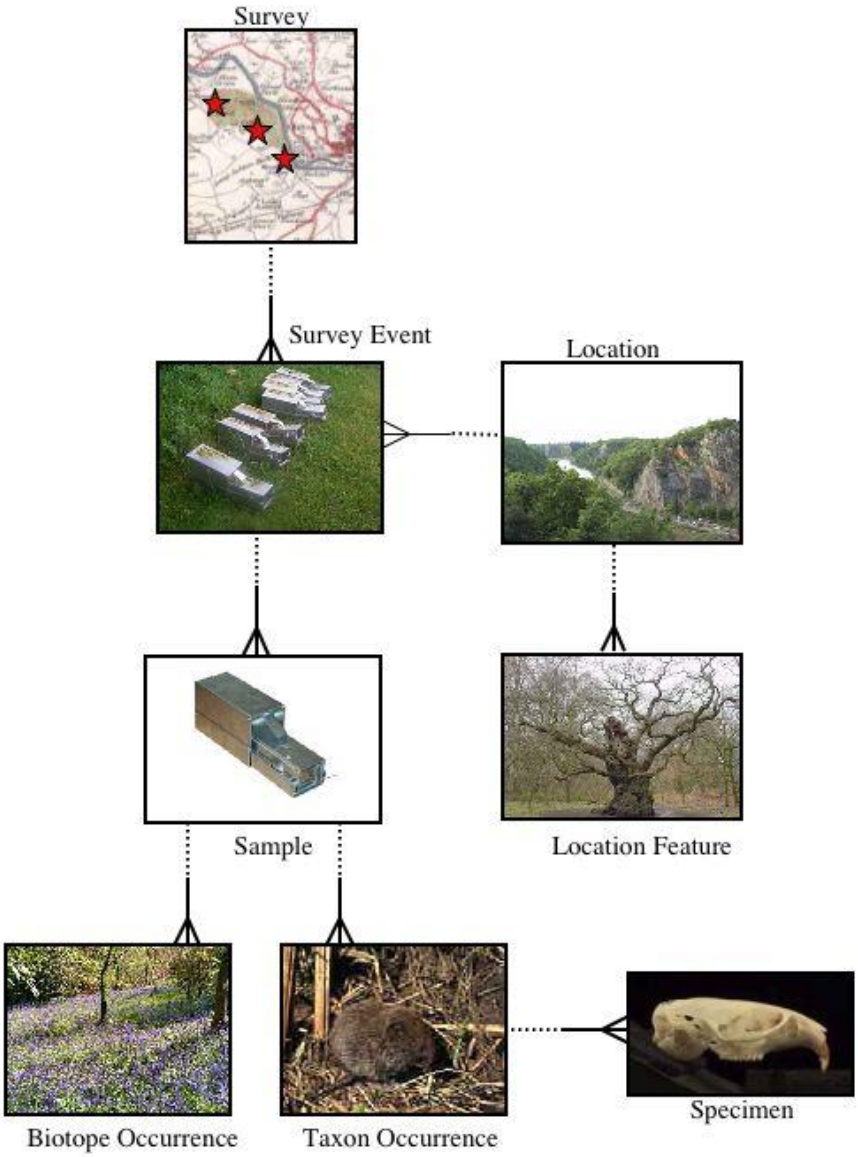
The NBN Data Model

National Biodiversity Network

Sharing Information about Wildlife



Part 1
Description of the Model

Charles Copp
Environmental Information Management

June 2004

defra
Department for Environment Food and Rural Affairs

The NBN Data Model

Part 1
Description of the Model

June 2004

Charles Copp

Environmental Information Management
8 The Paddock, Clevedon
North Somerset. BS21 6JU

Tel: 01275 874128
Email: *eim@globalnet.co.uk*

Contents

Special Notes

All Acronyms and all words in the text emphasised with italics are defined in the Glossary. Explanations and URLs are also given in footnotes

The NBN Data Model

Charles Copp

June 2004

Part 1
Description of the Model

Summary

This report describes the NBN Data Model. The NBN Data Model is an extensible logical model that incorporates a high level conceptual model of biodiversity and earth science data and individual logical models of subsets, or modules, of the wider system. The model is expressed principally in the form of entity relationship diagrams (ERDs) augmented with descriptions of the principal attributes associated with each entity.

It is not the purpose of the NBN Model to replace existing databases or control how data are managed. Its purpose is to provide the framework within which data collection and management can be interpreted and within which diverse efforts can be harmonised and put to new uses. At a time when it is increasingly expected that the majority of wildlife and earth science data should and will become widely available through electronic means, it is essential that there is a full understanding of the nature of the data and there are the standards to improve the degree of interoperability and improve the chances of reliable data retrieval.

The NBN Conceptual Model includes six key modules;

- *Surveys (field observations and gathering events)*
- *Locations (named places and collecting sites)*
- *Collections (specimens and their management)*
- *People & Organisations (Contacts and addresses)*
- *Sources (Publications and Images)*
- *Thesaurus.*

These top-level modules may incorporate further sub-modules, for instance, Sources includes references and images. At the lower level, modules are comprised of entities that group and describe the relationships of individual items of data.

The modules cross relate to each other, for instance, the thesaurus provides the controlled terminology that would be used in applications featuring the other modules. People and sources can be referenced from any other module. In addition to the main modules, there are a number of Common Entities that can appear in any module. These common entities include measurements, identifications (determinations) and spatial (geographic) references.

The NBN Data Model was first conceived in 1998 during the systems analysis and relational data analysis carried out for the Recorder 2000 database project. The Recorder Project augmented the

NBN Logical Model by developing a number of data syntax and format conventions including NBN 16-character primary keys and Recorder vague date formats. Parallel projects led to the development of the NBN Taxon, Biotope and Administrative Areas dictionaries for terminological control. Subsequent projects including the Lowland Parks and Woodlands Project, Marlin and the Luxembourg National Museum of Natural History's Collection Add-in to Recorder have extended the original data model and produced applications based on the extended model. Part 2 of this report includes physical models for application databases developed from the NBN Model which serve to illustrate how physical and logical models relate and demonstrate the way in which the model can be used and extended.

The NBN data model is compared to and shown to be compatible with, other data models and schemas including the ABCD schema that will become a formal standard and used by both BioCASE and GBIF portals for access to distributed, heterogeneous biodiversity databases world-wide. The final section of the report discusses how the NBN Data Model can be used in the development of applications or as a tool for developing data exchange projects.

The NBN Data Model

Charles Copp

June 2004

Part 1
Description of the Model

1. Scope of the Project

1.1 The Contract

This report was commissioned by the National Biodiversity Network. The contract (part of DEFRA Contract No. CR0241), funded under the Joint Agreement with DEFRA and placed by BioD Service Ltd, was awarded to and has been carried out by Charles Copp of Environmental Information Management.

1.2 Objectives

This report provides an update and explanation of a model for biological and earth sciences field and collection records, referred to as the **NBN Data Model** together with a description of a number of physical data models and applications based upon it.

The NBN Data Model was intended to be one of the under-pinning standards developed to enable reliable data exchange within the National Biodiversity Network[1]. The data model has been used in the development of the NBN Recorder software and to guide other software developers wishing to exchange records with Recorder or send data to the NBN Gateway.

The need to refer to NBN standards including the data model has frequently been referred to in guidelines and documents but there has been no readily available definitive statement of what compromises the full model or explanation of how this can be used in different applications. Copies of the original analysis, from which the NBN data model was derived, (Copp 1997) and a fuller description of the NBN Data Model as used in Recorder (Copp 1998) have been available from the JNCC and NBN websites and a set of notes is currently available (anon.) but this provides little guidance on the relationship between the model and physical implementations.

Significant changes and enhancements have been made to the model since the publication of the first version in 1998 and the purpose of this document is to definitively describe the NBN logical data model, to map it to recently published models and schemas created in Europe and elsewhere and to demonstrate how various physical models (software) relate to it. The benefit of carrying out this work will be a clear definition of the standard, which will provide the means for software developers and data managers to ensure data compatibility and ease of transfer, not only in the UK but in relation to European and world projects such as BioCASE[2] and GBIF[3].

[1] For details of the NBN see www.nbn.org.uk
[2] Biological Collections Access Service for Europe see www.biocase.org

In summary the objectives of this report are to:

- *Document the NBN logical data model and compare it with other models*
- *Describe the physical implementation of the model in Recorder 2002 and Recorder 6*
- *Discuss examples of other implementations and extensions to the model*
- *Provide guidelines on implementing the model and data transfer issues*

2. Introduction - Data Models and the NBN Data Model

2.1 The purpose of a data model

The purpose of a data model is to help us frame our ideas and express our understanding of the relationships between things about which we record information. Models provide a simplified view of the 'real world' and help us focus on those things that are important to our current situation. A good model enables us to place those current interests in a wider context so that we understand them better and can use that understanding to adapt to changing circumstances and extend rather than repeat our efforts. A model can be used to communicate our understanding of complex systems to others and act as a tool for information exchange and the development of standards.

The number of items of information (attributes) recordable about the 'real world' is infinite and so any model can only be an abstraction that represents a particular point of view or set of specific requirements. The model can, however, help us see how many of these detailed attributes fit into more generalised structures (elements) and identify elements that have wider application from those which are highly specific. A feature of a good data model is that it provides a framework that accommodates all the items of information required by our current interest but is both flexible and extensible enough to accommodate new items as they are identified or even completely new but related areas of interest. This is the aim of the NBN Data Model.

2.2 Types of data model

There are many types of data model and modelling techniques. Some models are directed towards defining the structure of knowledge and the meaning of relationships (*ontological models* and *semantic models*), some focus on the way that systems handle information (*data flow models* and *process* or *function models*), some are used to aid database design (*relational data models*, *data structure models*) and some are used to define data processing and user-interface applications (especially *object oriented* and *object relational models*).

Other modelling techniques combine aspects of each of these approaches to provide a conceptual overview of the system and its components, the chief technique used being the entity relationship diagram (ERD). Models that describe concepts and relationships are called *Logical Models* whilst those that define how to store and process the data are called *Physical Models*. Logical Models are normally system independent, which means that they are not constrained by considerations of what storage medium, database or operating system might be used to manage the data whereas physical models must take these in to account as well as performance, management and security

[3] Global Biodiversity Information Facility see www.gbif.org

issues. Physical models can closely resemble a logical model in structure but they may be very different especially where the physical database is optimised for functionality and user interface considerations.

Logical models provide a framework from which real applications may be built. They do not generally cover all aspects of the data or data management that are required to make working systems or to enable exchange of data. All working systems are subject to 'business rules' that place constraints on the type and format of data recorded or the way in which data might be processed (e.g. rules on how you apply and manage levels of record confidentiality). All data management systems also have to have rules or conventions governing data formats, syntax and terminology control. Format, syntax and terminology should be governed by clear conventions and where possible by published standards. Models may refer to such standards but do not define them.

The NBN Model is associated with a number of data format conventions (e.g. NBN Primary Key format) and semi-formal standards (including the taxon dictionary), which were developed under the auspices of the NBN or as part of the Recorder re-development project. These conventions are described in Part 2 of this work.

2.3 Origin of the NBN Data Model

This report describes the NBN Data Model. The NBN Model is a logical model which incorporates a high level conceptual model of biodiversity and earth science data and individual logical models of subsets or modules of the wider system. The model is expressed principally in the form of *entity relationship diagrams* (ERDs) augmented with descriptions of the principal attributes associated with each entity. Part 2 of this report includes physical models for application databases developed from the NBN Model, which serve to illustrate how physical and logical models relate and the way in which the model is extensible.

The NBN model was one of the products developed from a formal *systems analysis* (using SSADM[4] techniques) carried out for the Joint Nature Conservation Committee (JNCC) in early 1997 (Copp 1997 & 1998). The analysis was commissioned as part of a project to redevelop an existing biological recording application called *Recorder*. The analysis involved a fundamental review of the way biological data recorded for different purposes relate to each other and also of the things that people do with biodiversity data and how they were currently managing biological records.

In the report arising from the analysis a new modular, general model was proposed, which allowed for the integration of data derived from different survey types including earth science records (although earth science records were not included the Recorder application). One of the key features of the model lies in the concept of a recording sample, which can link many types of observation to surveys, places, people, physical data and specimens in any combination. The model also allows for records representing repeated sampling and surveillance data or related synchronous samples (e.g. trap lines and transects) that were not well handled in biological recording applications at the time. The model demonstrates the potential for integrating biological and earth science records including both observation and collection records, something of great interest to Local Records Centres (LRCs) and to the developing biodiversity information networks (now including NBN, BioCASE and GBIF).

The Recorder Analysis Model was adopted as the NBN Data Model in 1998 and was used as the reference for developing *Recorder 2000* (released in September 2000), a biological records

[4] Structured Systems Analysis and Design Method

application replacing the earlier *Recorder 3.3* Advanced Revelation application[5]. Recorder 2000 did not implement the full NBN model but the Recorder table structure (physical model) is derived from the suggested physical database structure published in the analysis report. The tables in the Recorder physical model closely match the entities described in the logical model although there are implementation differences, arising from the choice of database manager (Microsoft Access) and meeting performance issues. There are numerous detailed differences in table attributes (fields), for instance, arising from the way that vague dates and spatial references are handled and there are extra attributes for tracking data entry and edits. Information relating to Recorder and support documentation, including a description of the modules used by Recorder and a list of tables and attributes can be found on the NBN website (www.nbn.org.uk).

2.4 On-going development of the NBN Data Model

The modular nature of the NBN Model means that management of parts, such as individual dictionaries (e.g. the taxon, biotope and administrative area dictionaries) can be distributed and also specific applications can be built using only those modules required. Recent work in conjunction with the European funded BioCASE project have facilitated the development of a new Thesaurus Model, developed from the NBN Dictionary models, which allows for the integration of dictionaries in a common database and addresses some of the weaknesses in the original dictionaries. This change is reflected in the current definition of the model presented in this report and which places the separate dictionary modules of the earlier versions of the model in context (n.b. separate dictionaries are still used in Recorder 2002).

Modifications have also come about through the ongoing development of the Recorder Biological Records application and other database initiatives, including the Lowland Parks and Woodlands Information System[6] and the Marlin Project for marine records[7]. The Luxembourg National Museum of Natural History has funded a major extension to the model as part of a project to build a collections management add-in for Recorder that covers both biological and geological collections and extends the survey module to include earth science observations (Dorset Software 2003)

As part of the current contract, a review of the original relational analysis has been carried out and an updated physical model is included in Part 2 of this work. This model includes many tables and attributes (fields) necessary to include earth science data and museum collection management records but due to the need to maintain compatibility with existing versions of Recorder there is a degree of redundancy introduced. For instance, the new model describes a common thesaurus for all term lists and classifications but in the application, it was necessary to keep the separate taxon, biotope and administrative area dictionaries to avoid compromising the existing *Recorder* user interface and report wizard.

The NBN data model appeared slightly before (but overlapped in its development) the publication of a major work describing a comprehensive reference model for biological collections and surveys (Berendsohn et al. 1999) arising from the EU funded BioCISE project[8] and previous work on European floristic databases (Berendsohn 1997). Some features of the NBN model, notably those related to Locations were referred to in the European BioCISE model and the two models can be mapped to each other successfully. The BioCISE Model mixes both a logical model and detailed attribute lists including suggested database attribute names, field types and lengths. The BioCISE model, MDA data standard & SPECTRUM [9]and CIDOC Guidelines for Museum Object

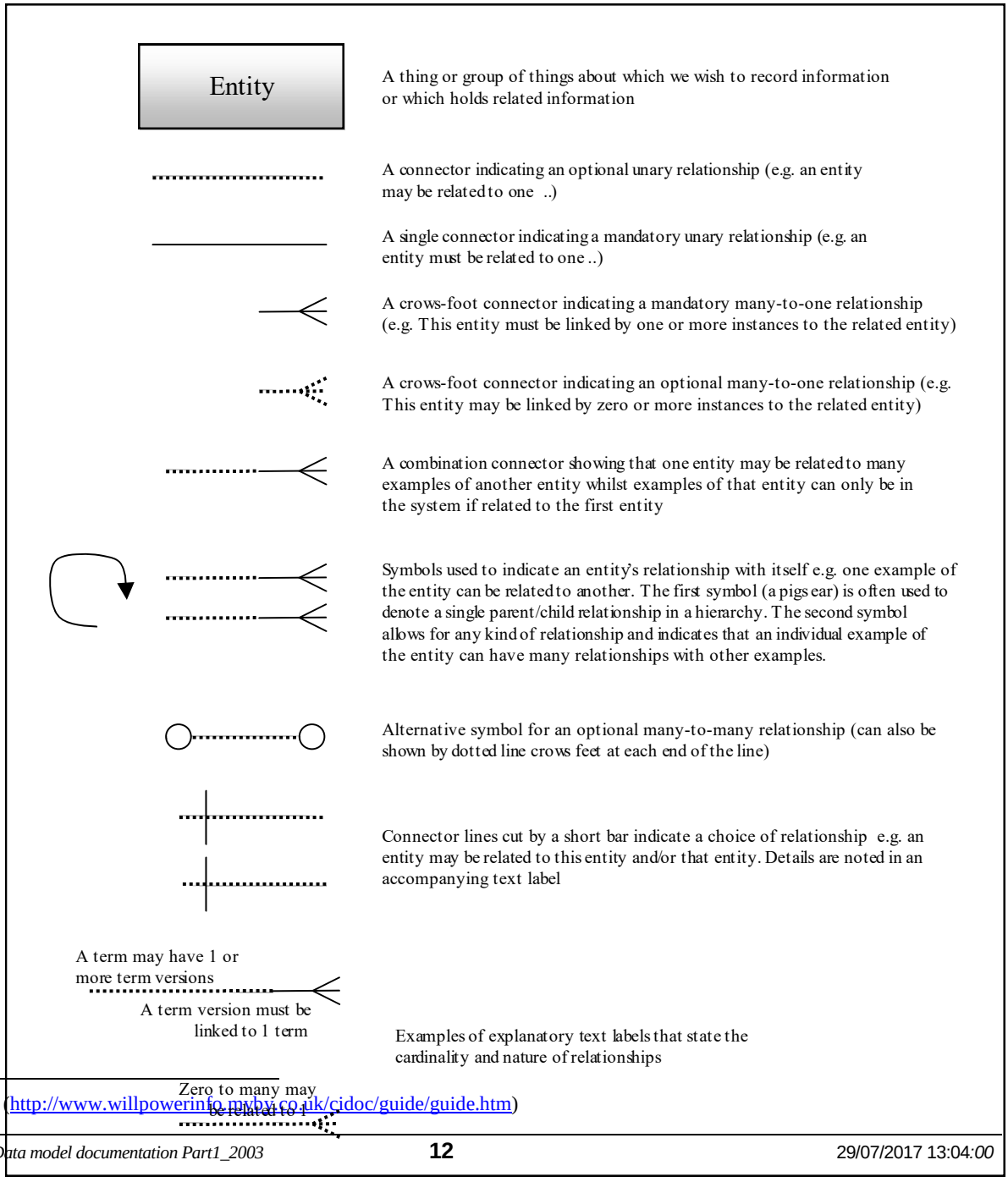
[5] Advanced Revelation is a DOS-based system, no longer supported – see http://www.nfbr.org.uk/html/recorder_3.html for details of current use of Recorder 3.3/3.4
[6] see (http://www.ukwildlife.com/metadata/parks/)
[7] see (http://www.marlin.ac.uk/)
[8] see (http://www.bgbm.fu-berlin.de/biocise/)
[9] see (http://www.mda.org.uk/spectrum.htm)

Information[10] were key references used during the analysis and modelling work for the collections extension to Recorder.

3. Overview of the NBN Data Model

3.1 Explanation of the diagrams

There are many methods used for depicting information systems and data structures; some are linked to design models, such as the relatively simple Entity Relationship Diagrams (ERDs) of the Structured Systems Analysis and Design Method (SSADM) and more the complex objects used in Unified Modelling Language (UML) diagrams.



[10] see (http://www.willpowerinfo.myby.co.uk/cidoc/guide/guide.htm)

Figure 1: Explanation of symbols used in diagrams in this report

Diagrams differ in the shapes they use for entities and objects, the shape and endings of connectors used to indicate relationships and the degree of label text allowed. Choice of shapes, text and connectors may be dictated by the diagramming software used and especially with CASE[11] tools. This range of diagramming methodologies and also codes used to signify options and relationships can make some systems very hard to understand for non-specialists. The diagrams in this report have been kept deliberately simple and are derived from basic SSADM ERD models. The intention is to relay meaning as clearly as possible without recourse to obscure codes. The simple diagramming model used lends itself to describing relationships from the level of the conceptual model of the whole system down to detailed table relationships in example physical models. The symbols and connectors used in diagrams in this report are shown in *Figure 1*.

3.2 The Scope of Biodiversity and Earth Science Records

The collection and management of information related to biodiversity and the earth sciences is complex. Typically the design of databases and data models has tended to focus on narrow aspects of the whole picture, to meet the requirements of individual projects. For instance many naturalists keep a list of sites, with species names and date seen, whilst detailed floristic surveys might be stored as a set of spreadsheets with sample numbers as column heads and species as rows with the cells holding dominance counts. Another database might store features of sites and record aims, threats and damage along with periodic monitoring statements. These different approaches might suggest that the data are very different and not readily manageable or accessible through a single system. The NBN data model demonstrates that the examples described can be interpreted as 'views' of a deeper data structure that can be used to map any form of environmental data.

The advantage of having a general model include:

- *It provides a template for mapping, merging or accessing data from heterogeneous sources*
- *It provides a guideline for the development of new data management software*
- *It can be used to highlight potential deficiencies in data sets, for instance missing metadata in datasets that are being made available through web portals or missing data required for successful transfer to another database application.*
- *It can be used in the planning and design of new surveys e.g. to ensure that data collected have a degree of re-usability, possibly by highlighting some extra data that could be collected at the same time as that needed for the primary purpose.*

It is not the purpose of the NBN Model to replace existing databases or control how data are managed. Its purpose is to provide the framework within which data collection and management can be interpreted and within which diverse efforts can be harmonised and put to new uses. At a time when it is increasingly expected that the majority of wildlife and earth science data should and will become widely available through electronic means, it is essential that there is a full understanding of the nature of the data and there are the standards to improve the degree of interoperability and improve the chances of reliable data retrieval.

Figure 2 is a schematic diagram of the relationship of the different kinds of data that fall within the scope of the NBN Data Model. It is also a model of how we see and understand the world.

[11] Computer Aided Software Engineering

Figure 2 is not an illustration of the NBN data model, its purpose is to show the relationships between the concepts from which the data model has been constructed.

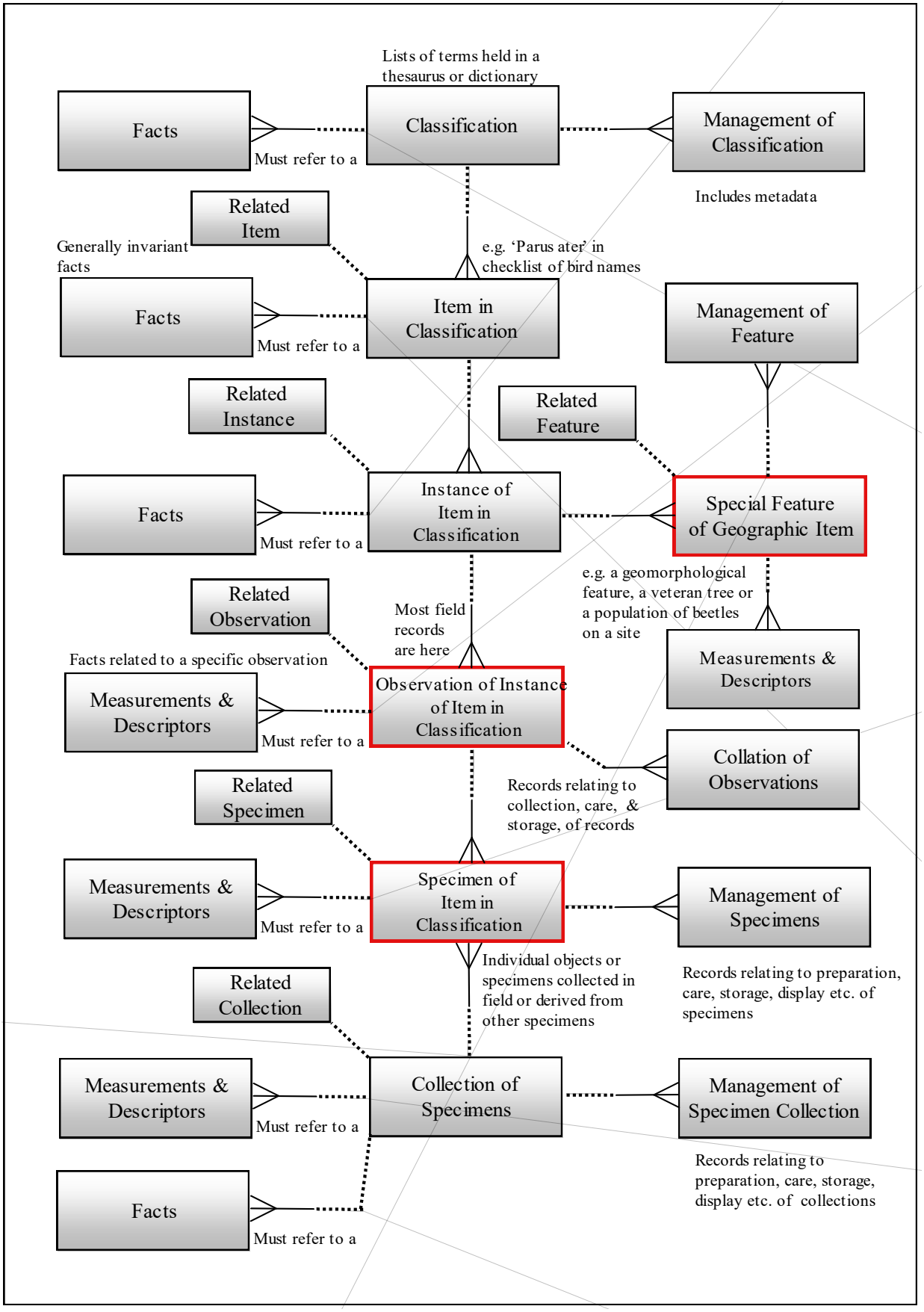


Figure 2: The relationships between the classification of things of interest, to observations, specimens, facts and management covered by the extended NBN Data Model

At the top of the diagram (*Figure 2*) there is a **Classification** entity, which represents the terminology through which we describe the world and our experience. In the context of biodiversity data, this is the store of plant and animal names, minerals, rocks, place names, habitat types, collecting methodologies and all other terms related to recording and collecting. These terms are associated in lists which belong to subject areas or domains. Each list has facts about the list and its management associated with it. In NBN terms, this is the area covered, for instance, by the Taxon Dictionary and the Biotope (Habitats) Dictionary.

Within each list are the individual terms (e.g. species names) which we use to describe what we have seen and where we have seen it. Each term is an **Item in Classification** and may also have facts related to it (e.g. Gastropods have a radula, Arctic Foxes have a white phase). The terms can be related to each other in many ways (e.g. broader term, narrower term, predator of, parasite of etc.).

An **Instance** of **Item in Classification** is a real world example of the term; where the term might be 'Oyster Fungus', the instance might be a particular example of Oyster Fungus[12]. The instance can have its own facts attached to it and be related to other instances (e.g. the Beech tree it was growing on). In the special case of geographic locations, the instance might be a named site and sites may include features (**Special features**) about which we might record information, measurements and descriptors. A special feature of the site could be the reason why the site has been given statutory protection (e.g. the presence of a population of a rare species) or it could be a feature that we return to often such as a particularly fossiliferous horizon in a quarry face. In *Figure 2*, this entity is outlined in red as it illustrates one of the main types of recording that the NBN model needs to address.

