ecn.ac.uk/CCI/cci.asp

  - Water Quality - http://www.ecn.ac.uk/freshwater2/index.asp

  - Biodiversity - http://www.ecn.ac.uk/CCI/composite.asp

- **Real-time access to weather data**: Direct links to ECN Automatic Weather Stations generate 'real-time' displays of weather conditions. The AWS generates hourly data that are transmitted to the ECN CCU via a modem link. These data are stored in a database and graphs and tabulations of selected data are automatically generated for display on the web (http://www.ecn.ac.uk/online_aws/index.asp). Direct links from AWS's

at other ECN sites are planned; as well as from other automatic monitoring instrumentation e.g. water quality loggers. Once these links are considered sufficiently reliable, they will be used to download data automatically into the database through the data validation and input software.

- **PRU:** The PRU meta-database (see section 5.3.1) holds information about ECN Publications, Research and Data Use. This interface (http://www.ecn.ac.uk/PRU/pru.asp) allows users to generate publication lists using various criteria (e.g. lists for a particular author, year, site or keyword) and create on-the-fly graphs of how ECN data are used. This information is an important component of the ECN performance indicators (see section 8).

- **Site information:** The meta-data on ECN sites can be queried and displayed on the web (http://www.ecn.ac.uk/sites.htm). These web pages display descriptive and contact information for the site. Information from the ECN database has also been incorporated to show general characteristics of the site (e.g. mean annual rainfall). Links to Multimap have been included to help users locate each site.

- **Weekly photos:** Some sites take weekly photos of the conditions at their sites. These are emailed to the ECN CCU and displayed on the website (http://www.ecn.ac.uk/photos/index.asp). The information on these images are stored in the ECN database allowing users to view the archive of images or view the photos for all available sites from the same week.

- **Images:** A sizeable collection of images relating to the ECN programme has been created. Information on these has been stored in the ECN database and they can be viewed and searched on the ECN extranet.

6.2.2 Data Digest

The ECN data digests are produced annually (in time for the annual autumn meetings). They aim to provide a snapshot of the data collected in the previous year and also provide a comparison with data collected in earlier years. Since 2003, an automated system has been used to create the digests. This was set up in MS Excel (to preserve the formatting used in previous years). Macros are used to query the data direct from the ECN summary database and automatically format them in the required digest format.

The digests are a useful tool for ECN participants to check the data held in the ECN database for their site.

6.3 Spatial Data

ECN is developing web-based interfaces to its spatial data using Arc/IMS tools. This development will link with the time-series interfaces above, allowing users to explore spatial and temporal patterns within a single system.

6.4 Relevant Legislation

There are several important pieces of information that inform how ECN should manage and release data. The Information Commissioner is responsible for administering this legislation in the UK. Full details of the rights and responsibilities provided by the legislation are available on their website (http://www.ico.gov.uk/). However as a quick summary:

6.4.1 The Freedom of Information (FOI) Act

The FOI Act (and the Environmental Information Regulations (see section 5.5.2)) were introduced on January 1st 2005 and have a major effect on how public bodies manage data access. This legislation was introduced as part of the government's plan to promote openness and transparency within all public bodies.

The terms of the FOI Act state that public bodies are legally obliged to respond to all written requests for information within twenty working days. Requests can be made by anyone from anywhere. Although its main aim is to increase the openness and transparency of all public bodies, there will be occasions when information cannot be released, for example where disclosing information would breach the Data Protection Act (see section 5.5.3). But the scales are weighted in favour of openness.

Anything written down, stored on a computer server or retained as email is potentially open to access by anyone. Notes scrawled in the margins of meeting minutes, emails sent from work email addresses and post-it notes are not exempt – embarrassment is not an excuse for non-disclosure!

Freedom of access to information doesn't automatically mean free of charge. Details of charging regimes have yet to be finalised by Parliament.

6.4.2 Environmental Information Regulations (EIR)

EIR are similar to FOI but obviously only apply to environmental information. Unlike FOI requests, those under EIR do not have to be in writing. In all other respects they should be treated in the same way. They cover any environmental data held by a public body, regardless of who owns the data.

6.4.3 Data Protection Act

The Data Protection Act requires organisations which handle personal information to comply with a number of important principles regarding privacy and disclosure. The Act works in two ways. Firstly, organisations dealing with personal information must comply with eight principles, which ensure that personal information is:

- Fairly and lawfully processed
- Processed for limited purposes
- Adequate, relevant and not excessive
- Accurate and up to date

- Not kept for longer than is necessary
- Processed in line with the individuals rights
- Secure
- Not transferred to other countries without adequate protection

The second area covered by the Act deals with the rights that individuals have, including the right to find out what personal information is held on computer and most paper records.