The central entity in *Figure 2*, also outlined in red, is the **Observation of instance of item in classification**. This represents a fixed point in time observation about a named item of interest such as an example of a species, at a grid reference, on a specific date. This is the commonest sort of biodiversity record and the one most recording applications handle. Observations may have many measurements (e.g. count or abundance) and descriptors associated with them and can be related to other observations (e.g. all things seen at the same time and place). These records of observations are commonly collated together and will have management data associated with them. This latter data is the sort of information needed by local record centres to collate and manage records from different recorders and also includes ownership and copyright data needed to share records through the NBN portal.

The observation may be linked to the actual collection of specimens (**Specimen of item in classification**, outlined in red) which although not covered by most biological recording applications is a major concern of museums and is vital where voucher specimens and taxonomically important specimens are concerned. Specimens have their own measurements and descriptors associated with them and will be associated with data describing their preparation, conservation, storage and management.

Specimens are most often associated together into collections which have a 'group life' in that they may be managed, sold, stored and described as a unit. In fact, specimens, collections and even the furniture that they are stored in, share so many common information attributes that they

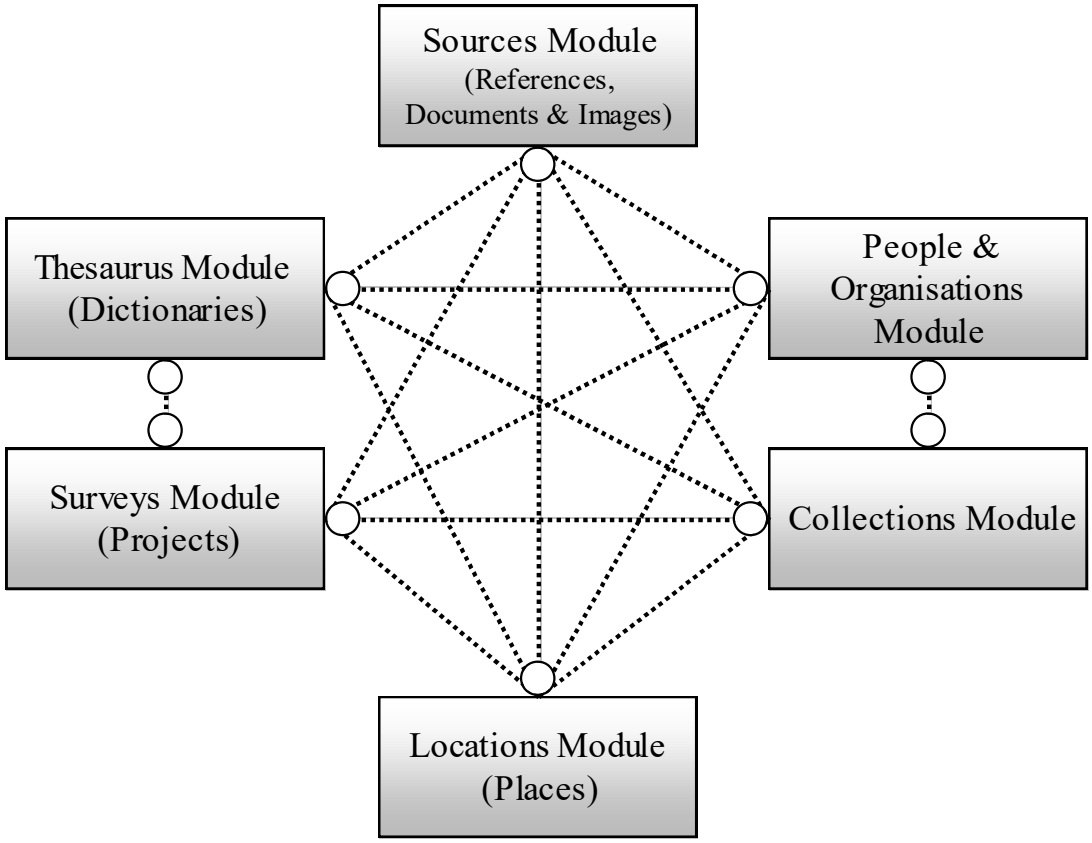
[12] Most lists of terms are 'generic', in that they are lists of group terms, which can be represented by many instances (e.g. there can be any number of Black Garden Ants (*Lasius niger*)) but some lists are specific in that the 'lowest level' terms are already unique instances, for instance a list of ancient woodlands, National Nature Reserves or local authorities. In this case the Item in Classification and the Instance appear to be the same thing – this is not a contradiction, all 'concepts' belong to some form of hierarchy where the highest terms are essentially abstract and the lowest terms are unique examples but we generally have a 'cut-off point' before the lowest level.

can be regarded as instances of a common entity, which the BioCISE[13] project dubbed a 'Unit'. They are kept separate here for clarity.

Thus, the information that we commonly regard as 'biological records' falls within a spectrum of information from our overall classification of the world at one end to collated collections of actual examples of classified items at the other. In the middle there may be field observations without physical voucher material, there may be field records linked to specimens, specimens without recorded field data and even features (such as a population of beetles) about which we maintain a longer term interest. This diagrammatic view enables us to analyse our data into related types which although managed differently and therefore indicate a potentially modular structure, fit together and indicate that many different applications could be built from the modular parts, each suited to its particular use yet compatible with all others. This is the starting point for the development of the NBN data model.

4. The NBN Conceptual Model

The NBN Data Model was developed for pragmatic purposes. Its primary function was to provide the conceptual and logical framework for the definition of the new *Recorder* biological recording database. In the process of development it became clear that the model offered a systematised description of biodiversity data that could be used as a basis for mapping databases to each other and the prospect of combining data from heterogeneous sources.



[13] Biological Collections Information System for Europe, a European funded collaborative project which was the forerunner for BioCASE.

Figure 3: The scope of the NBN Data Model (Note that each module may include elements that reference any other module.)

4.1 Logical Modules

The NBN Data Model is modular. The great range of information that might be recorded in relation to biodiversity and geodiversity is systematised into separately modelled units or **modules** (*Figure 3*). These units may be constructed of **sub-modules** which may, themselves, have further modular sub-divisions. The modules and sub-modules represent groups of **entities** that are linked to each other through a variety of relationships. Entities are simply things about which we record information or which provided information to the system. In the NBN Model we are recording facts about places (Locations), Surveys (Field Records), Collections (Specimens), People, References and also using a thesaurus (or dictionaries) to supply controlled terminology and supporting information. This is not an exhaustive list of the potential extent of the model. The modules illustrated in *Figure 3* could be added to by many others including Business Management, Financial Control, Manufacture & Production, Sales & Marketing, Publishing and Exhibition Modules. These modules might be appropriate for extensions relating to managing museums and LRCs but do not form part of the core model for biodiversity and geodiversity data described as the NBN Data Model. There is a separate metadata model used for describing the content, ownership and access for collections of data and specimens[14].

The conceptual diagram in *Figure 3* is a high-level entity relational diagram describing the wider interactions in the system. For instance, elements or entities within the Surveys Module and other modules might refer to people or organisations whose details are maintained in the People & Organisations (or Contacts) Module. This means that where a person or organisation is referred to, only a pointer to the Contacts Module is required to link to not only the name but to any personal, biographical or communication information that the system stores associated with that name. Note that this is a what is called a **logical model,** in that it describes the extent and relationships of things we are interested in; it does not require that all databases (or **physical models**) include every possible data attribute that might be described within a module. For example, the Collections Module is very large and includes several sub-modules but for some applications the only part of the module represented in a database might be one or two attributes such as 'specimen number' and 'storage location'. This was the approach taken in Recorder 2000 and 2002, an application designed primarily for field records, although a very extensive collections management add-in is available for the soon to be released Recorder 6.

5. Common Entities

In addition to the Modules and their associated sub-modules there is a further element type in the model, referred to here as '**Common Entities'**. In an XML schema[15] they would be referred to as *Named Complex Types*, that is, a group of elements and attributes that are linked together and can be included in to any other Type. In the MDA data standard (MDA 1994) they correspond to 'Common Elements'. In the NBN data model Common Entities represent entities that cannot exist in isolation (unlike the modules) but can appear in any module where needed, providing a standardised method for handling similar data such as measurements and identifications.

Common Entities cover parametric (measurement), descriptive, identification and attribution (attributing an item to a person, place, 'school', or time period) data that can be related to other entities. Grid-based spatial references such as latitude & longitude, UTM and O.S. National Grid

[14] The NBN Metadata Standard is available from the NBN website: http://www.nbn.org.uk
[15] See http://www.w3.org/TR/2001/REC-xmlschema-1-20010502/

can be regarded as belonging to the class of Common Entities because they are a form of spatial measurement.

One example of a Common Entity is 'Measurement". It is a common experience when looking at physical database designs to find that many of the table attributes (fields) represent specific measurements, for instance, length of wing, weight, thickness of shell, pH of soil, mineral hardness and many more, without limit.

This use of individual fields for specific measurements along with others for specific descriptive terms (e.g. eye colour, plumage phase, moult state etc.) is a major reason why databases can appear to be very different even when dealing with similar types of field or collection record. In the NBN model all of these specific measurements and descriptors can be mapped to a single measurement entity that provides the options for use of controlled terminology (e.g. linked to measurement units, methods, accuracy etc.) and for providing a standardised structure that makes data merging and retrieval much easier. In the NBN model, Common Entities can be extended to cover all of the possible ways in which the associated data may be recorded, for instance, geo-spatial references may be sub-typed to cover several different referencing systems.

It is not usual, in logical models, to be specific about the recording format of data items (attributes) but most physical database models prescribe the syntax and format that should be used. A growing number of collation databases[16], however, have to allow for different formats of items such as measurements, taxonomic names (atomised, non-atomised, zoological, botanical etc) and geo-spatial references. The development of XML schemas as an aid to data transfer in electronic systems (such as linking partner databases to a portal over the Internet) has had to recognise the need to harmonise different ways of recording data and this is represented in the way in which they may use 'Complex Types' to represent these concepts.

The existence of the Common Entities such as *Measurement* in the NBN model does not limit database designers to their implementation in physical tables. Physical databases might still have individual fields such as 'sward height' or 'altitude' but the model provides a means of mapping the data between applications. A database application might even show such detailed fields on a screen form but use a table derived from the measurement type entity to store data in the actual database. *Recorder 2000/2002* and *Recorder 6* both use a standard structure to record measurement data relating to sites, taxon observations and habitat occurrences but not a common physical table. There are thus many ways in which the model can be implemented, its value lies in enabling the mapping of data between applications and serving as a reference to what, for instance, a measurement implies (e.g. its method, accuracy, duration or scale; which may only be assumed in some applications).

The following sections examine the potential structure and content of Common Entities and how this might translate into actual database structures as represented by versions of the *Recorder* application.

One of the most important recent developments in analysing biodiversity and collection data is the collaborative development of an XML schema for the transmission of data related to individual or groups of observations or specimens (referred to as Units, as defined in the BioCISE model). The schema, referred too as the ABCD Schema[17], has been developed under the auspices of the Taxonomic Database Working Group (TDWG), the Committee on Data for Science and Technology (CODATA) and the BioCASE Project (Biological Collections Access in Europe). The ABCD Schema makes extensive use of Named Complex Types and provides a valuable way of visualising the Common Entities used in the NBN Data Model and for understanding the way they are used in database applications such as Recorder.

[16] databases designed to be able to bring together data from other databases, not necessarily of the same structure. E.g. the use of the *Recorder* database in Local Records Centres.

[17] see http://www.bgbm.org/TDWG/CODATA/Schema/

5.1 Measurements Common Entity

A measurement common entity is used in *Recorder* (versions 2000, 2002 & 6), in the form of separate tables including **Taxon_Occurrence_Data**, **Biotope_Occurrence_Data** and **Location_Data**. *Table 1* and *Table 2* illustrate the use of a measurement type (Common Entity) in the *Recorder* application. Table 1 is the measurement type used for taxon_occurrence_data and is part of the original Recorder 2000 development. It does not allow for the recording of range values (lower and upper readings) or duration for timed measurements (e.g. number of birds counted in one hour). This has been addressed in the Collections and earth science extensions to Recorder 6 where the measurement type is brought into compatibility with the ABCD Schema MeasurementAtomized Type (as in *Figure 4*).

COLUMN NAME	DESCRIPTION	TYPE	SIZE
TAXON_OCCURRENCE_DATA_KEY	Unique identifier for the TAXON_OCCURRENCE_DATA table. Required.	char	16
TAXON_OCCURRENCE_KEY	Identifies the occurrence which the measurement is associated with. Foreign key to the TAXON_OCCURRENCE_KEY field in the TAXON_OCCURRENCE table. Required.	char	16
DATA	Actual data value, stored as free text. Required.	varchar	10
QUALIFIER	Qualifier for the data. For example, may determine that a size measurement was taken of a wing's length. Required.	varchar	20
ACCURACY	Text detailing the accuracy of the measurement. Not Required.	varchar	10
MEASUREMENT_UNIT_KEY	Identifies the measurement unit, and therefore what is being measured, for the occurrence. Foreign key to the MEASUREMENT_UNIT_KEY field in the MEASUREMENT_UNIT table. Required.	char	16

Table 1: Measurements table (relating field observations of species) in the Recorder 2002 & 6

The table definition illustrated in *Table 2* (below) is from the collections management add-in to *Recorder 6*. The table structure differs in several ways from that used in the original version of *Recorder*. Note that method, parameter and measurement unit are all referenced by keys to lists managed in the Thesaurus Module and that measurements may now have duration and range values. A further change is that there is now a descriptor flag (**Is_Descriptor**), which enables the entity to be used for non-parametric descriptive terms; For instance, shell colour or degree of

openness of a shell umbilicus could be recorded alongside scalar measurements such as height and width.

COLUMN NAME	DESCRIPTION	TYPE	SIZE
Collection_Unit_Data_Key	Unique identifier for the Collection_Unit_Data table	char Not Null	16
Collection_Unit_Key	Collection unit the data is associated with.	char Not Null	16
Applies_To	Qualifies what the measurement applies to. For example, a measurement of temperature may be a surface, air or sub-surface measurement.	varchar Not Null	50
Method_Concept_Key	Concept from the Measurement Methods concept group that identifies the method used to make the measurement.	char Null	16
Duration	Free text duration of the measurement period.	varchar Null	50
Accuracy	Free text accuracy of the measurement.	varchar Null	50
Parameter_Concept_Key	Concept from the Measurement Parameters concept group that identifies the parameter that is being measured (e.g. length, height, altitude, temperature).	char Not Null	16
Unit_Concept_Key	Concept from the Measurement Units concept group that identifies the measurement unit.	char Null	16
Lower_Value	Lower value of the measurement range, or measurement value if a single value is specified.	varchar Not Null	50
Upper_Value	Upper value of the measurement range, or null if a single value is specified.	varchar Null	50
Is_Descriptor	Flag indicating if the record is for a descriptor (1) or a data measurement (0).	bit Not Null	

Table 2: Measurements table (relating to specimens, collections or storage units and areas – collectively called collection units) in the Recorder 6 Collections add-in

The use of a measurement type has been adopted in the ABCD Schema, currently being developed as a joint project by TDWG, CODATA and BioCASE, and which will be one of the two principal

unit-level[18] data transfer standards supported by GBIF[19]. The ABCD schema is an XML structure derived from the analysis of a number of data models and standards including the NBN model.

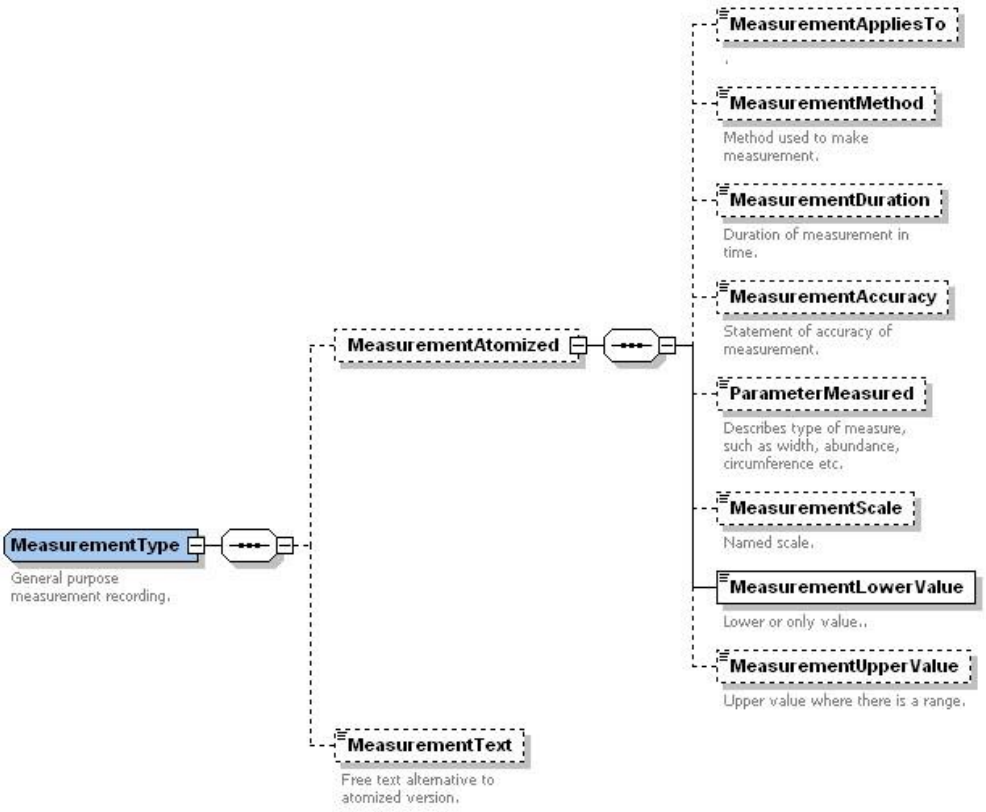


Figure 4: Structure of the MeasurementType Element in the ABCD Schema

The ABCD element **MeasurementType,** illustrated in *Figure 4*, demonstrates how a logical entity can represent different approaches to the format and syntax of data in actual databases. In *Figure 4* a measurement can be represented by a simple text description (**MeasurementText**), which might even include the subject of the measurement (e.g. "bill length 20 mm") or it can be atomised (**MeasurementAtomized**) into a number of related elements ('database fields'). The elements in the diagram with dotted outlines are regarded as optional. The fully atomized measurement allows the recording of what thing the measurement applies to, what parameter is being measured, the method of measurement used, its duration and accuracy, the scale (units) and actual value or lower and higher values.

There are a number of elements missing from the model in *Figure 4* and *Tables 1* and *2*. Each measurement should be linked to the person or persons doing the measurement together with the date and time the measurement was taken. The reason that these elements do not appear within the measurement definition is that they are normally taken from the context within which the measurement is recorded, for instance in the basic form of Recorder, measurements are linked to occurrences and samples, which have recorders, dates and times linked to them. With specimens, measurements might be taken at any time by any person and so the measurement should be linked to an event.

[18] Data at the level of a single unit e.g. observation, specimen, named collection – instance data not metadata.
[19] See http://www.gbif.org/links/standards

5.2 Identification (or Determination) Common Entity

The NBN Identification Common Entity is best characterised by the way in which it has been used in 'core' *Recorder* and the new Collections Management Add-in (see *Table 3*). In The NBN Data Model, the central premise relating to identifications is that once a determination or identification is applied to an observation or specimen, it cannot (normally) be changed or removed. Any change of name, either through disagreement over the correctness of the identification or for simple nomenclatural update must be added to the identification history of the record or specimen and may be set as the preferred identification. This constraint is intended to protect the identification history such that the original intentions of the observer, collector and later determiners can all be traced and, if necessary, later name changes accurately undone (For instance in disputes related to the lumping and splitting of taxa or ability of the determiner).

Table 3: Data elements in the NBN Identification (or Determination) Common Entity

The Identification Common Element can, theoretically, be applied to any type of identification in any domain with perhaps, domain specific controlled terminology for elements such as identification method. In the core *Recorder* application, identifications are restricted to taxa and biotopes and the data are stored in different tables. The extensions to Recorder associated with the Collections Management Add-in include a new generalised determination entity that can be associated with any type of occurrence or specimen, including rocks, fossils, minerals, soils and stratigraphic horizons.

The structure of the **Taxon_Determination** table in *Recorder 2002* is shown in *Table 4*. Any record in the Taxon_Determination table refers to a 'Taxon Occurrence' (i.e. a record of an observation) and each Taxon Occurrence record can be determined many times.. Each determination is detailed in this table. The current preferred determination is indicated by the PREFERRED flag. The fields relating to 'Vague Dates' refer to the Recorder-specific convention

Determiner	The name of the person or persons responsible for making the identification.
Determination Date	The date on which the identification was mad. This can be a vague or imprecise date
Determination Type	What sort of identification was made (e.g. original observer's identification, validation check by county recorder, nomenclatural determination during taxonomic revision)
Determiner Role	The role of the determiner (e.g. original recorder)
Determination Method	Method of determination, which may include reference to voucher specimens, chemical or genetic tests etc.
Determination Reference	Publications or reference works used to make the identification
Identification	The actual name applied to the record or specimen by the determiner. The name element can be a simple string, pointer to a thesaurus/dictionary entry or atomised name. In Recorder all names are held as pointers to a dictionary entry. The NBN convention is to refer to an entry in the UK (Recorder) Taxon Dictionary, the master copy of which is managed by the Natural History Museum, London.
Nomenclatural Status	If the determination is part of a taxonomic revision or review, any nomenclatural status proposed for the specimen by the determiner e.g. holotype, syntype, paratype, figured specimen etc.
Confidence Level	Measure of the confidence in the determination (e.g. high)
Inferred Flag	Marker that may be applied to any field indicating that the information is inferred and not original
Preferred Flag	Marker for the identification that is to be used as the preferred term in any lists or output
Other Sources	Links to any other source or corroborating material (can include papers, books, documents, photographs, film clips and other media).

for depicting dates, approximate dates and date ranges. The Recorder application parses the contents of a 'vague date' field on a data entry form and records a date type for it together with a calculated beginning and ending date. Vague dates are further explained in Section in Part 2.

In *Table 4*, where the column name (database field) refers to a foreign key (e.g. Determiner is **NAME_KEY** and the identification (taxon name) is referred to through **LIST_ITEM_KEY**), this is a physical way of implementing the use of complex types like those (marked with '+') in the

ABCD diagrams shown in *Figures 5, 6 & 7*. In the NBN Model the determiner's role is recorded in order to augment the information given by the different types of identification; for instance a validation type of identification might have been carried out by a determiner in their role as a Vice-county Recorder.

Note that in the original Recorder application the focus of the application was on field records and so the 'core' *Recorder* determination records do not allow for the determination of nomenclatural status although this is covered in the Collection Management Add-in.

COLUMN NAME	DESCRIPTION	TYPE	SIZE
TAXON_DETERMINATION_KEY	Unique identifier for the TAXON_DETERMINATION table. Required. Primary Key.	char	16
PREFERRED	Flag indicating the most recent (and therefore 'preferred') determination of each taxon occurrence. Required.	bit	
VAGUE_ DATE_START	Specifies the vague date for a taxon determination. Earliest possible date for a range specified by a vague date. Not Required.	integer	
VAGUE_ DATE_END	Specifies the vague date for a taxon determination. Latest possible date for a range specified by a vague date. Not Required.	integer	
VAGUE_ DATE_TYPE	Specifies the vague date for a taxon determination. Indicates the type of vague date. Not Required.	char	2
COMMENT	Comment for the taxon determination. Not Required.	text	
TAXON_OCCURRENCE_KEY	Identifies the taxon occurrence which is determined by the record. Foreign key to the TAXON_OCCURRENCE_KEY field in the TAXON_OCCURRENCE table. Required.	char	16
DETERMINER	Indicates the person who performed the determination. Foreign key to the NAME_KEY field in the NAME table. Required.	char	16
DETERMINATION_TYPE_KEY	Indicates the type of determination, for example Original, Confirmation. Foreign key to the DETERMINATION_TYPE_KEY field in the DETERMINATION_TYPE table. Required.	char	16

TAXON_LIST_ITEM_KEY	Identifies the taxon and checklist which was determined. Foreign key to the TAXON_LIST_ITEM_KEY field in the TAXON_LIST_ITEM table. Required.	char	16
DETERMINER_ROLE_KEY	Indicates the role of the person performing the determination. For example, Vice County Recorder. Foreign key to the DETERMINER_ROLE_KEY field in the DETERMINER_ROLE table. Required.	char	16

Table 4: Fields used in the Taxon_Determination Table in the core version of Recorder 2002 & 6

Table 5 shows the slightly different Determination table used in the *Recorder 6, Collections Management Add-in*. These determinations refer to a new Thesaurus rather than the separate species and biotope dictionaries used in 'core' Recorder. They also differ from the original Recorder determinations in that they can refer to specimens and may therefore be dealing with specimens that have a nomenclatural status. In this case the **Concept_Key** in *Table 5* is equivalent to the **TAXON_LIST_ITEM_KEY** in *Table 4*. The table also includes a number of extra fields allowing the determination of a nomenclatural status for a specimen (e.g. holotype, syntype etc.) and the confidence with which a determination is given. A further attribute allows a determination and determiner to be flagged as inferred; this is an important requirement with museum specimens.

COLUMN NAME	DESCRIPTION	TYPE	SIZE
Determination_Key	Unique identifier for the Determination table	char Not Null	16
Concept_Key	Identifies the actual thesaurus concept that has been determined.	char Not Null	16
Occurrence_Key	Occurrence the determination is linked to if the determination was made against a field observation.	char Null	16
Specimen_Collection_Unit_Key	Specimen the determination was made against, if Used_Specimen=1.	char Null	16
Determination_Type_Key	Identifies the type of determination from the Determination_Type table.	char Not Null	16
Nomenclatural_Status_Concept_Key	Linked to the Nomencaltural Statuses concept group, identifies the status of the determination (e.g. figured or typed).	char Null	16

COLUMN NAME	DESCRIPTION	TYPE	SIZE
Confidence	Identifies the confidence of the determination. Possible values are : 0 = Uncertain confidence 1 = No confidence 2 = Low confidence 3 = Medium confidence 4 = High confidence	tinyint Not Null default (0)	
Determiner_Name_Key	Identifies the person who performed the determination from the Individual table.	char Not Null	16
Inferred_Determiner	Inferred data flag for the Determiner_Name_Key field.	tinyint Not Null	
Determiner_Role_Key	Identifies the role of the determiner in making the determination from the Determination_Role table.	char Not Null	16
Vague_Date_Start	Vague date start field for the date of determination (integer, days since 30/12/1899).	int Null	
Vague_Date_End	Vague date end field for the date of determination (integer, days since 30/12/1899).	int Null	
Vague_Date_Type	Vague d determi		
Used_Specimen	Flag ind was mad the field		
Preferred	Flag ind the pref preferre		
Method	Free tex determi		
Notes	Free tex determi		

Table 5: Fields used in the Taxon Det
R

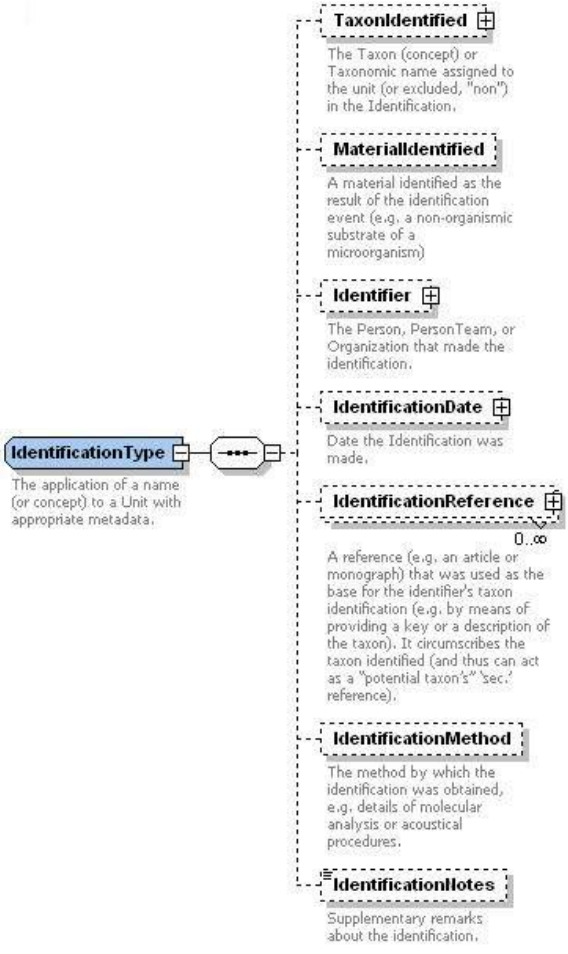


Figure 5 shows the content model for the AB
NBN Data Model. The ABCD Schema also u

information that may relate to the application of an identification (giving a name to) things observed or collected. The Type is made up of several other complex types, each of which may be variably atomised.

The first element represents the actual name given in the identification if it is a taxon name otherwise the identification is for a material substance. Note that the ABCD schema does not yet relate to earth sciences data other than fossil taxa and therefore the content model does not include mineral and rock identifications. Each identification is associated with an identifier, the date the identification was made, any published references used to aid the identification, the method used and any further notes. Elements marked with a '+' are broken down into further elements or choices.

Figure 5: Content of the ABCD Schema Identification Type. Elements marked with a + can be further de-composed, into further elements and complex types.

The **TaxonIdentified** element is shown *Figure 6.* The content model includes an element to record a higher taxon (e.g. Family or above) so that the general group to which the identified unit belongs can be quickly recognised. The actual name given by the person(s) making the identification is represented by either a common name (**InformalNameString**) or a classified 'scientific' name (**ScientificName**) and both may be applied. The ScientificName element is made up of a choice of further elements as illustrated by *Figure 7.*

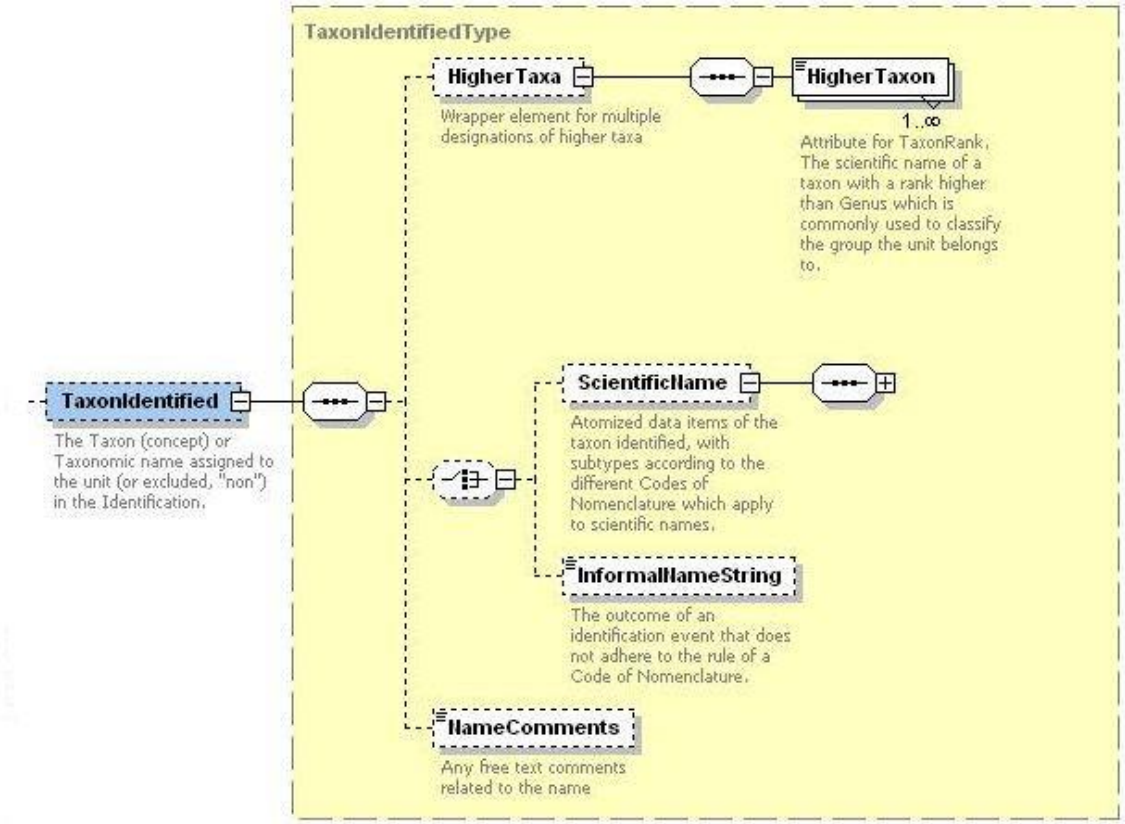


Figure 6: ABCD Schema, TaxonIdentified Element

The **ScientificName** element is made up of a number of elements that may be used in databases or act as a means for mapping different ways of recording names to one another. The **FullScientificNameString** represents a whole taxon name with author and date (e.g. *Cymbiola rutila* (Broderip 1826)) whereas **ScientificNameString** represent the name without the 'author string'. For the purposes of the ABCD schema, when used for transferring and merging data, these elements may be copied directly from a database or created by concatenating the name parts, if they are stored separately. Within the ScientificName element there is also a further complex type (**NameAtomized**) that allows formal taxonomic terms to be broken down into separate fields (e.g. genus, sub-genus, species, sub-species etc.) These choices of atomised element are listed separately for the different nomenclatural codes as there are significant differences in the formatting and allowed content of names between codes.

This subtlety does not represent part of the NBN Data Model Identification Common Entity, which simply refers to a name and expects that name to conform to a declared standard dictionary (e.g. the NBN Taxon Dictionary). In this instance, the ABCD schema sits above the NBN Data Model as a unifying tool for data recorded according to different models and databases.

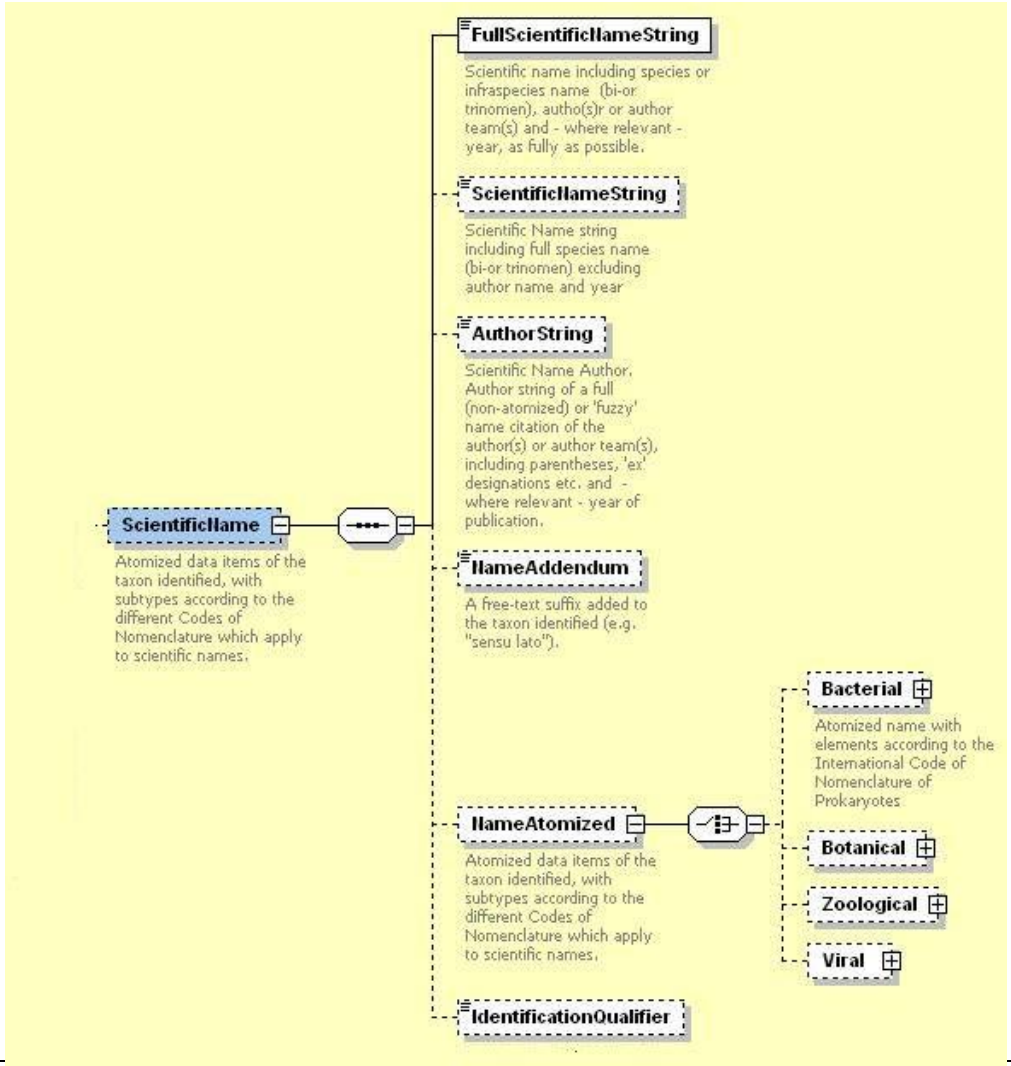


Figure 7: ABCD Schema, content model for ScientificName element. Note that the NameAtomized element can be further decomposed into elements representing different codes of nomenclature such as the botanical and zoological codes and these elements will contain data elements (fields) relevant to the particular code.

5.3 Spatial Reference (or Geospatial Coordinates)

The NBN Data Model includes many references to spatial data, for instance, in its Survey and Location Modules. In the model, the citing of spatial coordinates (e.g. grid references) uses a common entity, as the format is used in several places. The logical model includes entities as placeholders for geospatial elements including points and squares within spatial reference systems, vector polygons and raster map images but the formats of the data are not prescribed. It is regarded as the responsibility of application developers to ensure that data are capable of transfer in standard formats regardless of how they choose to store them.

The recording and transfer of spatial data is a highly complex issue which has been addressed by numerous groups around the world, with the consequent development of many geographical information metadata standards and data transfer standards (e.g. the UK National Transfer Standard[20]). In the UK these issues are addressed by the Association for Geographic Information[21] and the National Geospatial Data Framework[22]. The NBN Metadata Standard uses guidelines from the NGDF to ensure compatibility of data recorded describing datasets indexed on the NBN Gateway. The approach used in the NBN data model is a deliberately simple; only point and square coordinates have defined structures, all other digital map elements including line vectors, polygons and raster maps are regarded as external to the system and only references to them managed within the applications built using the model (e.g. *Recorder*).

The format for the storage of geospatial coordinates in *Recorder* is shown in *Table 6*. All spatial references are converted from the input format given by the user (e.g. OSGB) and stored

[20] National Transfer Format (NTF): British Standard BS 7567, defines an exchange format for the transfer of vector data. NTF is the main form used by the UK Ordinance Survey.
[21] AGI see http://www.agi.org.uk
[22] NGDF see http://www.ngdf.org.uk Link may be broken

internally as decimal latitude and longitude in degrees[23] and the SPATIAL_REF_SYSTEM field is used to lookup the output conversion needed for reports and display. The standard spatial reference systems provided with *Recorder 2000* are:

- *OS National Grid for Great Britain (uses 2 letter 100Km codes)*
- *OS National Grid for Ireland (uses 1 letter 100Km codes)*
- *Decimal Latitude and Longitude coordinates*
- *Universal Transverse Mercator coordinates*

Other spatial reference systems can be added as 'Add-Ins'. Only one spatial reference system can be used in any working session.

In Recorder Spatial References are stored under five fields:

COLUMN NAME	DESCRIPTION	TYPE	SIZE
SPATIAL_REF	Spatial reference for the 'square' as recorded by the user. Required. E.g. ST 423 718	varchar	20
SPATIAL_REF_SYSTEM	Original system used to record the spatial reference. Required. e.g. OSGB	varchar	4
SPATIAL_REF_QUALIFIER	Qualifier for the spatial reference of the grid square. Defaults to "Entered" but can be "Map" if retrieved from Recorder's internal mapping function. Required.	varchar	20
LAT	Decimal Latitude coordinate, stored as a double precision floating point value. Required. e.g. . 51.4417602287966	float	
LONG	Decimal Longitude coordinate, stored as a double precision floating point value.	float	

[23] The American Spatial Data Transfer Standard, part of the US National spatial Data Infrastructure (see: http://mcmcweb.er.usgs.gov/sdts/) provides the following definition of decimal latitude and longitude:
Latitude and longitude are ellipsoidal coordinate representations that show locations on the surface of the earth using the earth's equator and the prime meridian (Greenwich, England) as the respective latitudinal and longitudinal origins.
Latitude and longitude are angular quantities, and according to the standard, should be expressed as decimal fractions of degrees.
Degrees of latitude, according to the standard, should be represented by a two-digit decimal number ranging from 0 through 90.
Degrees of longitude, according to the standard, should be represented by a three-digit decimal number ranging from 0 through 180.
When a decimal fraction of a degree is specified it, according to the standard, should be separated from the whole number of degrees by a decimal point.
Latitude north of the equator, according to the standard, should be specified by a plus sign (+) or by the absence of a minus sign (-), preceding the two digits designating degrees. A point on the equator, according to the standard, should be assigned to the northern hemisphere. Latitude south of the equator shall be designated by a minus sign (-) preceding the two digits designating degrees.
Longitudes east of the prime meridian, according to the standard, should be specified by a plus sign (+) or by the absence of a minus sign (-), preceding the three digits designating degrees of longitude. Longitudes west of the meridian, according to the standard, should be designated by a minus sign preceding the three digits designating degrees. A point on the prime meridian, according to the standard, should be assigned to the eastern hemisphere. A point on the 180th meridian shall be assigned to the western hemisphere.
Any spatial address with a latitude of +90 or -90 degrees specifies the location of the North or South pole, respectively. The longitude component may have any legal value.

	Required. E.g. –2.83028793601795		

Table 6: Table structure for spatial references used in Recorder

In the ABCD Schema, Site coordinates are restricted as an included complex Type within the Gathering Site element (see *Figure 8*). The approach taken in ABCD is to have separate element types for Latitude/Longitude and UTM coordinates with a third, generalised element for other coordinate systems. This allows for the transfer of a wide range of coordinates and also the recording of original and transformed coordinates (e.g. as in Recorder where OS Grid References are converted to Lat/Long.) although there is not an attribute for which is the original reference.

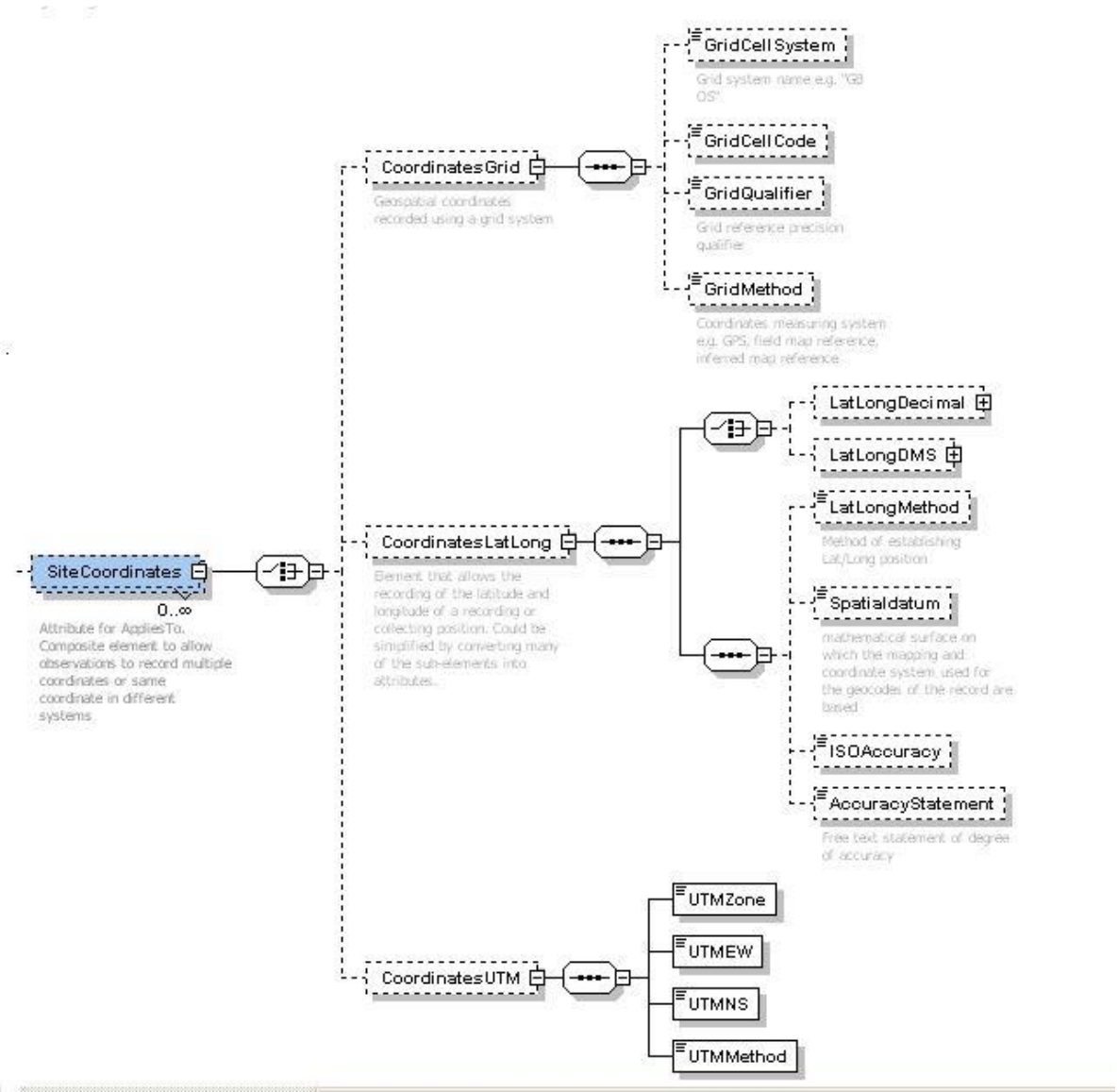


Figure 8: ABCD Schema vs. 1.30 content model for SiteCoordinates Complex Type Element

5.4 Date & Time

Date handling in Recorder

Dates appear throughout the NBN Data Model and in the logical model dates may be represented by exact dates, date ranges, imprecise dates and dates using period classifications (e.g. geological time periods). The way these different date forms are managed is left to the application developer to decide and in the NBN context this has lead to the development of the NBN Vague Date convention, described below.

In the *Recorder* application, dates may be stored as exact dates or Vague dates.

The format of exact dates is set by the Regional settings of the version of Windows upon which *Recorder 2000/2/6* is run. Ordinary date fields must be valid dates and are displayed using slashes (e.g. 26/8/1949).

The format for processing and storing vague dates is a convention and not a mandatory part of the NBN data model. It represents a practical solution to a widespread problem, that of being able to record dates and date ranges with variable degrees of precision.

Vague dates are represented in Recorder by 3 fields. These fields are used to delimit the range of a vague date e.g. a year such as 1977 would be stored as type= Y start = 01/01/1977 end=31/12/1977

- *VAGUE_DATE_TYPE: 2 character text field describing form of vague date (e.g. Y or D)*
- *VAGUE_DATE_START: Date/Time field*
- *VAGUE_DATE_END: Date/Time field*

The vague date format can accept a number of types of entry including date ranges, month or season, year or year range. Vague date values are internally interpreted by the software as exact starting and ending dates. The start and end dates are coded in the database as the number of days since 30/12/1899. If a two digit year (e.g. 99) is encountered the software interprets the century according to a user set cut-off year in the software options (e.g. <40 is the next century >=40 is the last century)

COLUMN NAME	DESCRIPTION	TYPE	SIZE
VAGUE_ DATE_START	Specifies the vague date for a taxon determination. Earliest possible date for a range specified by a vague date. Not Required. (e.g. 26/08/80	integer	
VAGUE_ DATE_END	Specifies the vague date for a taxon determination. Latest possible date for a range specified by a vague date. Not Required. (e.g. 30/08/80)	integer	
VAGUE_ DATE_TYPE	Specifies the vague date for a taxon determination. Indicates the type of vague date. Not Required.	char	2

Table 7: Vague date field format used in Recorder

The need to be able to express date ranges and imprecise dates has been recognised in other models. *Figure 9* shows the date and time elements used in vs. 1.3 of the ABCD schema. This arrangement allows for the precise recording of dates and date ranges and the indication if an event falls within is fully coincident with the given range. Where the date given is not formal calendar date (e.g. Summer 1997) this is limited to a single text field (DateText), whereas in the NBN/Recorder model the three field vague date format allows for application 'intelligence' to parse the fields according to date type and stored information such as the applications understanding of Spring or Autumn in date terms. A simplified Date Type is proposed vs. 1.36 as shown in *Figure 10*.

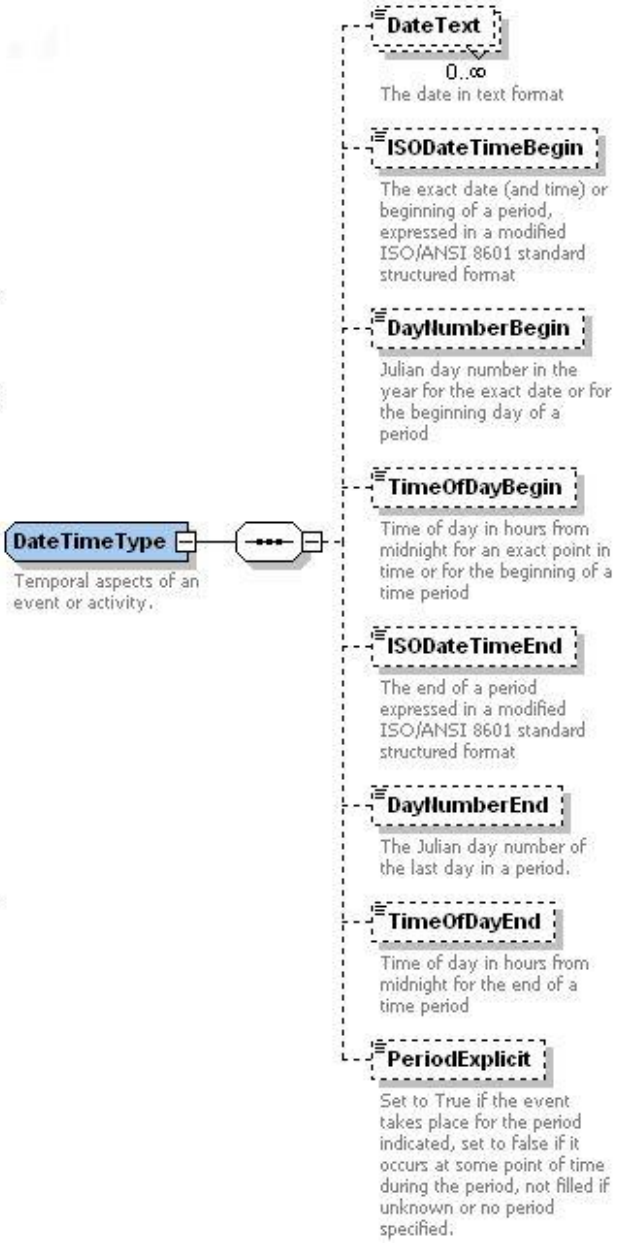


Figure 9: DateTimeType Named Complex Type from the ABCD Schema vs.1.3

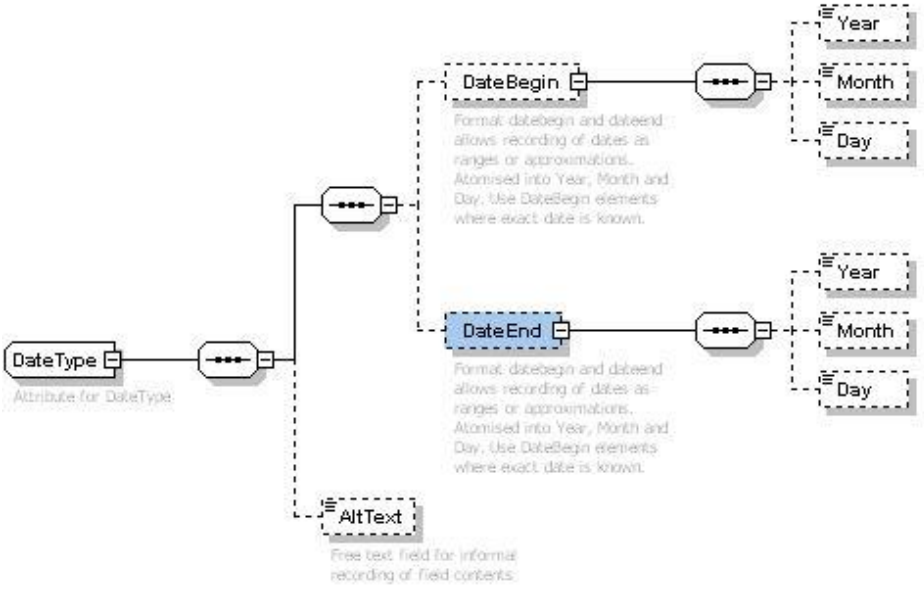


Figure 10: Simplified Date Type in ABCD vs. 1.36 gives choice of an exact date range or a free text field.

6. Thesaurus Module (including NBN Dictionaries)

6.1 The Thesaurus Module – logical model

The Recorder application utilises separate dictionaries for species, biotopes and administrative areas but all of these dictionaries (although varying in structure) are based on a single logical model. This is the NBN Common Thesaurus Model has been further developed and adapted under the BioCASE project and which is used in the specimen and collection management add-in for Recorder 6.

The principal purpose of the Thesaurus is to provide controlled and classified terminology to enable standardised data entry and reliable data retrieval from databases. The NBN/BioCASE Thesaurus (or its constituent dictionaries) is not intended to hold just a single authoritative term list for each discipline, instead it is intended to hold multiple lists (reflecting those that are or were used by original recorders) and to relate those terms to each other. It is an NBN principle that original data should never be altered, the original and any subsequent determinations associated with records should be preserved and name changes (even changing an old name to a current one) should be regarded as a re-determinations or identifications. The thesaurus can, of course, flag those terms that are currently regarded as preferred, a job which, for instance, is being undertaken for the UK Taxon Dictionary by the Natural History Museum London [24].

The Logical Entity Relationship Diagram (ERD) for the Thesaurus model is given in *Figure 11*. The Thesaurus Model is primarily concerned with the relationships of terms (e.g. taxon names, place names, habitat names) with the concepts that use them. The same term may be used many times in many different contexts and with differences of meaning, Lists of terms such as taxonomic revisions, habitat classifications and gazetteers are, therefore, regarded as collections of concepts that use terms in specific ways; this is why lists have been re-named **Concept_Groups** in the revised thesaurus model. Concept_Groups (Lists) may be simple lists, hierarchical lists or complex classifications. The terms, stored under the **Term** entity can be referred to concept_groups deriving from many different disciplines (e.g. taxonomy, biotopes, stratigraphy, gazetteers etc.) and so the thesaurus includes entities for **Subject_Area**s sub-divided into **Domains**.