7. Data Integration

In order to participate in global change research, ECN must interact with other existing thematic, historic, national and international databases, which will have captured data in different ways according to different criteria. Research should theoretically be unconstrained by the way data are organised in these databases. Grid technology for dynamic linking of distributed heterogeneous databases is only just developing. Environmental monitoring and research programs are initiating partnerships to develop data grid services and semantic mediation systems for environmental data.

The issues of data integration and access in a heterogeneous data environment are not trivial, particularly given the volume and complexity of environmental data. The solution lies in devising methods for describing terms and their relationships between databases which can be used to translate and transform data into comparable integrated forms (semantic mediation). To achieve truly 'interoperable' databases these descriptions must be embedded in and used dynamically by the system to create a unified data discovery, query and delivery interface across distributed data sources.

One of ECN's strategic developments involves the use of new technology to dynamically link its database with other distributed national and international long-term monitoring databases. ECN is working with European partners, through the EU Framework VI 'ALTER-Net' Programme, to develop a core facility to link such data sources, including a core extensible ontology that will handle the exchange of semantics necessary to integrate these data across Europe. The aim is to provide single gateway access to a range of data and analytical tools. For more details of this work see Work Package I6 of ALTER-net (http://www.alter-net.info).

8. Performance Indicators

The ECN Data Centre measures its performance using a number of indicators:

- Number of Data Licences Issued – These data are available from the PRU database (http://www.ecn.ac.uk/PRU/pru.asp).
- Summary Database Usage – Use of the summary database interface (including downloads) is logged in the ECN Meta-database.
- ECN Website – Software is used to track the usage of the main ECN website. Summary figures are stored in the ECN meta-database.

- Number of Datasets Received/Processed – These data are available from the ECN data processing meta-data tables (see section 3.3).

9. Bibliography and References

Lane, AMJ. (1997). The UK Environmental Change Network Database. An Integrated Information Resource for Long-term Monitoring and Research. Journal of Environmental Management, 51 (1), 87-105.

Lane, AMJ. and Parr, TW. (1998). Providing Information on Environmental Change: Data management strategies and Internet access approaches within the UK Environmental Change Network. Proceedings of 10th International Conference on Scientific and Statistical Database Management, IEEE Computer Society, Los Alamitos, California.

Lane, A.M.J., Rennie, S.C. and Watkins, J.M. (2004). Integrated Data Management for Environmental Monitoring Programmes. In: Environmental monitoring, edited by G.B. Wiersma, 37-62. Boca Raton: CRC Press.

Morecroft, M.D., Sier, A.R.J., Elston, D.A., Nevison, I.M., Hall, J.R., Rennie, S.C., Parr, T.W. and Crick, H.Q.P. (2006). Targeted Monitoring of Air Pollution and Climate Change Impacts on Biodiversity. Centre for Ecology and Hydrology.

Parr, TW. and Hirst, DJ. (1999). The UK Environmental Change network and the Internet: their role in detecting and interpreting environmental change. In: Advances in Sustainable Development. Environmental Indices: Systems Analysis Approach, EOLSS, 223-236.

Rennie, SC. (2000). ECN meta-database of research, data use and bibliographic information at sites. ECN report.

Rennie, SC., Lane, AMJ. and Wilson, M. (2000). Web Access to Environmental Databases: a Database Query and Presentation System for the UK Environmental Change Network.. Proceedings of the 2000 ACM Symposium on Applied Computing, 2, 894-897. Available online.

Rennie, SC. and Jackson, D. (2001). Using environmental data on the Internet. Teaching Geography, 26(1), 33-35.

Rennie, S.C. (2002). ECN Data Processing. ECN report.

Scott, WA. and Hallam, CJ. (2003). Assessing species misidentification rates through quality assurance of vegetation monitoring. Plant Ecology, 165, 101-115.

Sykes, JM. and Lane, AMJ. (1996). The UK Environmental Change Network: Protocols for standard measurements at terrestrial sites, The Stationery Office.

Sykes, JM., Lane, AMJ. and George, DG. (1999). The UK Environmental Change Network: Protocols for standard measurements at freshwater sites, ITE.