The same term may be used in different concept_groups (lists) and may even be used in different subject domains (e.g. plant names, animal names and fossil names may be homonyms). Many terms have only one meaning but others such as taxa and geographic place names may be used in several contexts, for instance, through the process of taxonomic revision (e.g. lumping and splitting) or redefinition of boundaries. It is thus necessary to have a **Term_Version** entity so that the right 'meaning' of a term can be linked to its use in a specific term list. A term list, such as a taxon checklist or gazetteer of place names, is referred to as a **Concept_Group** (or **List** in Recorder Dictionaries**).** Concept Groups may be revised or added to and released in different versions, referred to in the model as **Concept_Group_Version** (or **List_Version)**.

The central entity in the model and the link between terms stored in the Term entity and the Concept Group entity, is the **Concept** entity **(List_Item** in Recorder). The Concept entity enables the same term (e.g. a species name) to appear in many lists. Many term lists include synonyms, homonyms, vernacular and multi-language versions of terms. To handle the situation where a concept can be represented by many terms and variants, there is a **Term_Version_in_Concept** entity (**Term_in_Item** in Recorder Dictionaries), which links term versions to the Concept. This allows a single 'place' in a list or hierarchy to be occupied by one or more 'equivalent' terms, one of which may be flagged as the preferred term. It is possible to have different types of preferred term such as preferred binomial and preferred English vernacular name.

[24] see http://nbn.nhm.ac.uk/nhm

Concepts are linked to list versions, but this too is a potential many to many relationship, as terms may be added, deleted and added again during various revisions of term lists (e.g. lists of protected species in legislation). This relationship is handled by the **Concept_In_Concept_Group_Version** entity.

Hierarchical relationships of terms and term synonymies are functions of individual versions of lists. For instance, it is quite common for 'competing' taxonomic checklists/revisions to have quite different relationships for taxa. This can be handled in a number of ways. If the term list is hierarchical each term can store a pointer to its 'parent term' in the hierarchy. The nature of the hierarchical relationship can be stored in a separate **Hierarchy** entity that, among other things, can store the order and sort positions of hierarchical terms (e.g. phyla, classes orders, Eons, Periods, Stages). Hierarchical relationships within a list can therefore be handled by a combination of declaring the term's rank and declaring the immediate parent term for the current item.

List items that need to be sorted in other than alphanumeric order can also have a numeric concept (list item) sort code to ensure listings come out in an appropriate order. Any further details of relationships between List Items (groups of equivalent terms) either within or between lists can be recorded in the **Related_Item** entity. This entity could be used with weightings to record more 'fuzzy' links between terms (e.g. Molluscs might have shells and shells might belong to molluscs).

Facts about terms, such as a taxon's association with given biotopes or other taxa are linked to the taxon version in the **Taxon_Version_Fact** entity if the fact is true regardless of the list a term might appear in (e.g. The Death Cap (*Amanita phalloides*) is always poisonous). There may be information about a term which is specific to an individual list (e.g. a taxon description from a specific checklist) and this should be linked to Concept (List_Item) in the **Concept_Fact** (**List_Item_Fact** in Recorder) entity.

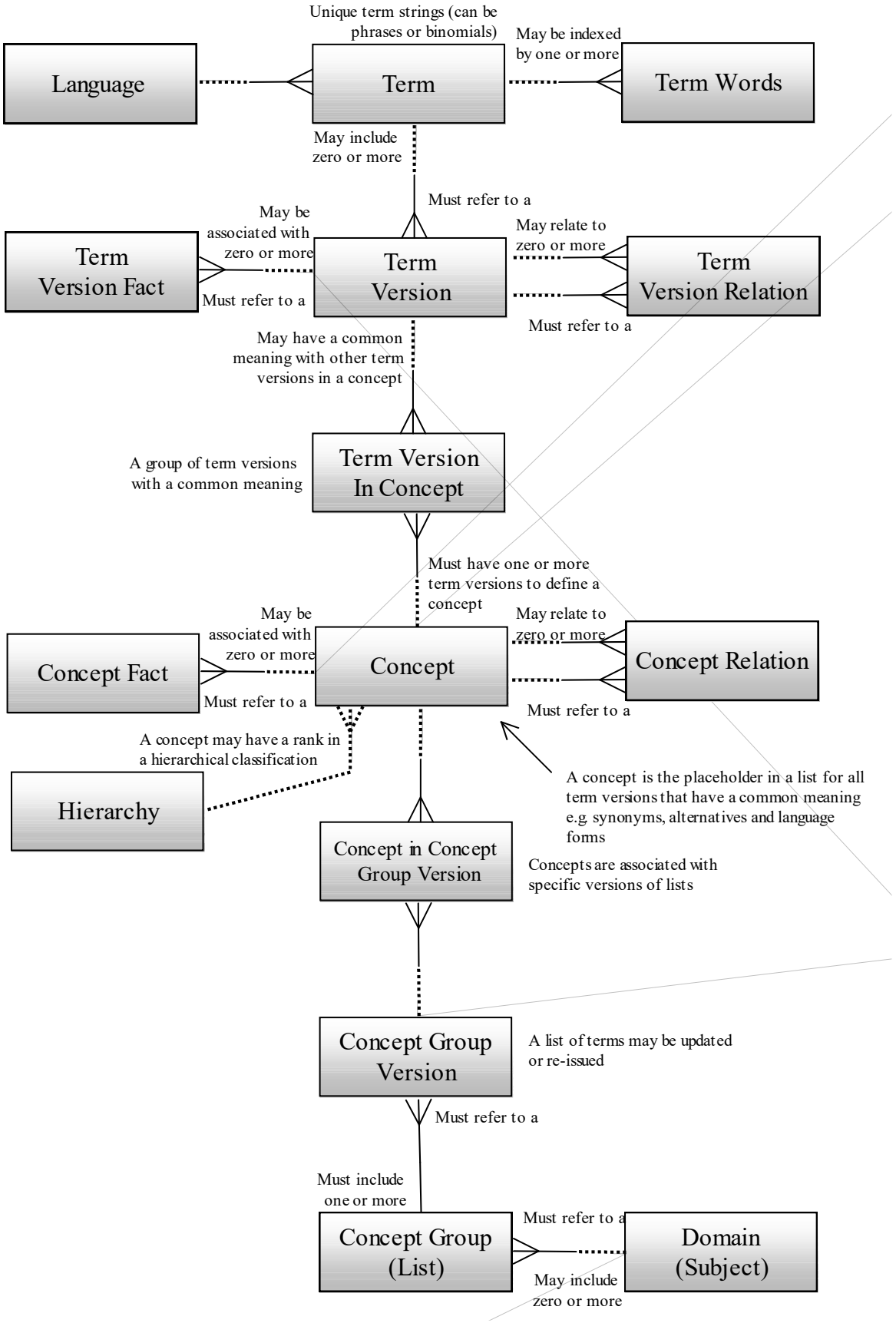


Figure 11: The Logical Model for the Thesaurus Module

Table 8: Description of entities in the new NBN Thesaurus Logical Data Model

Thesaurus Entity	Role in Model
Term_Words	Not an essential entity. Holds an index to every meaningful word in the Term table useful for indexing multi-word entries.
Term	Holds unique term values together with other attributes including language and term type. Terms may be single words or entire phrases including taxonomic binomials or trinomials.
Language	The Thesaurus needs to manage both formal nomenclatural terms and multi-lingual vernacular terms
Term_Version	Essentially every use of a term that has a different interpretation e.g. in geographic extent or taxonomic interpretation. Term and Term Version could be all in one table but this would increase data redundancy.
Term_Version_Relation	Maps the links between versions or interpretations of terms – Can relate terms within and between lists. Its used here rather than in Concept because it means for instance that a version of a taxon that is related to two other taxa by merging can be incorporated into many different lists without repeating the relation information.
Term_Version_Fact	Any factual information that is linked to a specific meaning of a term e.g. the boundary data for the 1974 version of the County of Somerset
Term_Version_In_Concept **(Term Version in List Item)**	Essentially a convenient way of listing a synonymy and common or alternative names. Also allows attributes such as preferred term or synonym list order. Could be done using the related item table but kept separate in the the model so that the Concept_Relation table is used for inter-list relationships. This table covers 'Use' and 'Use For' thesaurus relations.
Concept **(List Item)**	Links unique terms (term versions) with lists or references. Essential for tracking origins and use of terms. Can include both parent term and sort order to enable reconstruction of original list structure. Parent term gives Broader Term/Narrower term thesaurus relations.
Concept_Fact **(List Item Fact)**	Any factual information linked to a term in a specific list e.g. taxon codes or list specific descriptions.
Concept_Relation **(Related Item)**	This table is here specifically to cover Thesaurus-style related item links e.g. links outside of the current 'tree'. e.g. Psiloceras planorbis zone in the biostratigraphy tree can be linked to Hettangian in the Chronostratigraphy tree.

Concept_In_Concept_Group_ Version **(List Item in List Version)**	This entity links list items (terms) to specific versions of a list – items may be added to or deleted from lists by revision and update. This relationship can be physically built in various ways.
Concept_Group_Version **(List Version)**	Many standard lists are reviewed and revised with terms being added and deleted. Could be merged with lists but increases data redundancy.
Concept_Group **(List)**	The source of a list of terms or names. Can be a taxonomic review but could be any source of terms e.g. existing thesauri, gazetteers, legislation lists or even an *ad hoc* list for pragmatic purposes.
Domain **(List_Type)**	Provides a higher classification of subject areas so that lists can be readily sorted or filtered in thesauri covering multiple subject areas. Domains can be further divided into child domains e.g. lists of taxa used in legislation, taxonomic revisions and checklists used in recording schemes. In the Recorder collections add-in this entity is divided into three physical tables: Subject Area, Domain and Local Domain (for local implementation specific filtering e.g. only French lists).
Hierarchy	Lookup table – hierarchical terms related to term types (e.g. phylum, class, family for taxa)

The logical model described is deliberately highly normalised and allows for the various relationships between terms and the lists in which they occur to be readily visualised. The logical model can, however, be expressed in many physical ways depending on the intended use and database software employed. For some purposes the number of tables can be drastically reduced and the physical build can appear quite different. The Recorder Taxon Dictionary and the version used by the Natural History Museum, London to manage the UK Taxon Lists is very close to the logical model in structure (enhanced with various index and view tables) but the BioCASE Thesaurus looks very different because it has been built to enhance data retrieval speeds, which required taking the version tables out of the central table linking path and a degree of de-normalisation.

6.2 Implementation of the NBN Dictionaries

6.2.1 NBN Dictionary Overview

The first version of the NBN Model, and that which was used in the development of Recorder 2000/2, includes different dictionaries for taxa, biotopes, administrative areas, and general term lists; this is the approach taken by most biodiversity application databases. Subsequent work by the author (Copp 2002, 2003) for the European BioCASE project has shown that all of these dictionary structures can be readily mapped to a common logical model and implemented in a common physical thesaurus database

The principal design objective for the NBN dictionaries is that they are intended as mechanisms for storing and managing many term lists and versions of lists together with the means for translating or relating from one to another. The purpose of the dictionaries is to enable users to enter data using exactly the terms (and meanings of terms) as used by recorders; even with old records using ‘out-of-date’ nomenclature. Changing the taxon or biotope name associated with a

record to a different or newer name is seen as a new determination or identification, which must be attributable to the person making the change. The purpose of the dictionaries is also to improve retrieval of information through the ability to find synonyms, broader terms or narrower terms and to widen or refine queries.

The NBN Taxon Dictionary (or Species Dictionary, as it is more frequently referred to) has the most complicated structure of the Recorder dictionaries. It is represented here by the physical table structure used In Recorder 2000/2 (*Figure 12*). The Biotope Dictionary physical model (*Figure 13*) represents a simplification of the Taxon Dictionary Structure except for the addition of tables to map the requirement for allowing the mapping of a habitat in one classification to habitats in other classifications to vary geographically. The Administrative Dictionary physical model (*Figure 14*) is even simpler being reduced to term, term type and term relation entities. In Recorder there are many simple term lists which are stored as single tables with a minimum attribute (field) content of short name, long name and description although some such as **Measurement_Unit** include other fields such as data_type. All of these simple term lists could be handled easily within a single thesaurus and use of such would have reduced the number of application specific tables and joins used.

The Recorder dictionaries as presented in the original NBN model and incorporated into Recorder 2000/2 are described in the following sections. Recent work by the author (Copp 2002, 2003) for the BioCASE project has shown that all of these different dictionary structures can be readily mapped to a common logical model and even to a common physical thesaurus database. It is proposed that the new logical model be adopted in the updated version of the NBN Data Model presented in this work. The BioCASE and proposed new NBN Thesaurus Model is discussed in section 2.3 and in Part 2 of this work.

6.2.2 The NBN Taxon Dictionary – physical model

The NBN Taxon Dictionary is an example of how a physical data structure may be modified from a logical structure for the purposes of an application and to match aspects of the data being stored. In the case of the species dictionary, the physical model was derived from an earlier version of the thesaurus logical model and matches the earlier model very closely. One of the key differences is the lack of the equivalent tables relating to **Term_Version_in_Concept** (this would be **Taxon_Version_in_List_Item**) and **Concept_In_Concept_Group_Version** (**List_Item_in_List_Version**), which are handled in a simpler way within the **Taxon_List_Item** Table.

It was not the intention that the NBN Taxon Dictionary should be a single, comprehensive and authoritative list of UK taxa; that such a list should exist is a desirable thing, but is a different project[25]. The purpose of the Taxon Dictionary is to store and relate as many lists and versions of lists as are necessary for the unambiguous referencing of species names supplied by recorders or gleaned from collections and literature. The key to the Taxon Dictionary structure is the linking of taxon names to taxon lists via the **Taxon_List_Item** table (the Concept entity in the new model); the same name can appear in many lists. The physical table model used for the 'Species Dictionary' in Recorder is illustrated in *Figure 12* and a summary of the tables given in *Table 9*.

[25] The building of a UK list is currently being coordinated by the Natural History Museum London (NHM) using the Taxon Dictionary model to store the information.

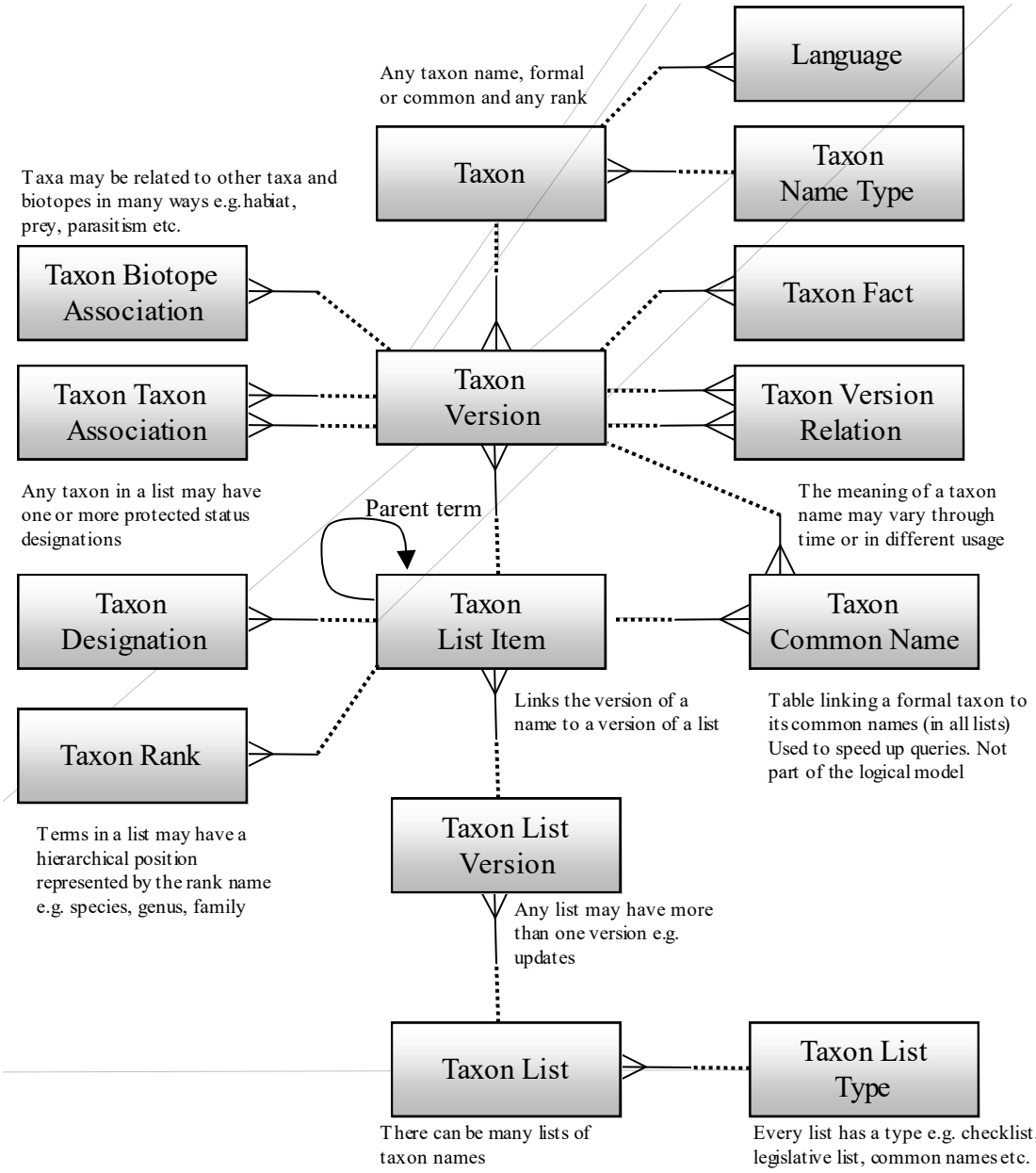


Figure 12: Physical Data Model for the Taxon Dictionary used in Recorder 2000/2

A **Taxon_List** may be any defined grouping of taxa, most commonly a published list but it can also be an informal list used for some particular recording project or even locally used common names. The dictionary can also represent lists of taxa given in schedules and annexes to conservation legislation, or for any other purpose and in any language. The advantage of storing lists in this fashion is that the 'schedules' can be recorded with the actual taxon names used in the printed legislation (in which the names used may not be either current or correct!).

In this model, the legal conservation status of a species is not, therefore, held in a single list associated with a single taxon name but could be constructed for any species by looking at the status attached to it or its synonyms in a selection of lists representing different acts of legislation (and their amendments). This method also allows for the recording of status given in other types of list including red data books and biodiversity action plans. The model takes account of the possibility that a species may be given more than one status in a particular list (e.g. a Red Data

Book might list World, European and UK status) and also geographic limits to protection or status.

In the model represented in *Figure 12*, the **Taxon** table holds all unique name and author combinations with date of introduction. This table should include all synonyms, genus/species combinations, infra-specific and sub-specific names as well as hybrid formulae. These names are interpreted and related through other tables.

All unique names have at least one version but the same term may be used in several contexts through the process of taxonomic revision (e.g. lumping and splitting). It is thus necessary to store taxon keys in a **Taxon_Version** table so that the right 'meaning' of a taxon can be linked to its source reference and specific **Taxon_List(s)** or **Taxon_List_Version(s)**. Note that lists may be released or updated in many versions and individual taxa may be added, removed or re-added in any version. It is also possible by using fields (attributes) for date added and date removed to have dynamic versions of lists that can be edited and expanded over a period of time rather than in single updates.

Earlier versions of the model held common names and informal groupings of taxa (e.g. the term waders) in a separate table linked to taxon but there is no difference between formal taxa (ones created within the rules of the nomenclatural conventions) and informal or vernacular ones in terms of versions, relationships and associated information. For instance, bird recording in the UK is almost exclusively carried out using common names and protected status, distribution facts and family relationships are all linked to the common names. All naming terms are therefore placed in taxon and taxon version. This does not preclude separation for convenience in particular applications and in Recorder a 'view' table is automatically generated that links all known common names to a binomial, in order to speed up retrieval times in queries and reports.

As discussed in the description of the logical model, hierarchical relationships of taxa and taxon synonymies are functions of individual versions of checklists. It is quite common for 'competing' checklists or revisions to have quite different relationships for taxa. In the Recorder Taxon Dictionary, this is handled by representing each taxon in a particular checklist version in the **Taxon_List_Item** table. Hierarchical relationships within a list are handled by a combination of declaring the taxon rank (from the **Taxon_Rank** lookup table), declaring the immediate parent term for the current item (as an attribute of Taxon List Item) and storing a checklist item sort code to ensure listings come out in an appropriate order. Synonyms are identified by having an entry for the List_Item_Key of the preferred term.

Facts about the taxon such as its association with given biotopes or other taxa and general information (biology, behaviour etc.) are linked to the taxon version. There may actually be information about a taxon which is specific to an individual list and this is linked to the Taxon_List_Item (e.g. the scientific description) but this can also be accomplished by including a link to the individual Taxon List in the **Taxon_Fact** table (e.g. include a Source Key in the record).

One of the major problems of managing taxonomic lists is sorting them into order for retrieval and reporting. Taxonomic hierarchies are not fixed and levels vary widely from group to group. Most applications overcome this problem by the use of either fixed taxonomic hierarchy levels and/or using taxonomic sort codes. Sort codes may be simple integers or complex codes with several parts which attempt to codify a taxon's hierarchical position (e.g. a code for phylum + code for class + code for family + code for genus + code for species). Some code systems are confined to the original list for which they were created, whereas others have been utilised in more than one database or on a variety of recording cards. The taxon dictionary stores species codes as an attribute in the Taxon_List_Item table, linked to the original list which created the coding system. Where other lists use a code from another system (e.g. a recording card which refers to the BRC taxon list) this is pointed to by a taxon code source attribute in Taxon_List_Item.

Table 9: Description of entities in the Recorder Taxon Dictionary Model

Table in Recorder	Purpose
Taxon	The master list of all taxon names
Taxon_Name_Type	Type of name e.g. informal name or zoological binomial, botanical binomial etc.
Taxon_Version	Different usage of names in Taxon usually associated with taxonomic revision e.g. splitting or lumping of species.
Taxon_Version_Relation	How the different uses of the same name relate – e.g. used for linking segregates to an aggregate name. Can also link named hybrids to parent taxa.
Taxon_Taxon_Association	Used to record specific (e.g. symbiotic, parasitic etc.) links between two taxa. *Polysiphonia lanosa* grows on *Ascophyllum nodosum*. (special form of the Term_Version_Relation table in the logical model)
Taxon_Biotope_Association	Links between a species and a habitat. (special form of the Term_Version_Relation table in the logical model)
Taxon_Fact	This entity can hold any general information about a taxon together with a pointer to the source of information.
Taxon_List_Type	There are many possible types of list e.g. taxonomic checklist, recording card, schedule to legislation, red data book, species of conservation concern etc.
Taxon_List	Name and details of any particular list of taxa
Taxon_List_Version	All lists have at least one version. Most lists are periodically reviewed, amended or otherwise altered
Taxon_List_Item	The actual names occurring in any list – (foreign keys to Taxon Version and Taxon List Version). Note that several names may represent a single position (i.e. taxon) in the list. This table stores the preferred name Taxon_List_Item_Key for each Taxon_List_Item as a means of indicating groups of synonyms and their preferred name.
Taxon_Rank	Lookup table of hierarchical ranks – can be used to hold formatting information for output.
Taxon_Designation_Type	Within Acts and Conventions this entity covers the various schedules and Annexes. In non-statutory lists it represents the designation classification e.g. RDB
Taxon_Designation	Actual designations applied to a taxon in any list. There may be more than one status/designation for any taxon e.g. national status and world status

In addition to the above tables, the taxon dictionary includes a number of indexing tables updated directly from the dictionary, in order to simplify and speed up data retrieval. These include the **Index_Taxon_Group**, **Index_Taxon_Name**, **Index_Taxon_Synonomy** and **Taxon_Common_Name** tables. The NHM have developed other view tables to aid their work but all of these tables are populated with copy (redundant) data and do not form part of the data model. The full table and attribute description for the Recorder Taxon Dictionary is given in Part 2 sections 2.8 and 2.9.

6.2.3 The NBN Biotope Dictionary – physical model

The NBN Biotope Dictionary is currently under review and may possibly be re-structured for management purposes. The version described here is that deployed in Recorder 2000/2002. The information currently held for biotopes is somewhat simpler than that for species lists and the physical model is correspondingly simplified.

As in the taxon dictionary, the key table is the list item (**Biotope_List_Item**), which represents a selection of individual biotope or land-cover terms held in the **Biotope** table and links to a **Biotope_Classification** (a checklist) through a specific **Biotope_Classification_Version**. Unlike the taxon dictionary, there is no biotope_version table as they are rarely redefined and biotope names are not often shared between different classifications. Biotopes may have facts and statutory or informal designations attached to them and this is done through the Biotope_List_Item.

There are differences in detail of the field (attributes) values of the tables between the taxon and biotope dictionaries. For instance, the Biotope Dictionary allows a long name and a short name to be recorded for biotopes in the Biotope_List_Item table because many habitat classifications have longer and shorter versions of the entries as well as codes. In the new thesaurus model the long and short names would be held in the term table and regarded as synonyms.

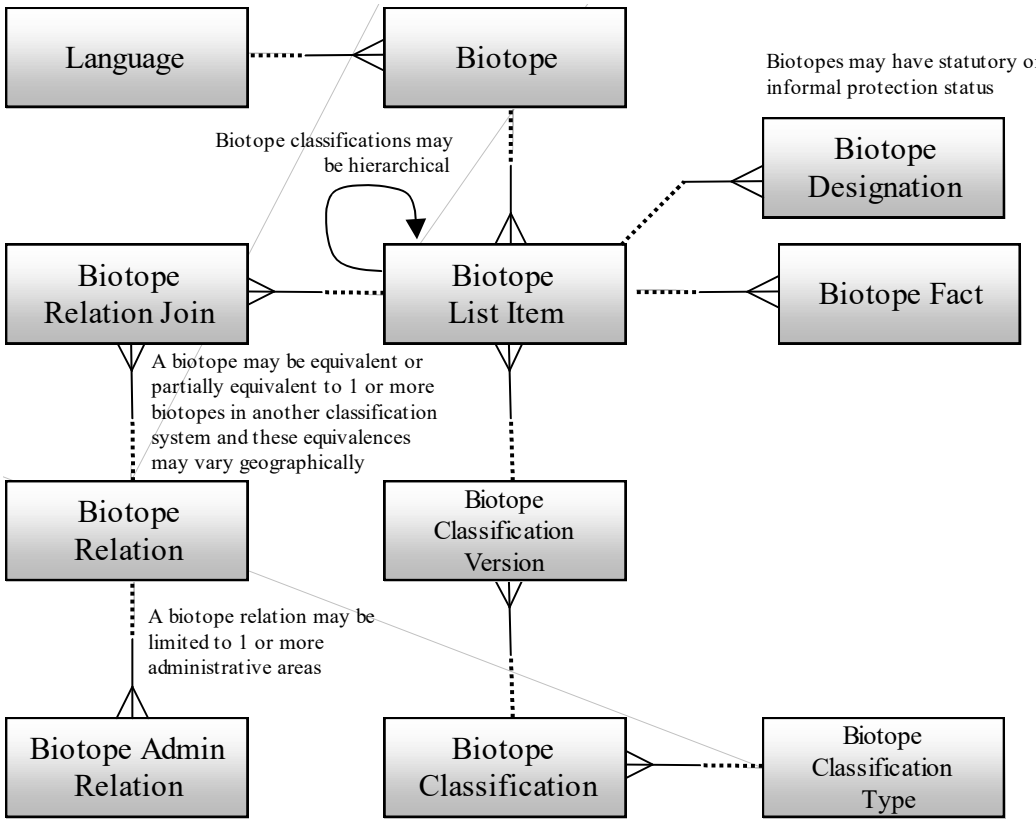


Figure 13: The Recorder Biotope Dictionary uses a simpler version of the Taxon Dictionary model but also adds extra elements for handling the complexity of biotope equivalences

The order of biotopes within checklists may be recorded by a sort number and their hierarchical position by a parent term code, represented by the curved arrow ('pigs ear') in *Figure 13*. Most biotopes have a single code within the original classification which is stored as an attribute of the Biotope_List_Item.

Biotopes may be related to one another (e.g. for the purposes of mapping equivalents in different classifications) through the **Biotope_Relation** table. One of the problems with equating habitats or biotopes from one list with those in another (e.g. UK NVC[26] habitats with EUNIS[27] habitats) is that the mappings are rarely on a one-to-one basis; normally several habitats may partially relate to one habitat and the degree of relationship may vary geographically. The way this is handled in the Recorder application is for the Biotope_Relation table to be linked to a **Biotope_Admin_Relation** table so that any relationship can be linked to one or more places or administrative areas (e.g. counties or countries – the Admin refers to an entry in the NBN Admin or Administrative Areas Dictionary). The need to relate several biotopes in a relation is handled by the **Biotope_Relation_Join** Table. The logical model allows for the statement of a measure of the completeness of the relationship (e.g. Habitat A and Habitat B respectively map as 30% and 70% equivalent of Habitat C in Wales). Unfortunately, the fact that it might be possible to record such fine levels of discrimination does not make it easy to either obtain the data or to write applications that can use this information! This is why the latter feature has not been implemented in the Recorder application.

6.2.4 NBN Administrative Area Dictionary

The NBN Administrative Area Dictionary, as implemented in Recorder, represents an even greater simplification of the general thesaurus model. In this dictionary all terms representing place names or geographic areas are stored in the **Admin_Area** table with a link to a list name held in the **Admin_Type** Table. Versions are treated as new lists. Areas may have pointers to stored boundaries associated with them and relationships are dealt with by a parent key in the Admin_Area table for mapping hierarchies and links in the **Admin_Relation** table for joining items in lists (e.g. linking districts to their containing counties or counties to countries).

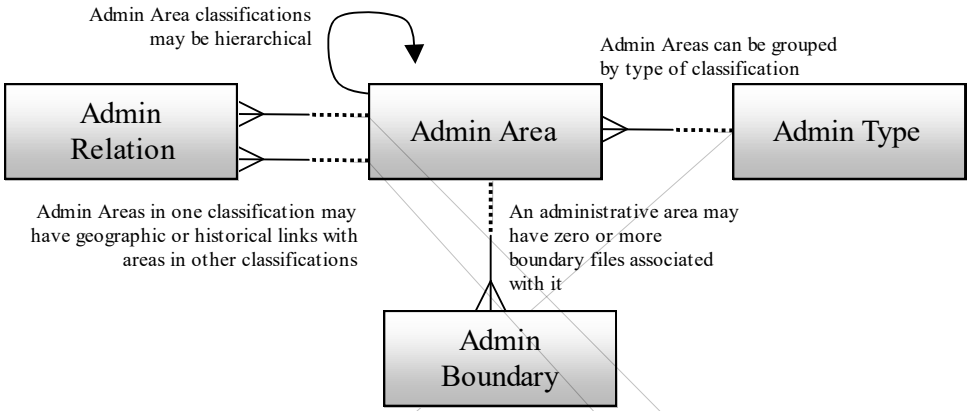


Figure 14: The Recorder administrative area (Geographic) Dictionary uses a very cut-down version of the dictionary (thesaurus) model

[26] National Vegetation Classification: The NVC is published as a five volume series entitled *British Plant Communities* by Cambridge University Press. Summary descriptions for the *woodland* and *grassland* and *montane* communities are published by the JNCC.

[27] EUNIS: The European Nature Information System, developed and managed by the European Topic Centre for Nature Protection and Biodiversity (ETC/NPB in Paris) for the European Environment Agency (EEA) and the European Environmental Information Observation Network (EIONET). The EUNIS Habitat Classification is a comprehensive pan-European biotope classification, covering all types of habitats from natural to artificial, from terrestrial to freshwater and marine. (see http://mrw.wallonie.be/dgrne/sibw/EUNIS/home.html)

6.3 Recorder 6 Collections Add-In Thesaurus (BioCASE Thesaurus)

6.3.1 Thesaurus physical model

Collaboration over three years between the author, Luxembourg National Museum of Natural History and Dorset Software Ltd. has resulted in the development of a number of major extensions to the Recorder application under the general name of the **Collections Add-in**. The Collections Add-in not only adds a range of functionality for documenting and managing specimens and collections but extends the scope of recording from living taxa and biotopes to earth sciences and related disciplines and adds a completely new Thesaurus module that can replace the old separate dictionaries.

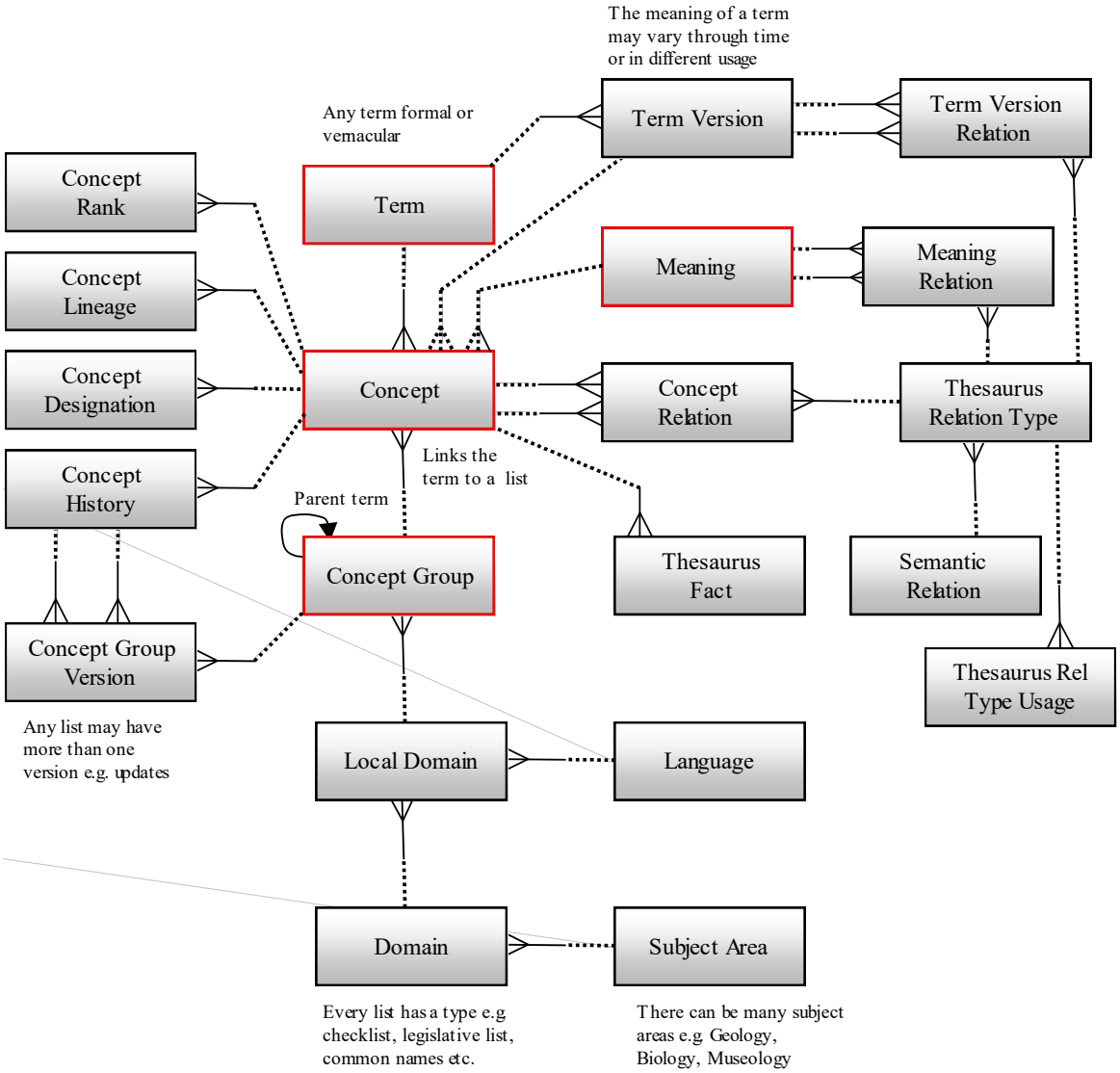


Figure 15: Physical model for the Thesaurus database in Recorder 6 Collections Add-in and Thesaurus used in the BioCASE Project. The tables outlined in red are the ones most used in queries.

The great extension of functionality provided by the new collections application required the management of many new term lists and new dictionaries to cover palaeontological taxa, mineral

nomenclature and others and also addressed the need to internationalise dictionaries to deal with the issue that most museum collections contain specimens from many countries. This latter requirement, for instance, introduces the need for much bigger species lists and gazetteers.

The Recorder 6/BioCASE Thesaurus physical model deals with concepts, how concepts relate to one another and how they may be arranged into classifications that reflect those relationships. A single concept can be represented by many terms or names and these are linked through the MEANING table. Concepts can be clustered into a variety of organisational structures including plain lists and hierarchical trees, so for this reason the LIST_ITEM entity has been re-named CONCEPT and LIST has been re-named CONCEPT_GROUP. *Figure 15* shows the physical model with re-named entities. The key difference between this physical model and the Recorder Taxon Dictionary model is that the version tables (TERM_VERSION and CONCEPT_GROUP_VERSION) are moved to the side of the main tables to cut down the number of links between tables for most queries; this works because most terms and lists only have one version and it is quicker to look up version data once the main query has been run.

A potentially confusing aspect of the new model is the rather different use of CONCEPT between the logical and physical models and also the introduction of the MEANING table. In the logical data model the LIST_ITEM (CONCEPT) entity is a placeholder for a position in a list and the TERM_VERSION_IN_LIST_ITEM (TERM_VERSION_IN_CONCEPT) holds the foreign keys to the TERM table. In the new physical Model, CONCEPT also holds the TERM foreign keys and MEANING provides the link between synonyms and equivalents. MEANING holds no data but provides a Key attribute that can be applied to all terms (CONCEPT) that are linked as a single item in a list (CONCEPT_GROUP). MEANING can also link to FACTS so that information can be applied to all the terms covered by the same meaning and to the relationships between equivalent terms (e.g. multi-lingual versions of common names of taxa, synonyms etc.).

Because the thesaurus can manage concepts and concept groups relating to any subject, the model includes DOMAINs. In the physical model presented in *Part 2* of this report, this is further refined into SUBJECT_AREA (e.g. Geology, Biology etc.) and the DOMAINs represent subject disciplines (e.g. Mineralogy, Taxonomy etc.). For the physical build in the Luxembourg Project[28] we have introduced a LOCAL_DOMAIN which allows CONCEPT_GROUPS to be preferentially linked to a local (e.g. country specific) implementation of the thesaurus (e.g. Taxon Lists specific to a country).

A description of the function of each of the tables in the Recorder/BioCASE physical model and details of the fields (attributes) in each table is given in *Part 2* of this report.`

6.3.2 Comparison with the Berlin Taxonomic Dictionary Model

During 2003, Berendsohn et al. published an important contribution to the development of taxonomic dictionaries and the handling of information linked to taxonomic concepts under the name of MoReTax[29]. This work, derived from developments in projects such as the Euro+Med database managed by Berendsohn's team at the Berlin Botanic Garden and Botanical Museum was carried out independently of the BioCASE Thesaurus Project. The table structure for the MoReTax model is shown in *Figure 16*.

[28] The Luxembourg Project is developing a fully integrated field records and collections management application that covers both biological and earth sciences. The application is an extension of the UK National Biodiversity Network biological recording software called Recorder (see: http://www.nbn.org.uk/). The version being modified is Recorder 6 which is a client server application written in Delphi and using a MicroSoft SQL-Server 2000 database backend.

[29] See ***The Berlin Model: a concept-based taxonomic information model*** Berendsohn et al. 2003 In ***MoReTax****: Handling Factual Information Linked to Taxonomic Concepts in Biology* Schriftenreihe für Vegetationskunde Volume 39 Federal Agency for Nature Conservation 2003.

Both the 'Berlin' model and the Recorder 6 / BioCASE model separate terms (or Names) from the Concepts that refer to them. The 'Berlin model' is optimised for handling botanical information and separates the detail and structure of names (nomenclature) from their interpretation (taxonomy). The Recorder/BioCASE Thesaurus also addresses the names and taxonomy issue but is intended to be a generic concept-based management tool relevant across all disciplines. The new physical model described here for the Recorder/BioCASE Thesaurus bears a number of similarities to the 'Berlin Model'. Both models use devices including concatenated fields and view tables and a degree of relational de-normalisation to improve data retrieval speeds.

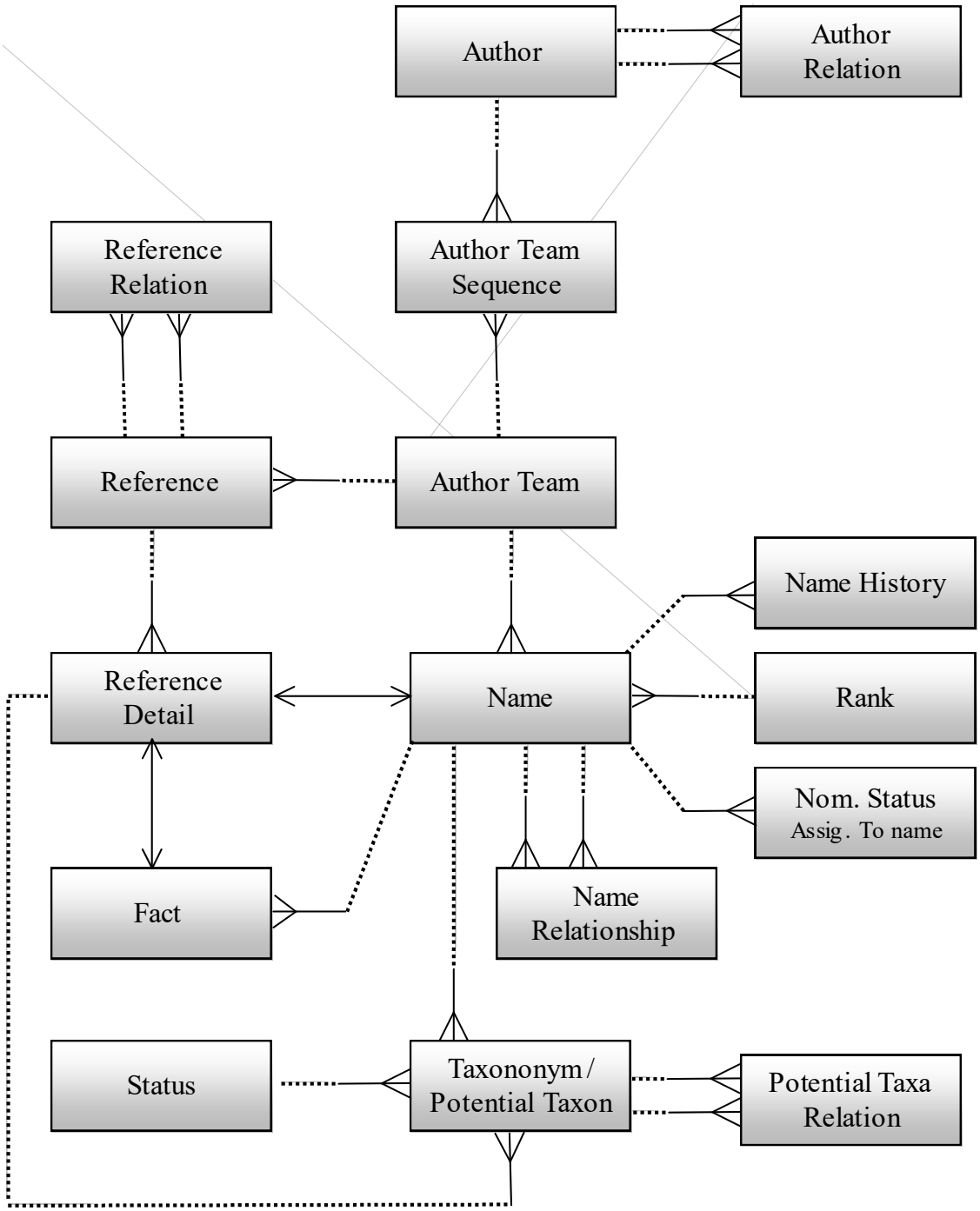


Figure 16: Simplified Table structure in the Berlin MoReTax (Berlin) Model

Both the 'Berlin Model' and the Recorder/BioCASE Thesaurus support the separation of names and the concepts to which they may be applied. In the Berlin model taxon names are atomised into Genus, species and other epithets whilst in the BioCASE Thesaurus taxon names are treated as full strings although the author/date citation is held separately. Author citations are atomised in the Berlin Model. The Recorder/BioCASE model allows for sub-typing of the TERM table; should atomisation of terms be required in the future but this is not yet implemented.

In the 'Berlin Model', NAME equates to TERM in the Recorder/BioCASE Model (although differently structured) and POTENTIAL_TAXON equates to aspects of TERM_VERSION (because this links to published reference) and CONCEPT (holds link to TERM) in the Recorder/BioCASE Model. MEANING and MEANING_RELATION in BioCASE equate to POTENTIAL_TAXA_RELATION in the Berlin Model. A difference is that the Berlin Model assigns RANK to the NAME table whereas in the BioCASE model RANK is a function of a particular list and are therefore attached to CONCEPT as is the hierarchical position of the term in a list. In the 'Berlin Model' nomenclatural status and taxonomic status are attached to NAME and CONCEPT respectively whereas in the BioCASE Thesaurus they are attached to TERM_VERSION and MEANING_RELATION.

The Berlin model and the BioCASE/Luxembourg model differ in many details including how they represent published references, details of individuals and the atomisation of terms. The models have been designed for different purposes but are sufficiently close to ensure that data can be readily mapped between them. The 'Berlin Model' (MoReTax) is a nomenclatural and taxonomic tool whereas the purpose of the BioCASE Thesaurus is to provide a multi-disciplinary 'lexicon' of terms that can enhance the indexing of data sets and improve data retrieval through the BioCASE Portal. Perhaps, as the use of the MoReTax system is extended the two applications will converge.

The current physical structure of the BioCASE Thesaurus is readily extensible to allow for the extra functionality provided by term atomisation or for linking to other modules.

7. Contacts Module (People & Organisations)

The Contacts Module recognises that links are needed to both individuals and organisations and that much of the information required is common to both. For this reason, the central entity of the contacts module is called **Name** which is linked to subtype entities for individuals (**Person**) and organisations (**Organisation**) each with its own appropriate attributes. Organisations may be divided into **Departments**[30]. Each subtype has attributes relevant to either individuals or organisations; for individuals it is possible to record birth, death and floreat (time most active) dates and this can be extended to any other biographical information; for organisations details include when founded. The logical model for the Contacts Module is illustrated in *Figure 17*.

Names of people or organisations are referred to in many places in biodiversity and collections applications; as recorders, determiners, authors, owners, participants in events etc. In the Recorder physical model listed in Part 2 of this work, the most obvious links are 'hard-coded' as attributes using **Name_Key** or a similar named foreign key link. Where the link might include more than one person or organisation, a linking entity, such as **Survey_Recorder** is used. Logically it is possible to generalise the relationship by creating a multi-purpose link entity which could associate any name with any table or attribute but this would have very few build advantages.

[30] Departments were not included in the first version of the NBN data model but have been added as part of the analysis work related to the development of the 'Luxembourg Collection Add-in'.

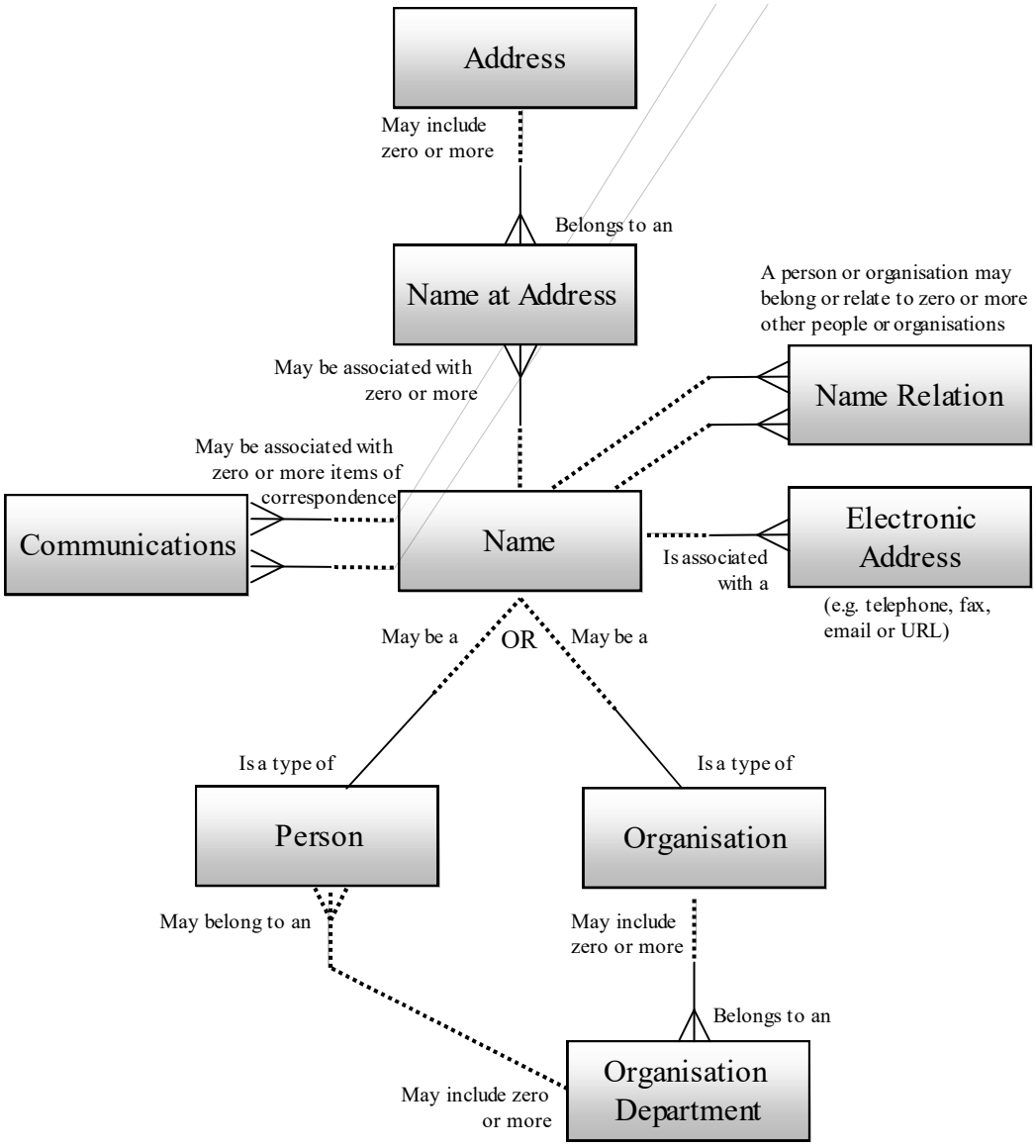


Figure 17: Simplified LDM for the Contacts Module (People & Organisations)

Individuals and organisations may also be known by one or more codes (e.g. National Insurance number and BRC recorder code) each of which can be stored in **Name_Code** (not shown in figure 14) or the information could be stored as an attribute of Name_Relation**.** This could become important as the wider availability of data through enhanced exchange facilities or the delivery of information over the Web could lead to a need for unique codes for every recorder (There could be hundreds of people called John Smith).

Details of addresses (**Address**) are kept separately and associated with individuals or organisations through a linking entity, **Name_at_Address** which includes dates. In this way several 'names' can share one address and one 'name' can have several addresses either simultaneously (home & work) or sequentially (changes of address). Any name can be linked to

any number of electronic communication numbers such as telephone, fax or email through the **Comms_Number** (Electronic Address) entity.

Names (individuals and organisations) can be linked together through a **Name_Relation** entity which records dates and nature of the association. This entity can be used to record anything from membership of societies, marriage and employment to corporate mergers. In the Recorder 6 Collections Add-In, individuals can also be linked directly to departments of their employing organisation or similar primary link. A separate entity called **Name_Role** (not shown) allows individuals and organisations to be classified for various purposes such as listing all local record centres, wildlife trusts or vice county recorders.

It is possible to use the Contact module as the basis for tracking interactions with any individual or organisation. This is represented in the model by the **Communication** entity, which includes attributes for communication type and file reference. It could include copies of files or pointers to them in other systems. This could include notes of telephone conversations, copies of letters or data supply agreements and could also be linked to a data transfer-tracking module (not yet modelled). The use of name keys from Name in the same way as the Name_Relation table (logically this is a sub-type of Name_relation) means that communication between any individuals or organisations in the database can be recorded.

8. Survey Module

8.1 Surveys, events and samples

The Recording or Survey Module, together with the Location Module forms the core of the Recorder 2000/2/6 application. The basic logical data model for the Survey Module is shown in *Figures 18 & 19*. The model attempts to overcome the perceived differences between species recording and habitat recording and also introduces the sampling element, which was missing from the earlier versions of Recorder (e.g. Recorder 3). This is a powerful model, which allows the integration of most forms of recording including extensions to cover earth science records and links to collection management software.

The division of a simple field record into the constituent parts of survey, survey_event, sample and occurrences, is confusing and seen as unnecessary by many recorders, usually because they are recording a very limited number of elements of information (e.g. date, species, grid reference) and even storing most contextual information (survey metadata) in their heads. Few amateur recorders are familiar with the concept of samples and cannot relate this to activities such as walking through a wood and listing the birds or flowers seen. This issue can be overcome by designing simple recording card-style data capture software (or forms within more complex software, such as Recorder) that allow recorders to quickly enter simple data but which also prompts for the extra information needed to make it possible to transfer data and make it compatible with data from other recorders and surveys. The storage or transfer application could generate event and sample data automatically, this is already done with some data import routines for Recorder 2002 (e.g. transfer from Recorder 3.3 and from spreadsheets)

The roles and relationships of the various entities in the Survey Module are described below. The way in which the model has been implemented in Recorder 2002 and extensions developed for the Recorder 6 Collections Add-in together with table and attribute descriptions is given in Part 2.

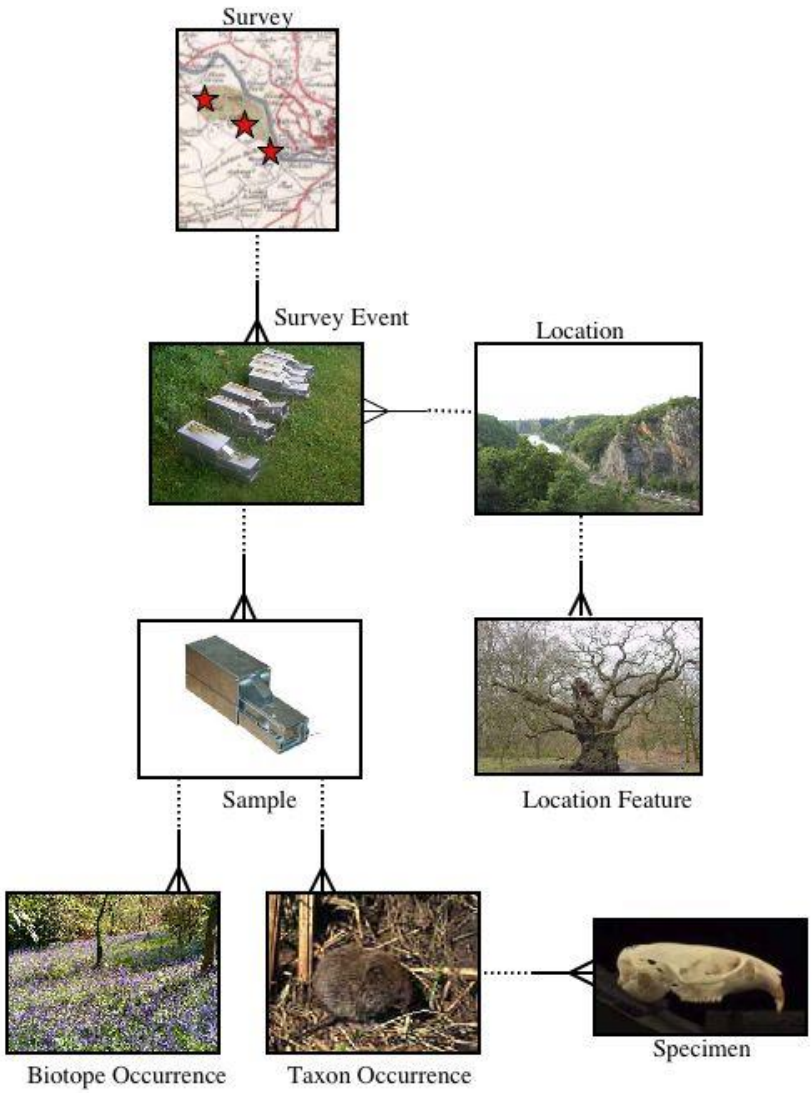


Figure 18: A schematic view of the Survey Module showing its relation to the Location Module and to specimens (including the Collections Add-in). Note how different types of occurrence can be linked to a single sample

8.1.1 **Survey**

The Survey Entity represents the need to record metadata about the source, quality and ownership of records. In simple terms the metadata groups together common data, recording, among other things, who organised the survey, its geographic range, ownership and date span. As far as the model is concerned a survey can be anything the organiser wants to call a survey, it could be one person's life-time records (e.g. "Charles Copp's Mollusc Records"), the combined records of a

local society or a much more detailed linked set of events such as a single year's field work (e.g. "Phase II Chalk Grasslands Survey of Sussex").

In the Recorder application this entity is represented by the Survey table and links to the Source Module to provide a means of recording basic metadata and links to documents or references. It is clear that not all of the information is part of an organised detailed survey but some thought will demonstrate that there is still a need to record the source and data constraints and that there are logical links which group records together – if only for validation or administrative convenience; this becomes very important once records are passed on to local record centres or made available to the NBN system.

There is still an outstanding task to be under-taken to link the metadata standard developed for the NBN to the NBN data model and re-assess how the metadata being recorded for inclusion in the NBN Gateway and Index can be harmonised with that stored in the Recorder software.

Table 10: Items of Information covered by the survey entityin the NBN Model and implemented in Recorder

Survey Information	Definition	Entity	Validation
Survey Name	Many recording events are organised in relation to wider ranging surveys. Surveys may be set up to achieve a specific target e.g. to publish a 1K square flora atlas of a county or for more general purposes e.g. a natural history society entomological group. Surveys may be geographically wide ranging or confined to a single location.	Survey	None (possibly spellcheck)
Date started	The date when the survey was formalised or when the first records were made.	Survey	Valid date - but Recorder uses its be Vague Date format
Date ended	Date survey was completed or last records made.	Survey	Valid date - but Recorder uses its be Vague Date format
Survey Status	Is the survey static or ongoing?	Survey	choice of terms linked to Survey Status table in Recorder
Survey Description	Text description of survey and its aims	Survey	Spellchecking
Geographic coverage	Statement of the general geographic area the survey relates to. e.g. county or region.. In Recorder the Geographic Coverage is covered both by a text field and the ability to enter the corner coordinates of a bounding box which can be used to validate Sample Gfrid References.	Survey	Link to administrative dictionary (gazetteer). May use controlled terminology list for improved retrieval.
Responsible	Organisation or persons responsible for organising the survey. [If this involves several organisations or individuals then a name/survey link entity would be needed – only by extension in Recorder].	Survey or Survey/Name Link	Valid Name keys
Survey Type	Many surveys fall into well-known types e.g. Phase I, Phase II, Flora, National mapping scheme etc.	Survey	Controlled terminology list - Survey Type
Survey Methods	Note of the method or methods employed in the survey to obtain data. e.g. mist nets, pit fall traps, Pollard walk, timed observation at fixed points etc. A survey may be based on more than one method.	Survey Method may need to be separate list – not used in Recorder where method is linked to Survey Event	Controlled terminology list
Periodicity	Is this a one-off or repeated survey, if so how often is it carried out?	Survey	None
Recording Media	How the data are normally recorded e.g. BRC card, Phase I card and map etc.	Survey	Controlled terminology list
Ownership	Who owns the records? Copyright and IPR. Not implemented in Recorder but is part of the Metadata standard.	Survey	

Text References	Any publications, manuscripts, letters or agreements referring to this survey.	Link to References in Sources (Internal and External sources)	Valid Source keys
Survey Quality	Measure of level of survey quality e.g. thorough, adequate, superficial. This may be supplied by the survey organisers but any such judgment will be subject to the data protection act.	Survey Not implemented in Recorder	Keywords
Validation	How quality of survey is maintained and how records are validated.	Survey Not implemented in Recorder	Keywords
Volumetric	Number of records, recording events and samples in the survey dataset. Not in Recorder but part of Metadata standard.	Survey Not implemented in Recorder	System calculated but sometimes supplied by users [may not tally!]
Record Management	Details of how the data are managed e.g. how and where cards are stored. If transferred to a database. When copied or archived. Not in Recorder but part of Metadata standard.	Survey Data Management Not implemented in Recorder	Keywords

8.1.2 Survey Event

Every survey, however defined, is made up of one or more discrete recording events which may be interpreted in a number of ways; for instance, an event may represent the efforts of one or more recorders at several locations on one day or it may represent efforts at a single location over a period of days. The controlling factor will be the common information, which links all the observations together. Survey event information includes the recorders, the date, the weather and the survey methods used. The locality at which observations are made can be linked to the Survey, Survey Event or Sample information but practical build reasons make the **Survey_Event** the most convenient, with perhaps sub-locations (if any) attached to the Sample as this allows for several locations or sub-locations to be linked within one event (e.g. several quadrats in one meadow).

8.1.3 Survey Recorder

This is a link entity between the Survey Event and the Name Entity (People & Organisations Module) and allows multiple names to be attached to a record or event. The recorder could be any valid name e.g. a person or an organisation (useful for some published records). The link might include information relating to the individual's role in the Survey Event so that it might be clear from group efforts who was responsible for covering plants, insects, birds etc.

8.1.4 Survey Sample

This entity links any specific data, which might be collected during a recording event to the overall recording event and general survey information. The survey sample can be made at any level from a general link between a location and a list of taxa (e.g. to create a species list for a 1-kilometre square) right down to the detailing of the contents of a pitfall trap. Survey Samples may be linked through a **Survey_Sample_Relation** entity (not shown in *Figure 19*) to enable the recording of, for instance, the positional order of samples in a transect. Survey Samples may also be related in time, which would enable time series observations to be recorded or for repeated surveillance and monitoring records to be associated.

The Survey Sample entity may have a link to location and conversely could be a link from the location module into the biological (or earth science) data recorded for that location. This link would also be the way that most repeatable physical data about the location would be recorded. This is clear when physical data likely to change are considered (e.g. soil or water pH) which allows the management of measurements used for monitoring purposes.

Instances of observations are made in relation to a Survey Sample record. This is useful device because it allows for the recording of any mixture of biotope and taxa records. This includes biotope land-cover data where taxa records are subordinate such as in Phase I Surveys through to site taxon lists where habitats are not noted. It also allows for the detailed recording of taxa for very detailed habitats such as the NVC classes used in Phase II surveys.

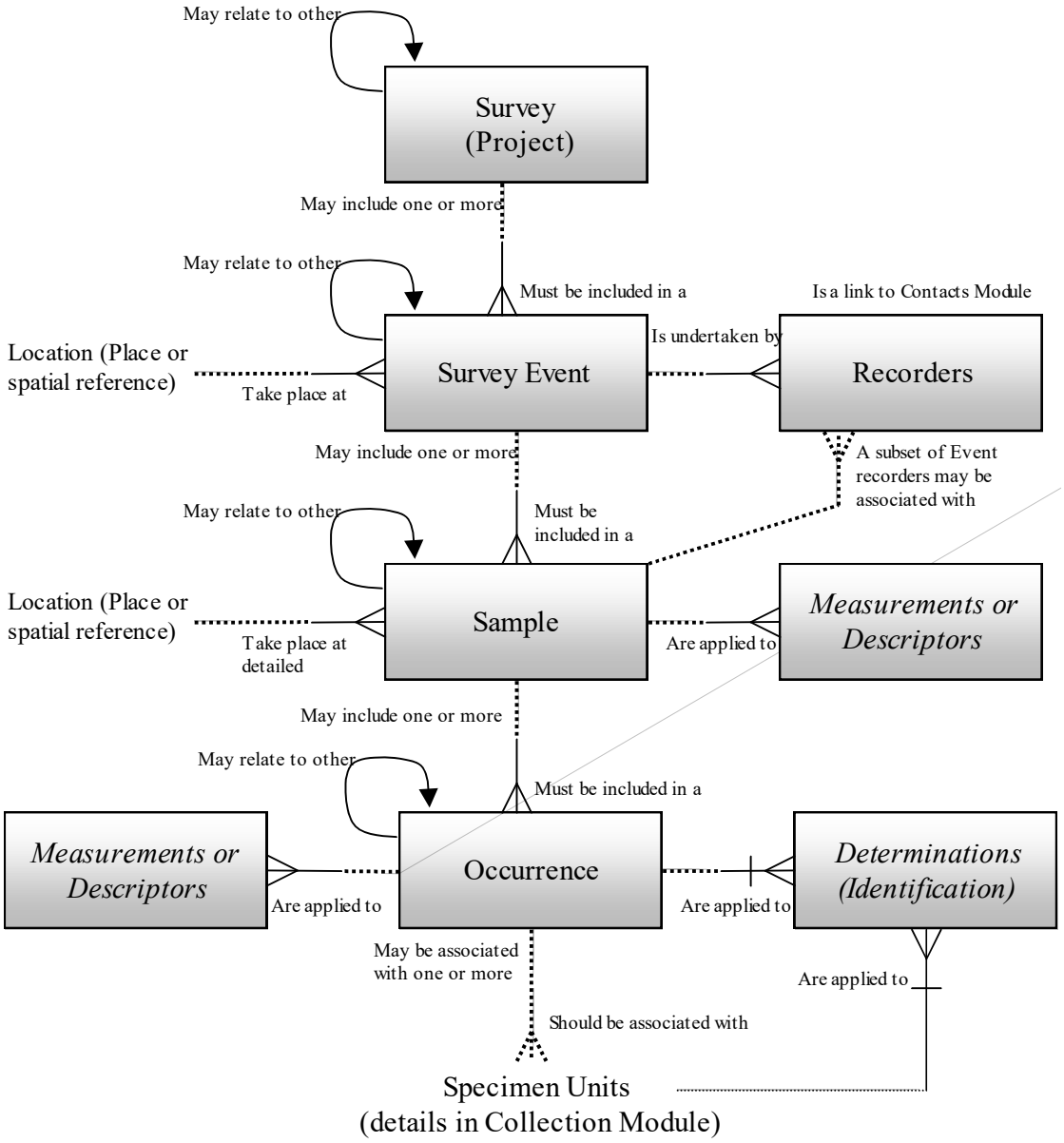


Figure 19: Simplified LDM for the Survey Module

Table 11: Information related to Survey Events and Samples in the NBN Data Model and in Recorder

Survey Event and Sample Information	Definition	Entity	Validation

Recording Event Date	Date (or range of dates) upon which this particular recording event within the overall survey took place. Individual samples may have their own date within a range. Can have both event and sample dates and times.	Survey Event & Sample	Valid date – Recorder uses Vague Date format. If Event has a range of dates, Sample date should fall within range
Location	The recording event may relate to one or more geographical locations, including sites and sub-sites. [i.e. Recording events are themselves divisible into discrete sampling events.] Can have links from both event and sample to Location Module.	Survey Event & Sample (also linked to Survey in Marine Recorder)	Valid location keys
Sample Scale	A cue to the scale that the sampling refers to e.g. whole site, subsite, community, stand , transect, quadrat etc. Not in Recorder but may be recorded as Sample Type or entry in comment field.	Sample	Controlled terminology list
Sample Area	measured area for the whole sampling location e.g. area of land parcel or size of quadrat	Sample Data (also Survey Event Data in Marine Recorder)	In appropriate measure e.g. square metres or hectares.
Recorder's Locality Reference	Any reference number or name given to the sample site by the recorders e.g. quadrat number or informal subsite name (Sample attribute Reference)	Sample	Individual applications may use a checklist or create sample references from a combination of site number and running number.
Grid Reference	Detailed grid reference for the sample site. Both the Survey Event and the Sample can have a spatial reference (e.g. OS grid ref.) associated with them as well as being linked to a location (optional).	Survey Event & Sample	Valid grid reference [may apply conversions to other referencing systems] Recorder converts all spatial refs to Lat/Long as well as storing original refs.
Grid. Ref. Source	Source of the grid reference e.g. original map reference, original GPS reference, subsequent GIS ref., inferred grid reference.	Survey Event & Sample (as Spatial Ref Qualifier)	Controlled terminology list. .
Weather	Record of the weather conditions at the time - free text	Survey Event	
Observation Period	Time spent collecting/observing. Survey Event has Start and End Dates, Sample has dates and a Duration attribute.	Recording Event or Sample	Recorder uses Vague Date format
Survey	Which survey this event belongs to.	Survey Key in Recording Event	Valid survey key.
Survey Type	This is necessary if the information is not available under the survey heading e.g. in the situation of general records from a natural history society or individual naturalist.	Recording Event But not in Recorder	Controlled terminology list
Survey Method	Sampling method used for an individual Recording Event or sample. e.g. Quadrat, Malaise trap, Satellite TM. Sample Type attribute in Sample	Sample	Controlled terminology list
Recorders	Individuals involved in collecting samples or making observations. The Survey Event has a list of Recorders associated with it and any subset of these may be linked to an associated Sample.	Survey Event Recorders & Sample Recorders	Valid name keys
Text References	Any publications, manuscripts, letters or agreements referring to this recording event. Can be any Source material including images and web pages, linked through Source.	Link to Sources	Valid reference keys
Related Samples	Within a survey event it may be necessary to link certain samples e.g. quadrats within a single biotope or traps in a trap line.	Sample Relations	Valid sample keys
Altitude/Depth	Specific altitude or depth measurement related to sampling activity. All Measurements and descriptors are covered by the measurements common entity.	Sample Data	Usually in metres but can be validated against thesaurus

Temperature	Specific temperature related to sampling activity. All Measurements and descriptors are covered by the measurements common entity	Sample Data	Measure dependent units
Relative Humidity	Specific rH related to sampling activity. All Measurements and descriptors are covered by the measurements common entity	Sample Data	0 - 100 %
Physical Data Measurements relating to the Locality	All Measurements and descriptors are covered by the measurements common entity which may be linked to any appropriate module and entity.	Sample Data	Usual checks
Comments	Free text descriptive data	Survey Event & Sample (No comment attribute in Recorder Survey Event)	None

8.2 Occurrence Records – observations and specimens

8.2.1 Occurrences

An occurrence is an instance of an observation which may also relate to the collection or gathering of one or more specimens. In the NBN model occurrences are linked to samples because this allows the relating of common information (e.g. the exact location, method, recorder, duration etc.) to be linked to a list of individual occurrences. The occurrences are most frequently, taxon occurrences (e.g. at this point I saw this list of species) but can also be linked to the habitat, geology, soil, minerals or any other observable or collectable items. In the BioCise model and in the ABCD schema, occurrences and their associated specimens are collectively referred to as **UNITs**[31]. The NBN model and Recorder physical implementation recognises the value of UNITs but separates field observations (OCCURRENCES) from COLLECTION_UNITS (which can be specimens, collections or stores) because it is easier to organise the recording and management applications in this way.

The systems analysis for the Recorder 2000 build described several types of occurrence including several earth science types including fossils, minerals, rocks, soils and stratigraphic occurrences. In the original and subsequent builds of Recorder, however, only taxon and biotope occurrences were included. The development of the Collections Add-in has now extended the ability of Recorder to store occurrences related to any recordable 'feature' through a generalised occurrence form and links to the new Thesaurus.

Figure 19 shows the generalised form of occurrences in the NBN data model as a single entity, although in Recorder 2000 there are separate tables for biotope and taxon occurrences. Logically, there is an occurrence entity, which may be one of any number of subtypes) and each occurrence entity may be linked to any number of measurements or descriptors (which are recorded using the measurement/descriptor common entity). Occurrences may also relate to other occurrences in the same sample or to occurrences in other samples.

8.2.2 Biotope Occurrence

The Biotope Occurrence entity one of the two types of occurrence supported in the core Recorder application. The Biotope_Key in the occurrence record is a link to the Biotope Dictionary, which covers all the various landcover and habitat classifications that might be recorded. This could also

[31] In fact, UNITs can also be whole collections or parts derived from other units.

include more informal descriptive and land-use classifications which would allow users to record hay meadows, ridge and furrow or ancient hedgebanks. In some applications it is important to know who actually identified the habitat particularly with some NVC types. It may be necessary to change the identification at some point but not lose the original record. This is achieved by linking occurrences to determination records in the **Biotope_Determination** entity.

Descriptive information relating to a single biotope or landscape type would be recorded in the **Biotope_Occurrence_Data** entity. Typical information might include minimum and maximum sward height, management keywords or damage records. For instance, in a Phase I Survey each landclass for a site might be recorded using an RSNC/NCC term and for each of these individual landclasses there might be a set of use, management, threat or damage notes and comments in addition to a more wide-ranging target note.

The model allows for several biotope terms to be linked to a single sample, which would be useful where the available information only allows a listing of types for a site or where overlapping or different classifications might be required. It should be noted that the calculation of areas for biotopes and sites would be an application problem not a data model one.

8.2.3 Taxon Occurrence

In core Recorder, taxa relate to the Survey Sample in the same way as biotopes. Thus it would be possible to make a list of taxa for a location (however it is defined) without reference to biotopes or a list of taxa could be provided for a specific habitat within a location. The **Taxon_Occurrence_Data** entity allows for the recording of facts about any taxon observation e.g. sex, stage, number of individuals (see below). The **Taxon Relation** entity allows for recording associated species, for instance a parasite and its host and even a parasite on the parasite!

8.2.4 Taxon Occurrence Data

There are many attributes, which users may wish to record about taxa, some of these such as sex and life stage are catered for in the existing version of Recorder but others are not. In particular there are some types of recording such as bird ringing with associated measurements or bird nesting records with clutch size, hatch rate and fledgling success. These more specialist classes of information could be treated as subtypes of the taxon occurrence data entity and as such could be built into Recorder or added as a third party applet. Each subtype, as illustrated in figure 11 above would have its own specific attributes and use controlled terminology as appropriate for instance, evidence of breeding would include singing and ovipositing whilst seasonal form would include summer plumage, winter coat and eclipse.

The name applied to a taxon observation is recorded in the **Taxon_Determination** entity which links to the Taxon Module. This arrangement allows for any number of names to be applied to the observation. This may arise through disagreement over the identification, reclassifying following a taxon split or use of a different synonym. The Taxon Determination would include the determiner, date and checklist used.

8.2.5 Generalised occurrences in the Recorder 6 Collections Add-in

The Collections Add-in to Recorder 6 extends the number of occurrence types that can be linked to a sample beyond the basic taxon and biotope occurrences. This has been achieved by adding a new generalised occurrence table to the database (and a form to the application) and linking it to the new thesaurus to pick up a concept_group (list) of available occurrence types. The list of occurrence types can be extended as necessary and determinations associated with occurrences can be validated against term lists in the thesaurus. In the extension built for the Luxembourg Natural History Museum, several new occurrence types relevant to geology (e.g. minerals, rock names, stratigraphy, palaeontological taxa, and soils) have been added. Adding further types of record would be simply a matter of adding it to the occrrence types list and providing appropriate term lists to the thesaurus. The new occurrence record is linked to the extended occurrence data tables (and application forms) that allow descriptors and measurements to be linked to observations. Details of the table and attribute structure for the generalised occurrence tables are given in Part 2 of this work.

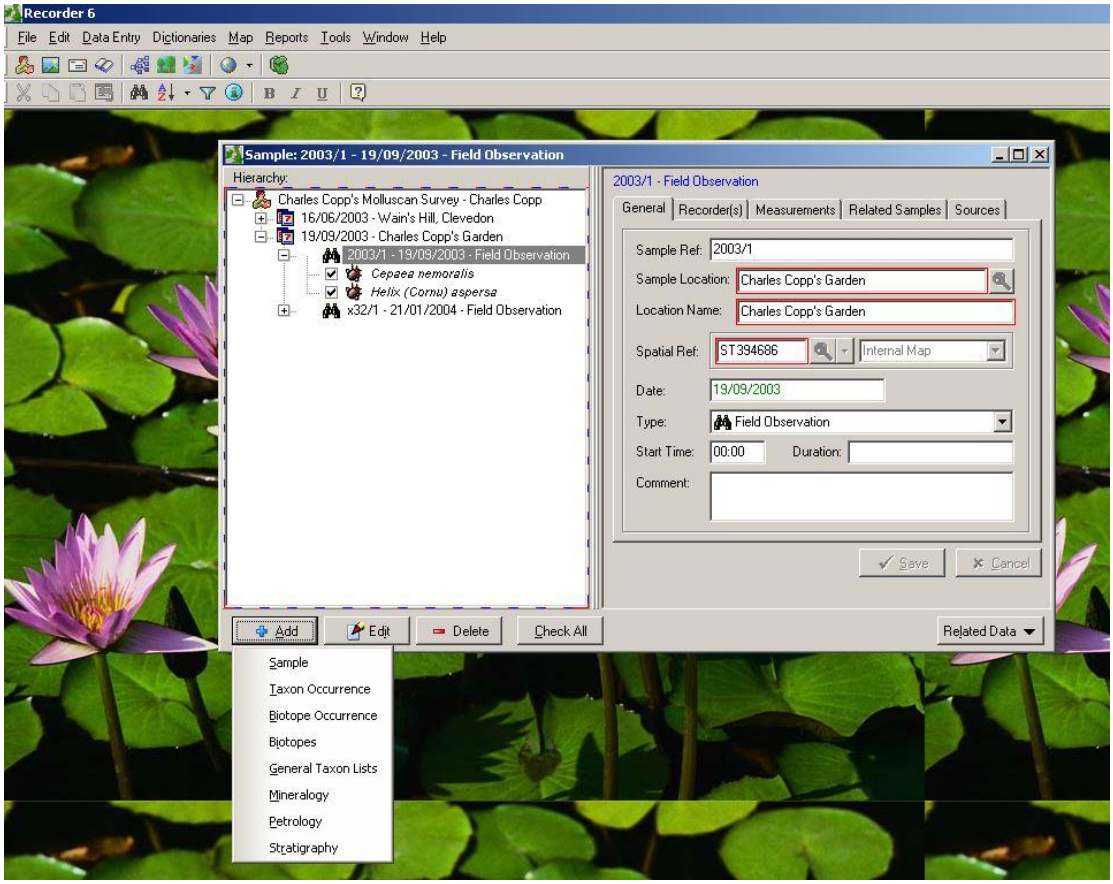


Figure 20: Observations form in Recorder 6 with drop-down menu for adding new occurrence observations. The double entries for taxa and biotopes reflect the use of either the separate taxon and biotope dictionaries or the new integrated thesaurus – an issue not resolved at the time of writing.

8.2.6 Biological field records in the NBN Survey Module

The following common survey types appear at first sight to be quite different but can be readily mapped to the NBN Data Model:

- **Standard 1K square or tetrad recording cards**
  Many local floras collect information about common species on a square basis, with the same card being used for a whole year. Ideally all records should be collected with detailed grid references and details of date and recorders which, in the model, would have individual entries in the survey_event and sample tables for each combination of date and recorders. Where these details are not recorded, each square becomes a location and each card represents an event and a single sample (which might be one year or more of data). In this second instance the reusability of the data is clearly less than the first option.

- **Single Site Visit Cards**
  This would include many of the basic 'BRC-style' record cards, which are based on single site visits. Older BRC cards have a single box for habitat whereas more recent examples (e.g. RA19 - Marine Isopods) have tick boxes for habitat, microhabitat, collecting technique, recording conditions and other information. Single visit cards represent single events and usually a single sample although differing grid references against taxon observations can become different samples.

- **Species Record Card for a Site**
  Many recorders maintain record sheets for sites, which include taxa seen on various dates e.g. species, date, and number seen. These are the sort of recording media commonly used to feed into monthly bird reports. If dates and details are entered then these can be extracted as events and samples otherwise the card represents a single event and sample with a longer time range.

- **Phase I Land-cover/land-use records**
  Phase I surveys are locality-based with the locations being a mixture of named sites and groups of one or more land-parcels which are related by a common land-cover or land use. For any 'cover'-type a number of associated facts may be recorded e.g. damage, threats, management and use. Some Phase I surveys also record species lists for interesting 'sites'. Dated observations of the occurrence and area of a named habitat constitute a biotope occurrence in a sample and the date marks the event. Information on damage, threats or notes on management relate to the biotope as a feature of the location. There are thus two types of record present.

- **Phase II Survey - Quadrat records**
  Phase II surveys are usually site-based. The site will be visited on one or more days and the site described in terms of biotopes/habitats, stands within those biotopes and detailed quadrat records within those stands identified by detailed grid references. Each quadrat or sample will include a list of taxa identified and accompanying measurements such as species dominance or sward height. The site visit becomes the event and each quadrat (or pitfall trap etc) becomes a different sample within the event.

- **Repeated surveillance and monitoring records**
  Repeated surveillance and monitoring records link observations together over a period of time with measurements and observations linked to set criteria or management priorities. For instance, in the case of a particular habitat on a site (e.g. *Calluna vulgaris - Ulex minor Heath* to see if it is spreading or decreasing), the data on area, sward height, grazing damage can all be linked to different survey event / sample records to give a time series. In this case the thing being monitored becomes a feature of the

location against which criteria and aims can be set whilst individual observations are samples within events that are linked to the feature.

8.3 Specimens

Biological recording applications have traditionally only sought to deal with field data that includes determinations whilst museum cataloguing applications have tended to concentrate on the physical specimen and be weak in the area of field data. It is logically possible to integrate both of these forms of record using the extended NBN data model. Specimens were included in the original analysis for Recorder, which lead to the development of the NBN Data Model but were not elaborated in detail.

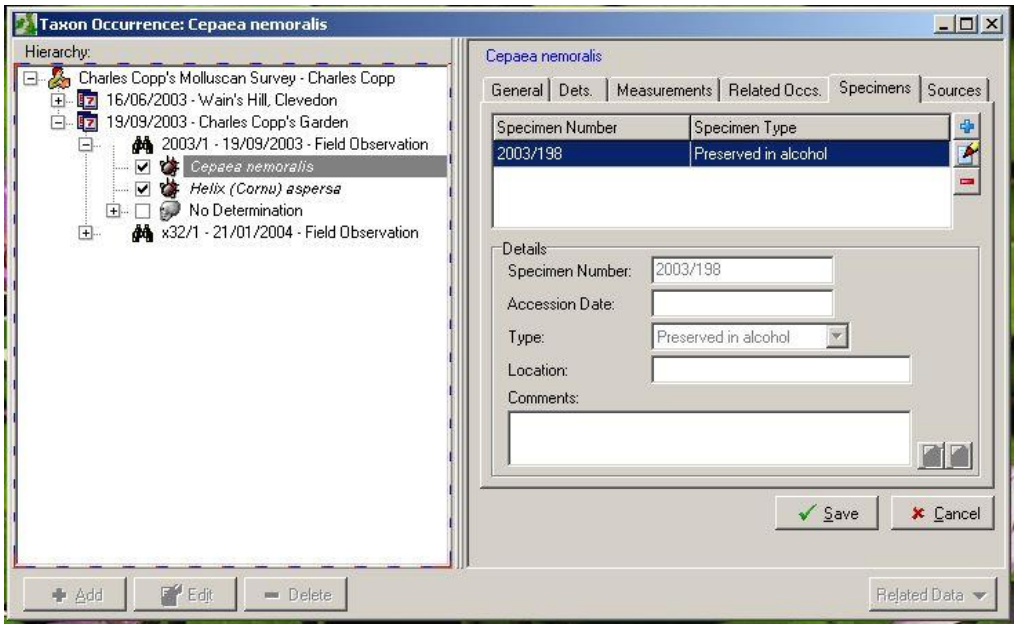


Figure 21: The Specimen Tab for Taxon Occurrences in Recorder 2000/2/6. Note the limited amount of information recorded - Registration Number, Date Accessioned, Storage Location and a comments field.

The core version of Recorder allows only minimal data to be recorded relating to specimens (see *Figure 21*) but the model has now been extended as part of the development of a Collections Add-in for Recorder 6 funded by Luxembourg National Museum of Natural History (see *Figure 22*) and further described in section X and Part 2 of this report. *Figure 21* shows how a specimen can be linked to a large number of data entities (folders displayed beneath the species name in the left-hand panel) and any entity will have several tabs of data (right-hand panel). Double clicking on any folder would reload the tree for that view of the data (e.g. double clicking Accessions will open the whole Accessions 'Register' at the point for the current specimen.

Including specimens into biological records applications can present a number of problems for data management and reporting. The specimens may not yet be identified although their location, date and collector are known, so that they represent a valid record (and voucher). Things can get complicated if the specimens originally lumped together as one record turn out to represent different species. In museum collections the specimens may be identifiable and valuable but other important parts of the record might be missing (e.g. a large lump of gold of unknown provenance is still a large lump of gold and future geochemical tests could determine its geographic origin).

In core Recorder the basic details of specimens are recorded in **Taxon Specimen**, which is linked to Taxon Occurrence (as in *Figure 20*). Identifications and facts (**Taxon Occurrence Data**) would normally be linked to Taxon Occurrence but the inclusion of the Collections Add-in allows alternative links from specimens to facts and determinations so that, for instance, individual specimens may be described and, if necessary, determined differently from the original field record. More traditional museum information relating to storage and conservation would be linked to the Specimen (or as it is called in the model **Collection Unit**) entity.

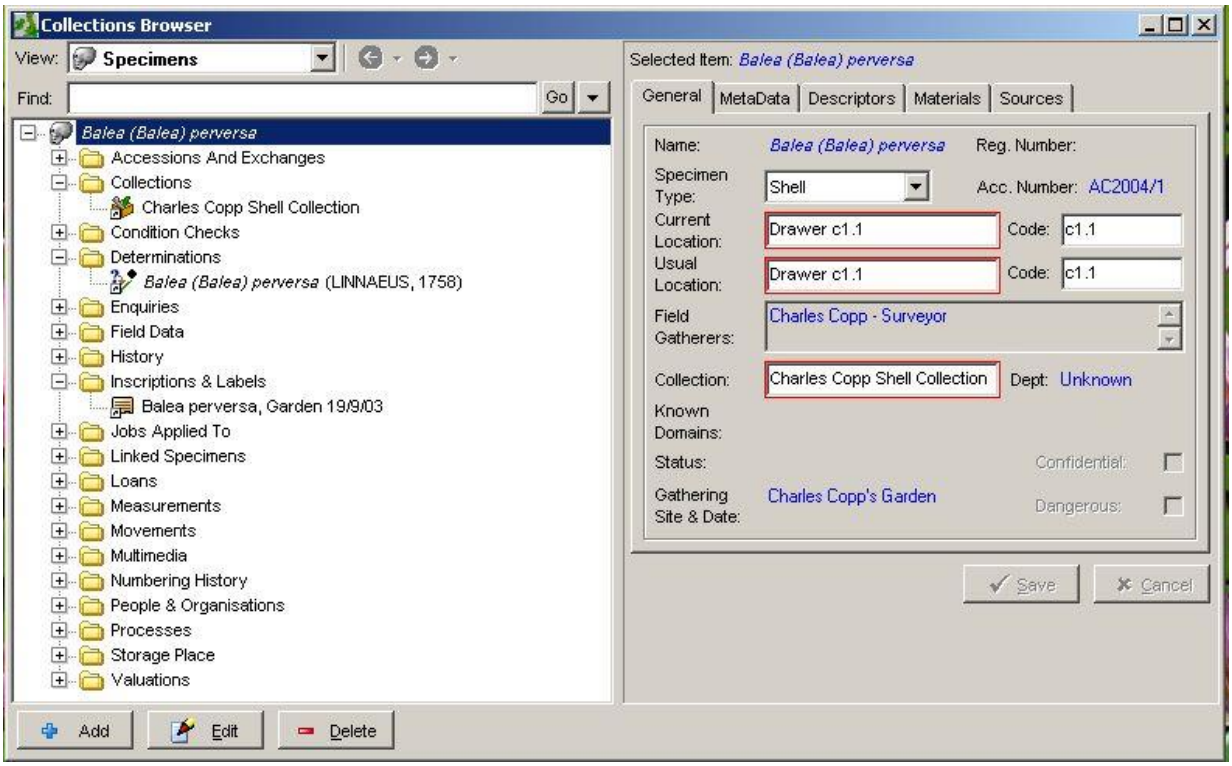


Figure 22: Specimen tree in the Recorder 6 Collections Browser (the Collections Add-In).

9. Location Module

In most environmental applications a distinction is drawn between sites and administrative areas. In many, a further distinction is made between ordinary sites and protected sites, for instance the JNCC Geological Conservation Review Database[32] used three tables to refer to administrative areas, GCR Sites and SSSIs. The original version of Recorder (3.3) used separate tables for Sites and District/Parishes, with Counties listed in a Codes table. Such arrangements are used to split up the information into easier managed units and to cope with the different attributes, which might be recorded about each type of location. In principal, however, there is no fundamental difference between administrative areas, sites and protected areas, they are all geographic areas identifiable by a boundary polygon or a grid square drawn on a map. This argument is true for any spatially referenced entity including sub-sites, grid references and transects.

It could be argued that all biological records should be linked to detailed grid references (spatial references) and that association with sites or administrative areas can be derived through the use of Geographical Information Systems (GIS). In the UK, it is now common practice to record 6-

[32] See http://www.jncc.gov.uk/earthheritage/gcrdb/background.htm

figure OS grid references for records but in practice, both spatial access and site name access to data are needed. Named sites figure highly in focussing recording effort, conservation action and protective designation. Digitised boundaries and the software to handle them are still not generally available to the majority of field recorders. The use of site-based data and the concentration of interest on named sites is therefore likely to remain a permanent feature of biological and earth science recording and forms a fundamental part of the NBN Data Model.

In practical terms it is sensible to make a distinction between sites that are linked to field or specimen records and gazetteer entries that supply controlled terminology and contextual hierarchies for place names (e.g. districts, counties and countries). In the context of a general biological recording model it is possible to argue that users may wish to record data for any spatial area, for instance, an old literature record may give only the vice county for a species. This is not precluded by the model, any named area or place can appear in both the Admin. Dictionary[33] and the Locations table. The model illustrated in *Figure 23* demonstrates how named sites (**Locations**) can link to administrative areas (actually any term in the gazetteer), to the survey module and also provides a means for storing information about key features for which threats, damage, facts, actions and plans may be recorded.

The way that the model is currently implemented is that a site in the locations table does not have to appear in the gazetteer although it might, for instance if the site were an SSSI[34]. Place names in the gazetteer are primarily used to provide text-based context and sorting for places of interest (e.g. sort all locations by Vice County). Individuals or local networks tend to create and define their own sites in the Locations table and keep the gazetteer for 'official lists' and contextual areas. Established and agreed local lists of sites can, of course, be added to the gazetteer if desired. For most current Recorder Users, using the Locations table works well except in cases where numerous 'local sites' are created for distribution to Recorders (to encourage consistent site recording). In Luxembourg, for instance, this created a huge list of over 17,000 sites, which populates the Location window regardless of whether there are associated observations (See *Figure 24*).

The purpose of the Location Module is to provide the means of managing information about 'sites' of interest to the system i.e. ones that have or are likely to have field observations or specimens linked to them or those with features subject to local management and monitoring. Sites can have sub-sites to any level and so can be used to create complex site hierarchies but the Location Module is not intended as a general gazetteer, rather, the data collected extend that available in the general gazetteer. The key entity is **Location**. In Recorder the Location table includes minimal data because so much of the information about sites may be multiple (e.g. alternative names for a site) which in a relational database is stored in separate tables. *Table 10* (below) illustrates the data that the NBN Data Model relates to Locations and how they are represented in both the Logical Model (*Figure 23*) and the Recorder Application (see Part 2 of this work).

[33] Core Recorder uses a separate Administrative Area Dictionary for geographic names whereas the new developments for the Collections Add-in allow for the use of an integrated Thesaurus, in which geographic place names belong to the Gazetteer Domain. For the sake of simplicity the word 'gazetteer' will be used to denote either in discussion of the Location Logical Model.

[34] The Admin. Dictionary and the Gazetteer Domain of the Thesaurus both hold multiple lists of place names e.g. UK SSSIs, UKCounties, districts and Parishes, Watsonian Vice Counties etc.

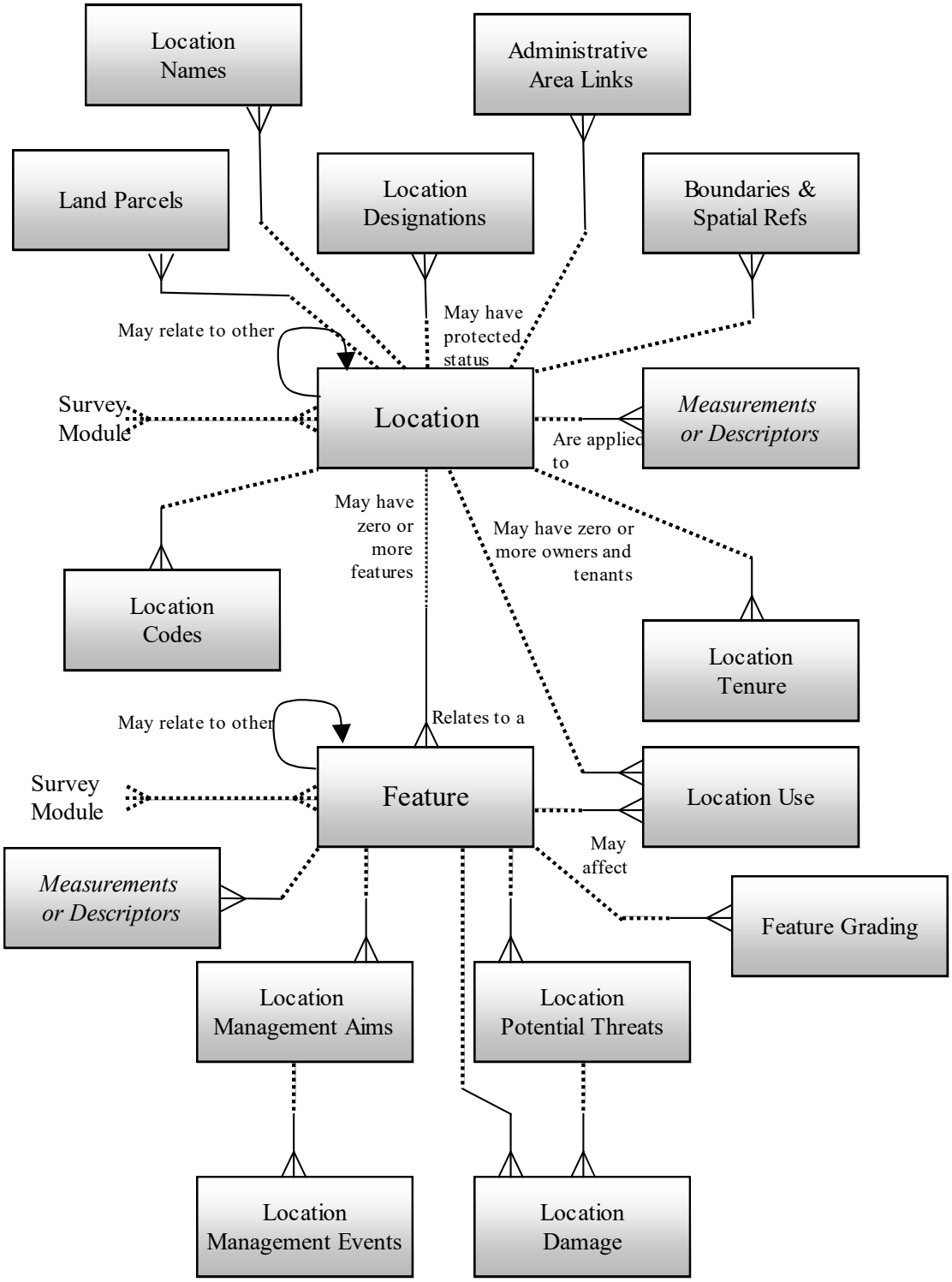


Figure 23: Simplified LDM for the Locations Module showing the Location Features sub-module

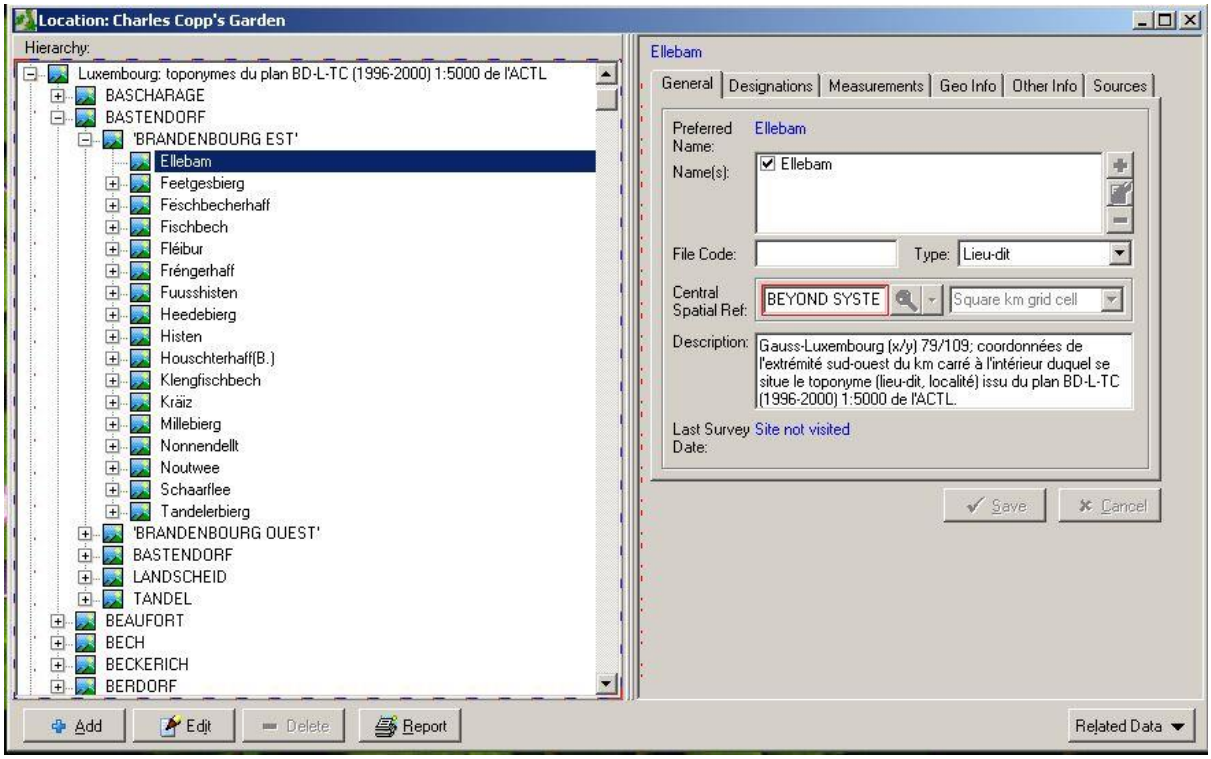


Figure 24: The Locations Window in Recorder, showing part of the Luxembourg list of locations. Note the various data tabs, including Sources, in the right hand pane.

(The erroneous grid reference message is due to changing the system base map to UK parameters!)

Table 12: Information attributes of Locations and how they are represented in the NBN Data Model and Recorder

Location Data	Definition	Logical Entity	Validation
Location Name	The name given to the site or region. Some locations may have more than one name or various versions of the name (e.g. spellings) at different dates.	Location Names	• Designated site names may be validated against a supplied dictionary of protected sites and areas. • Administrative areas may be validated against a system supplied dictionary. • Local site names may be validated against a locally agreed and managed list of site names.
Grid Reference (Spatial Reference)	In most systems it is convenient to record a centroid grid reference which may be used for locating the site on a map or depicting the location in simple map-based output. Many users wish to enter more than one grid ref. for linear or angular sites. In Britain, Grid reference is normally expected to be a 6 or 8 figure UK or Irish reference. Systems should, however, be able to use other referencing systems including Lat. / Long and UTM at any level of granularity. Where grid references are given, a statement of precision can be valuable.	Location – can hold a single centroid whilst vectors and polygons can either be stored in a boundary table or stored separately with pointers stored in the boundary table. Some applications, including, Recorder also store a list of 1Km squares within which a site falls.	Grid references may be checked manually or system checked. Primary check for whether it is a valid grid reference then checked for context e.g. does grid reference lie within the boundaries of stated administrative areas (county or vice county)or grid square list. Recorder has algorithms for this. In Recorder all spatial references are stored both as original references and converted into decimal Lat/Long for data manipulation purposes.

Boundary	Few current database systems record boundaries and even where GIS are used accurate boundaries for all 'sites' may not be available. Eventually, however, all 'sites' should be linked to boundaries. One useful approach is to store a scanned image of a site plan, aerial photo or sketch	Location Boundary can store pointers to vector or polygon files whilst scanned maps can be stored as either internal or external references or included in linked GIS.	Database entry might be a pointer to a boundary file which may need a check that it exists. Validation of stored vector boundaries must be done through a Geographic Information System
Administrative Area	Administrative areas are locations in the same way as sites. They relate to each other in a hierarchical fashion and have a number of relationships (e.g. contains or overlaps). Administrative areas may have many versions with changed boundaries. It is usually found convenient to record certain fixed administrative area items for any site e.g. its parish, district, county and vice county but these may change.	Location (some users create Locations for administrative areas) or Admin Area Link as used in Recorder to provide a contextual text list of associated areas derived from an Administrative Area Dictionary	By use of system supplied administrative area dictionary. Validation and controlled data entry may be done by entering the most detailed administrative area and having the system dictionaries provide the higher levels of Admin. Area. In other examples, the highest level is entered and progressively more detailed 'popups' offer the lower levels (e.g. county ⇒ district ⇒ parish). This approach can been complicated by the introduction of new types of area e.g. unitaries and makes for an inflexible application.
Region	Regions may be dealt with in the same way as administrative areas although the user may wish to define their own recording regions and add them to the Admin. Dictionary.	Location or Admin. Area	As above or by provision of a scrolling popup sensitive to typed entries - as in Windows help applications.
Sub-sites	Sites may be sub-divided in many ways. Each is effectively a new location record and may have all of the attributes of the master site.	Location linked by Parent_Key or Location Relation	Checks in relation record for valid location Keys and selection of a relation statement from a controlled terminology list.
Land Parcel	Land parcels are a feature of UK planning use and are marked on detailed maps. A location may span one or more land parcels as marked on small scale maps but the boundaries may not necessarily be coincident. Logically a land parcel is another type of location and could be stored in the locations table although it is more convenient not to.	Land Parcel (allows a list of land parcel numbers to be associated with a site)	Implemented for text purposes in Recorder but likely only to be successful through GIS and is probably currently rarely used.
Planning Authority	Name of the organisation responsible for judging planning applications relating to the location	Attribute of Location (not in Recorder)	Usually none but could use controlled terminology list and it should be possible to provide relevant authorities linked to administrative areas in the Admin. Dictionary.
Area of site	The actual land area covered by the location, usually recorded in hectares but could be in any other system or scale (e.g. acres or sq. metres)	Location measurements & descriptors (Location Data Table in Recorder)	Application may provided algorithms for converting from one system to another e.g. acres to hectares. Implies the need to record the measurement system in use, although most databases assume or ignore this. Recorder employs the Measurement Common Entity model. Area validation may be validated manually or through GIS.
Location Type	Locations may be classified for many purposes e.g. to group them or mark their position in a hierarchy (e.g. county, district, parish). Sites may be grouped by interest, ownership or protected status.	Attribute of Location	Controlled terminology list. In Recorder there is a Location Type table.

Description	Virtually all site or location records require a free text area to give a general description and/or history of the site. In practice there may be several types of description required which need separating out. Examples are: • General description • Legal statement (as in the SSSI legal statement) • Geological description • Simple explanatory note suitable for brief reports etc. For some applications it may be important for the same type of description to have more than one dated version.	Attribute of Location or separate table for multiple versions. Recorder has a single description field in Location table but there are other descriptive attributes linked to other entities e.g. features.	Spell checking Date and Author checking Some information provided with Admin. and Protected sites dictionaries.
Boundary Type	Boundary features e.g. hedgebank	Attribute of Location Boundary Not in Recorder where, for instance, a hedgebank might be made a subsite.	Controlled terminology list in thesaurus
Maps	In text-based system it is common to record the scale and reference number of any maps which depict the location. The commonest referred to are O.S. 1:50,000 and 1:10,000 series but any map or plan including topographic, geological, soil or land-use is equally valid.	Internal or external Source but in Recorder a site can be linked to a specific map location on the internal map.	Not usually validated but possible check against an existing list e.g. as entries in a reference file.
Conservation Status	Conservation status (e.g. SSSI, SINC or RIGS) is commonly associated with named sites although a number of status types are applied to larger areas (e.g. AONB, ESA and RAMSAR). Individual locations may have more than one status accorded them both simultaneously and over time. Boundaries of areas of specific status may contain one another, coincide or overlap and status may be removed. Details of conservation status should therefore include a start date and where appropriate an end date.	Location Designation	Validation against a dictionary of nationally (and possibly locally) recognised lists of sites and areas within each status type. Controlled terminology lists for input where checking is not done against a provided dictionary. Check that end date is after start date.
Location Code	Locations, especially recording sites, tend to have several classification numbers associated with them e.g. the file reference number for storage purposes, unique database number, Country Agency SSSI number, GCR number etc.	Attribute of Location also attribute of Conceot in the Thesaurus.	Not usually validated other than to check that code is in the right format and where a new one is expected, a check that it is unique. Some numbers e.g. SSSI numbers could be validated against a provided dictionary.
Measurements Length & Breadth	The main proportions of a site can help to visualise it and to more readily recognise it when looking on a map. Normally expressed in metres but could be in other units.	Location Data (Measurements & Descriptors) & Sample Location Data in Recorder	Imperial to metric conversion Implies the need to record the measurement system in use, therefore use of common measurement entity, (although most databases assume or ignore this).
Altitude Depth	For some purposes it is convenient to record altitude or depth for a location. This may be a mean or expressed as maximum and minimum. Normally expressed in metres O.D. but could be in other measures or to another datum. In the NBN model these are just further instances of Measurements.	Location Data (Measurements & Descriptors), Location Feature Data & Sample Location Data in Recorder	Max. > Min. Imperial to metric conversion Implies the need to record the measurement system in use, although most databases assume or ignore this.

Aspect (Slope & Direction) Exposure	General aspect of the location (its overall dip), usually expressed as degrees of slope and degrees from North. Other 'keywords' may be used to record aspect e.g. open or shady. In the NBN model these are just further instances of Measurements or descriptors	Location Data (Measurements & Descriptors), Location Feature Data & Sample Location Data in Recorder The Collections Add-In uses the extended measurements and descriptors common entity.	None usually for measurements although thesaurus can contain multiplicity data for measurements (e.g. 3 feet = 1 yard). controlled terminology list for descriptor 'keywords'.
Microrelief	Details of finer aspects of topography e.g. crack, fissures and drainage channels.	Location Data (Measurements & Descriptors), Location Feature Data & Sample Location Data in Recorder The Collections Add-In uses the extended measurements and descriptors common entity.	None usually for measurements although thesaurus can contain multiplicity data for measurements (e.g. 3 feet = 1 yard). controlled terminology list for descriptor 'keywords'
Physical Measures: pH Nitrates Pesticide Levels	A very wide range of physical measurements may be made at a location. These are all time limited and are therefore repeatable as physical observations.	Location Data (Measurements & Descriptors), Location Feature Data & Sample Location Data in Recorder The Collections Add-In uses the extended measurements and descriptors common entity.	To be defined as necessary Implies the need to record the measurement system in use, although most databases assume or ignore this.
Climate	Climate type prevailing on a site may be recorded. Data may be taken from climate maps.	Attribute of Location or associated Descriptor (not in current Recorder)	Selection from a controlled list of climate types.
Microclimate: Temperature Humidity	As above plus local data recordings	Location Data (Measurements & Descriptors), Location Feature Data & Sample Location Data in Recorder The Collections Add-In uses the extended measurements and descriptors common entity.	To be defined as necessary Implies the need to record the measurement system in use.
Soils	Soil information may be generalised for a site and derived from soil maps (and stored as a site feature) or it may form part of a detailed sample.	Soil Occurrence in extended model linked to Sample or as a Location Feature Core Recorder does not have soil records	Controlled terminology list of soil types in thesaurus.
Ownership/ Tenancy	Many organisations need to keep track of the ownership and/or tenancy of sites. A single named location or bounded area may have several owners and tenants and these are likely to change with time. Records, therefore require from and to dates. The Data Protection Act applies to the computerisation of this information.	Location Tenure	Check for person/organisation against the Name file. Check location against location list. Controlled terminology list of ownership/tenancy terms.
Biotopes/habitats	Biotope Occurrence occurs in Survey Module but a specific biotope e.g. a BAP habitat may be declared either as a Location Feature or if the boundaries are fixed, as a sub-site	Biotope Occurrence linked to Sample or a Location Feature	Valid Sample, Location and Biotope keys
Taxa	Taxon Occurrence occurs in Survey Module but a specific taxa e.g. a population of a BAP species may be declared either as a Location Feature	Taxon Occurrence linked to Sample or a Location Feature	Valid Sample, Location and Taxon keys

References	Any number of references may be relevant to a location either as a whole or relating to any of its attributes (e.g. ownership, management, geology etc.) See Source Module for details of text reference attributes.	Source Link to Internal Source – subtype Reference	Check for valid reference key Check for valid location key
Images	Any number of images may be relevant to a location either as a whole or relating to its attributes (e.g. panorama, specific feature etc.) Images may include slides, photographs, video, book-plates. It is possible to store actual images as well as references to them including URLs. See Source Module.	Source link to Internal or External Sources – subtype Image	Check for valid image key Check for valid location key
Features	A Location Feature can be anything that a user wishes to define for its interest and includes non-geographic items such as a population of bats or a biotope. Some features can also be listed as sub-sites e.g. a fixed biotope area. The key characteristic of a Location Feature is that it can have individual management aims, threats, damage or events linked to it.	Location Feature	Valid Location Key
Educational/ Recreational Use	This may be broken down into a number of different headings including: Present use, Potential Use. Location may be assigned a grading based on educational potential (e.g. RIGS grading). Should be qualified by a date and could therefore represent a separate entity.	Location Use or Feature Grading Location Use can relate to whole sites, sub-sites and to site features	May use controlled terminology lists e.g. use types and linked gradings.
Management Aims	Linked to Location Feature. Text statement of management aims for the location. Multiple aims as separate records. Subject to change - time related records	Management Aims linked to Location Feature	Usually none Spell checking could use a term list.
Location condition	Statement of the condition of the location. Should be linked to a feature of interest and a Management Aim (e.g. aim is to maintain quality of habitat and condition statement is linked to that as an attribute of a monitoring event)	Location Management Event linked to Location Management Aims linked to Location Feature	May use controlled terminology lists
Threats to location	Linked to Location Features. Statement of the perceived threats to the location. May be resolved under a number of headings - time related observations	Location Potential Threat	May use controlled terminology lists
Damage to location	Linked to Location Features. Statement of recorded damage to recorded features of the location. May be resolved under a number of headings - time related observations	Location Damage (Damage Occurrence in Recorder)	May use controlled terminology lists
Land Use	Current use of location - often more than one use and time related	Location Use	May use controlled terminology lists
Management Methods	Management methods used at location. Subject to change - time related records.	linked to Location Features Not implemented but could be employed from term list linked to Location Management Event	May use a controlled terminology list Valid dates
Management Agreements	Links between organisations, people and location. Subject to change - time related records	linked to Location Features. Implemented in Recorder by a single attribute in Location Management Aims and allows reference to Sources (Internal and External)	Valid keys

Management Events	Applied to Location Features. Events can be anything from condition monitoring to 'scrub bashing'. May be linked to a Survey Event. Should include a link to who carried out the event. Many records for events on different dates.	Location Management Event (May also link by date to a Survey Event)	Valid location and personal name keys Valid dates
Management Restrictions	Details of any restrictions to landuse or developments e.g. SSSIs would have a set of PDOs.	linked to Location Features but not in Recorder other than as a single text attribute (Restrictions) of Location	May use a controlled terminology list
Surveillance Frequency	Suggested frequency for checking location condition or repeating survey e.g. once per year.	Attribute of Location Management Aims	None
Next Appraisal Date	Reminder date for revisit, survey or monitoring.	Attribute of Management Aims	Valid date - later than yesterday
Access Route	Text description of approach route and entry to site. This can be useful.	Attribute of Location	Spellchecking
Access	Details of permissions needed for access to site e.g. permit from local wildlife trust or seek permission from farmer at ...	Attribute of Location	Spellchecking
Facilities	For sites likely to be visited by parties a note of the facilities for parking etc. can be useful.	Location Facilities (not shown in figure 23) Not implemented in Recorder	May use a controlled terminology list
Associated People	People may be associated with a location in many ways e.g. recorders, wardens, owners, managers, referee etc. There is a need to record a link between sites and people and their roles.	Name Link at various places	Valid location and name keys controlled list of roles
Associated Organisations	This is logically the same as associated people	Name Link at various places	Valid location and name keys controlled list of roles
Geology Geomorphology Soils Sediment/ Substrate Hydrology (Water Features)	Taxon and Biotope records may require additional information on geology, geomorphology, soils and hydrology to put them in the correct context. Many users wish to record these aspects of sites in their own right. A number of conservation classifications pertain specifically to earth sciences e.g. GCR and RIGS.	Occurrence Not implemented in Core Recorder but Collection Add-in extension to Model Recorder allows any type of occurrence to be added.	Validated by controlled terminology supplied in the thesaurus.
Selection Criteria	Many conservation or Site use classifications select sites according to fixed criteria (or should do!) This attribute allows the recording of which criteria relate to the present location	Selection Criteria linked to Location Features (Not implemented)	Controlled terminology lists related to different type of status.
Site Assessment	Assessment of the quality of the location e.g. in relation to selection criteria	Location Management Events linked to Location Features	Keywords

It is possible to define multiple versions of a site by storing boundary changes in the linked **Location_Boundary** table[35] but this is regarded mainly as a job for the Thesaurus and the sites in the Location Module are selected 'stable' entities.

Relationships between locations are stored in the **Location Relation** table. Relationships include relations between versions of the same location and relationships between different locations e.g. Site A overlaps SSSI B, district C lies within county D, unitary E replaced district F. Relationships may also be maintained as **Concept_Relationships** in the gazetteer (part of the Thesaurus).

[35] See Part 2 of this work for a full definition of the tables utilised in Recorder

The NBN Location Module model is powerful in terms of building an application which includes a wide range of site data and allows information to be linked to both static site features and to visits and surveys. The greatest weakness of the model is that within an application the information managed for some sites e.g. SSSIs, might be very different from that for ordinary sites and different again for administrative areas. Furthermore each type may be managed and distributed by different organisations. This can be handled by separating the sites for which the user is actively collecting and managing data in the Location Module and all other contextual sites (e.g. administrative areas and protected sites) in the gazetteer. *Figures 25 and 26* illustrate how Recorder has implemented the model in a practical application.

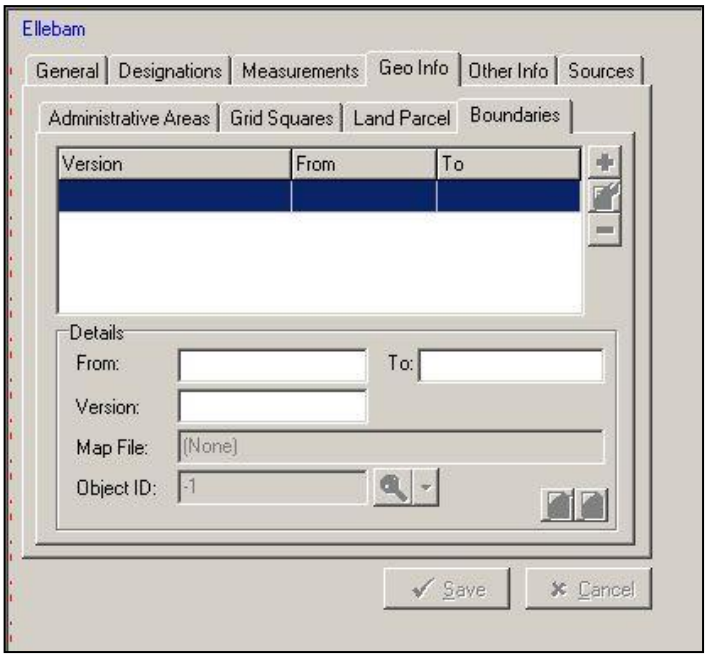


Figure 25: Data tabs in the Recorder Location Window. Note that some tabes (here the Geo Info tab) have sub-tabs handling related items.

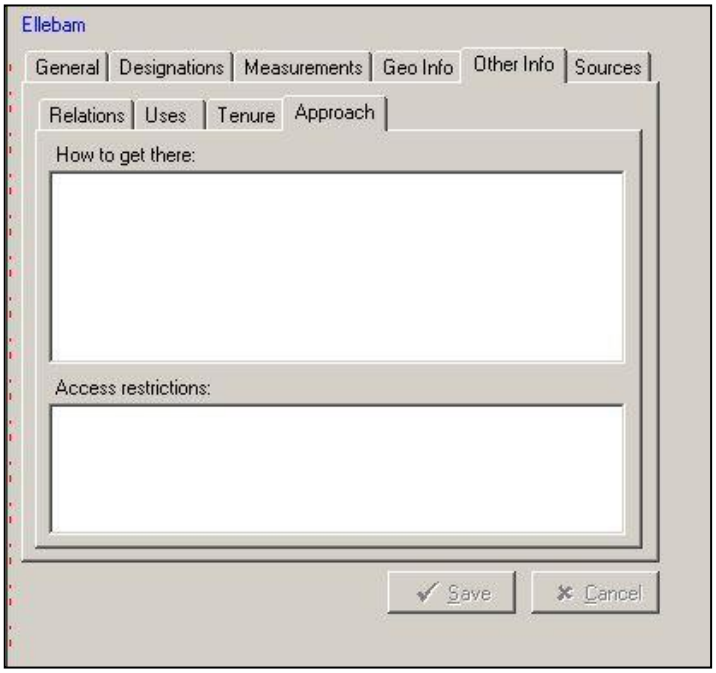


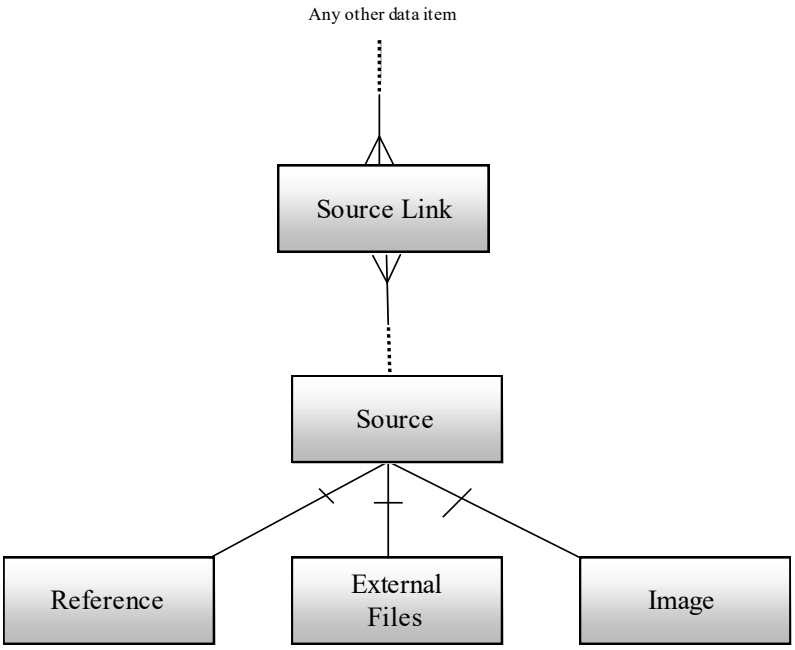
Figure 26: Locations in Recorder. The Other Info tab has sub-tabs for Related sites, Location Use, tenure and details of site access and restrictions.

There are a number of developing standards used for exchanging data between Geographical Information Systems and other applications and for describing geographical information resources. Most recent examples are based on extensions of Extensible Markup Language (XML) such as schemas using Geographical Markup Language (GML) or the Resource Description Framework (RDF). These schemas provide structured elements that can be used for describing the

machine rendition of geospatial features, locating items within spatial referencing systems and also elements that describe attributes such as accuracy and method in recording or digitising of spatial coordinates. The detailing of these attributes has not been fully developed in the NBN Data Model or in physical applications, such as Recorder, associated with the model.

10. Sources – References and Images

The source module provides a means of tracking the origin of any of the information in the system, ranging from a whole dataset to individual items of data in the dictionaries. There can be many types of source ranging from word-of-mouth, books and images, to metadata relating to collections of specimens. The information content of the source record may also need to change depending on what the source is referring to. This has been modelled using a simple **Source** entity which links to sub-modules for specific purposes. The two subtypes included in the Recorder physical data model cover **References (**called **Internal Documents** because the details are stored in the database**)** and links to other file types stored outside the database (called **External References**). In Recorder, external file types are linked to the system default applications for the file, so that, for instance, double clicking a pdf reference in a Source tab would load the highlighted file into Acrobat Reader. A sub-module for detailing images has been defined but not yet incorporated into any of the physical models or builds.



The Source record can link to the references, images and contacts modules through link entities as required. This means that, for instance, the source of a taxon record could be a reference and also a source can be linked to people with whom agreements are documented in the **Communications** table.

Figure 27: Logical Data Model for the Source Module with example sub-types

The source record can be linked by a Source_Key to any table and attribute in the database either singly as an attribute in the table or through a **Source_link** entity enabling multiple sources to be linked to a record. In the Recorder application, a Source tab is included on every record window as in *Figure 23*.

A recent development of the model for the Luxembourg Collections Add-In allows References (e.g. books and manuscripts) and Images to be treated as specimens. For instance, a photograph may be catalogued, stored and described and also linked to a digital copy as an external reference.

10.1 References

Text references refer to any written information including publications, manuscripts, letters and wordprocessor files. Many different attributes in the application may need text references attached to them. The Reference sub-module has a central entity (**Reference**) which includes attributes to match different types of publication and manuscript. In the physical model (Part 2 of this work) these occur in the one Reference table but they could be created as sub-types. It is be up to the application builder to provide the interface logic which displays only the attributes relevant to a single type of reference (e.g. the different fields needed for journal, book and manuscript references).

The model is essentially a very simple one (see *Figure 28*), which includes a dictionary of Journal and Serial names (**Journal**) to provide controlled data entry. **Reference_Number** allows various reference numbers to be associated with any publication or manuscript including shelf numbers, Library of Congress and ISBN numbers.

A **Publication_Keywords** entity can be included (not shown in *Figure 28*) because many users like to link publications to specific concepts such as a taxon and whether this is an identification text. Keywords are notoriously difficult to control and this area would benefit from a degree of controlled terminology.

References should be linkable to any other table and attribute in the database through a linking entity (Source_Link). This is incorporated in the Recorder physical model although in some tables a direct link to publication can been included as a marker that a reference is needed, again something which would be a build decision. [i.e. do you need complete flexibility of reference linking or can it be pegged down to a fixed number of essential points?]

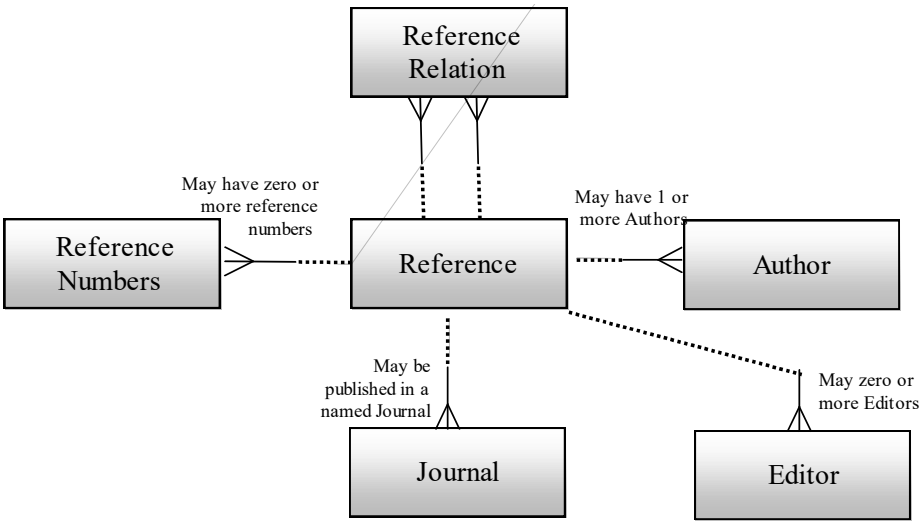


Figure 28: Logical Data Model for the References Sub-Module (note that Authors & Editors could be combined ina single Name Role entity)

Table 13: Information in the References sub-module

Publication & Reference Data	Definition	Entity	Validation
Reference Key	Unique identifier. Called Source_Key in Recorder	Reference	Unique key – In Recorder must use NBN 16 Char format
Authors	Can be implemented as a single 'field' using standard bibliographic format (e.g. Copp C.J.T.) even for multiple names although in Recorder a separate table is used for multiple authors with an attribute for position in the author string.	Reference Or Reference Author	format
Date	Date of text or publication date	Reference	Valid date – Recorder uses a Vague Date format
Title	Free text - title	Reference	None
Type	Type of text e.g. manuscript, book, serial publication. In Recorder attribute is Reference Type	Refwerence	Controlled terminology list
Serial	If the text is in a serial publication, link to subtype attributes including: Serial Name (from Dictionary), serial volume, volume part, page start and end, serial part, serial number, serial supplement. Recorder uses Journal_Key as a foreign key to the Journal Table.	Serial subtype. In Recorder attributes are all in Reference table and a FK to Journal	Valid key to serial publication dictionary (e.g. list of journals)
Serial Dictionary	A list of serial publications can be a valuable asset to a biological recording application. It would require the following attributes; Serial key, Serial name, serial abbreviation, start date, end date, publisher, country, links to associated people or organisation and ISSN. Recorder has a Journal Table recording Long Name, Short Name,and Description only.	Journal	Dictionary of serial titles and publication details.
Edition	Books, especially have edition numbers but wordprocessor documents may have version numbers.	Reference	None
Symposium Title	If the text is an article in a symposium volume this will include attributes for ; Symposium title and symposium editors. Recorder uses an attribute for Symposium Title in Reference table and lists editors in a linked Editors table.	Symposium subtype. Or Reference and Editors	None
Supplement	If the reference is released as a supplement	Reference	None
Volume	Volume number if any	Reference	None
Part	Part Number(s) if any	Reference	None
Pages	Number of pages or page range for the publication	Reference	None
Plates	Number of plates (could even use a link to images file for details of individual plates)	Reference	Valid key to images file (if used)
Figures	Number of figures (could even use a link to images file for details of individual figures)	Reference	Valid key to images file (if used)
Tables	Number of tables	Reference	None
Maps	Number of maps	Reference	None
Publisher	Name of the publisher	Reference	None
Publication Location	Where published. Recorder attribute is Place_of_Publication.	Reference	None

External Reference Numbers	A text or publication may be referred to by many classification or filing numbers e.g. ISBN, ISSN, Library of Congress, Shelf number etc. This entity could also store a wordprocessor file name and directory. Attributes in Reference Number Type table or term list include Number Type and description.	External Reference Number	Valid key

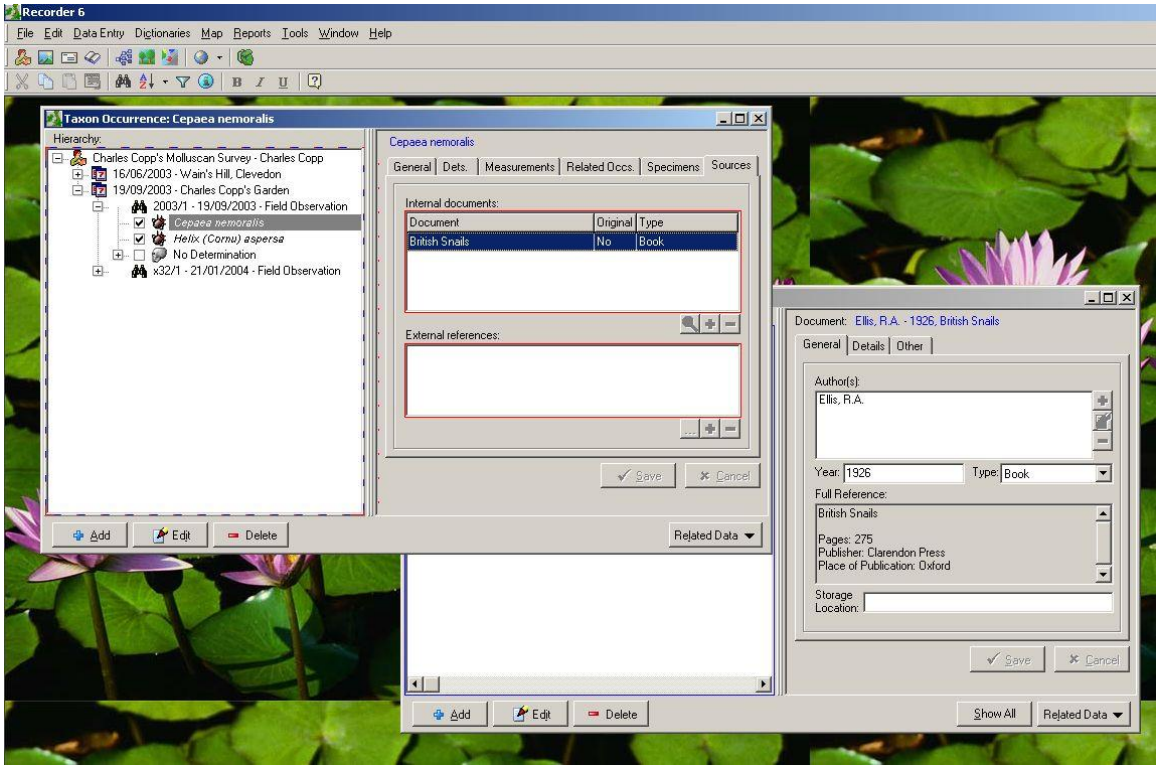


Figure 29: The Reference tab for an observation record in Recorder with associated reference detail window

References are dealt with in a very simple way in the ABCD Schema (*Figure 30*). The ABCD schema is used for transmitting data in response to database queries and for merging responses from heterogeneous sources. Recorder could easily provide Reference data in this format as a concatenated output from the Reference sub-module.

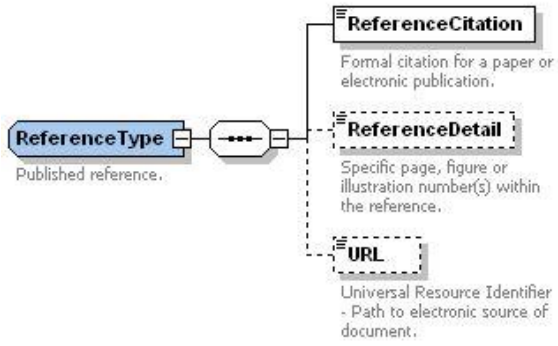


Figure 30: The ReferenceType Complex Type entity from the ABCD Schema

10.2 Images

An image module could become important in future extensions of applications based on the NBN data model, as machines become better able to store and display electronic images and pictures are integral to any attractive web-based delivery software. Since the analysis was carried out for the Recorder Data Model in 1998 there have been major advances in the development of Digital Collection Management models, linked particularly to the delivery of images through web portals. These developments include metadata models for describing a wide range of resources (including images) such as the Dublin Core Project[36] and the development of new image databases fronted by sophisticated ontological thesauri to aid retrieval, such as the BioImage project[37]. These advances have left the Image aspect of the NBN Data Model at an undeveloped state although it is included here for completeness and does provide a simple framework which can be elaborated for application development.

The image module offers an integrated way of dealing with images of all types from oil paintings and bookplates to digital video. The main identifying information is associated with the **Image** entity which could also store images or pointers to images in other systems. The Image entity is sub-typed to provide the appropriate attributes for different kinds of image, for instance **Moving_Image** includes information on format, duration and soundtrack whilst **Artwork_Image** would include information on the materials used and so on.

In addition to subtype information, other entities are linked to the main image entity, including **Image_Dimensions** and **Image_Reference_Codes**. Images can be related to each other through an **Image_Relations** entity which would cover copies of images in different formats, prints from negatives and so on. Image reference details are not dealt with in the core version of Recorder although images can be linked to records as external references.

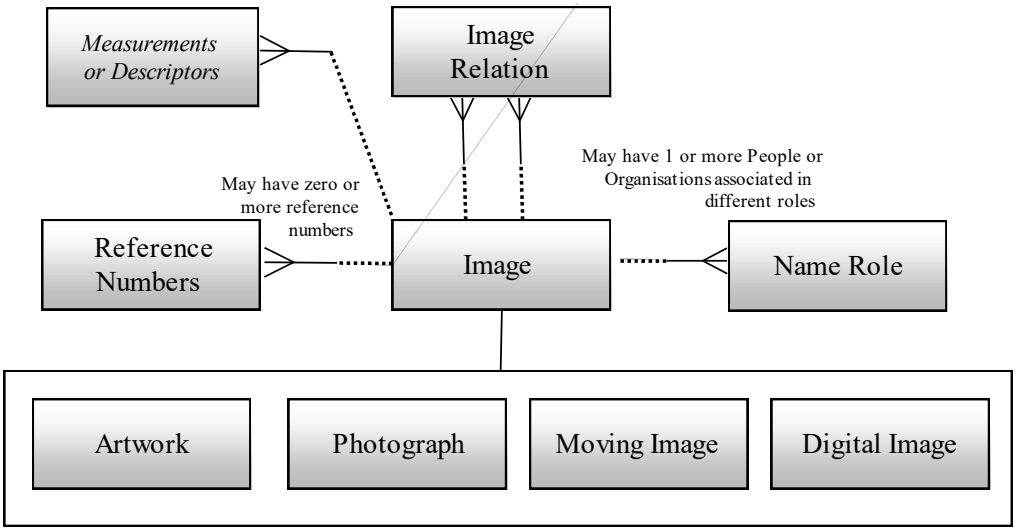


Figure 31: Simplified LDM for the Image sub-module with example sub-types (the model is essentially the same as for References although with different sub-type attributes)

An outline of the typical data attributes of images is given in *Table 14* and examples of how image data is handled in the ABCD schema is illustrated in *Figures 32* and *33*. Images can be treated as specimens in the Luxembourg Collections Add-in.

[36] see http://dublincore.org/

[37] The BioImage project, based at Oxford University is part the European funded ORIEL (On-line Research Information Environment for Life Sciences) – see http://www.bioimage.org/pub/index.jsp

Table 14: Outline of Data attributes of Images

Image Data	Definition	Entity	Validation
Image Key	Unique identifier	Image	Unique key
Image Type	What sort of image e.g. print, slide, photo, artwork etc.	Image	Controlled terminology list
Image title	Title if any	Image	
Image Description	Description of image and or its content	Image	
Name Role	People and organisations related to the image inv arious roles e.g. artist, photographer, conservator, owner	Name Role link to Contacts Module	Valid Keys
Image Date	Date image created	Image	Valid date - may be vague
Stored Image	The actual image may be stored in the database: May need to store info on image format and size. Could store a thumbnail image for reference.	Image or link to stored image	Format
Publication key	link to text/publication entity for linking images to publications. Need also to store information on link between image and publication e.g. plate in publication or this image scanned from this publication.	Image	Valid Text Key
Image Relations	Links between individual images (e.g. a slide of a painting, a series of etchings!)	Image Relations	Valid Keys
Image Reference Codes	Images may bear many reference codes e.g. accession numbers, plate numbers etc.	Image Reference Codes	
Moving Image	The moving image subtype needs to store information on format and duration	Moving Image subtype	Controlled terminology list of formats
Photographic Image	The photographic image subtype may need to store information on format, material, finish, negative number, negative type etc.	Photographic Image subtype	The MDA data standard provides a good model for attributes
Digital Image	Digital image subtype may need to store file format, resolution, colour depth, compression etc.	Digital Image subtype	
Artwork Image	Original artwork e.g. sketches and paintings may need to store information on size, format, material, technique, frame etc.	Artwork Image subtype	The MDA data standard provides a good model for attributes
Location Image	Link to a location - may need to describe content, format, film, speed, aperture etc. For monitoring purposes the camera position, height and direction need to be known. Not in Recorder or Collection add-in	Image of location	Valid Keys
Text References	Any publications etc. referring to or by this person or organisation.	Link to References	Valid reference and name keys
IPR	IPR and copyright information related to image and its use.	Source	

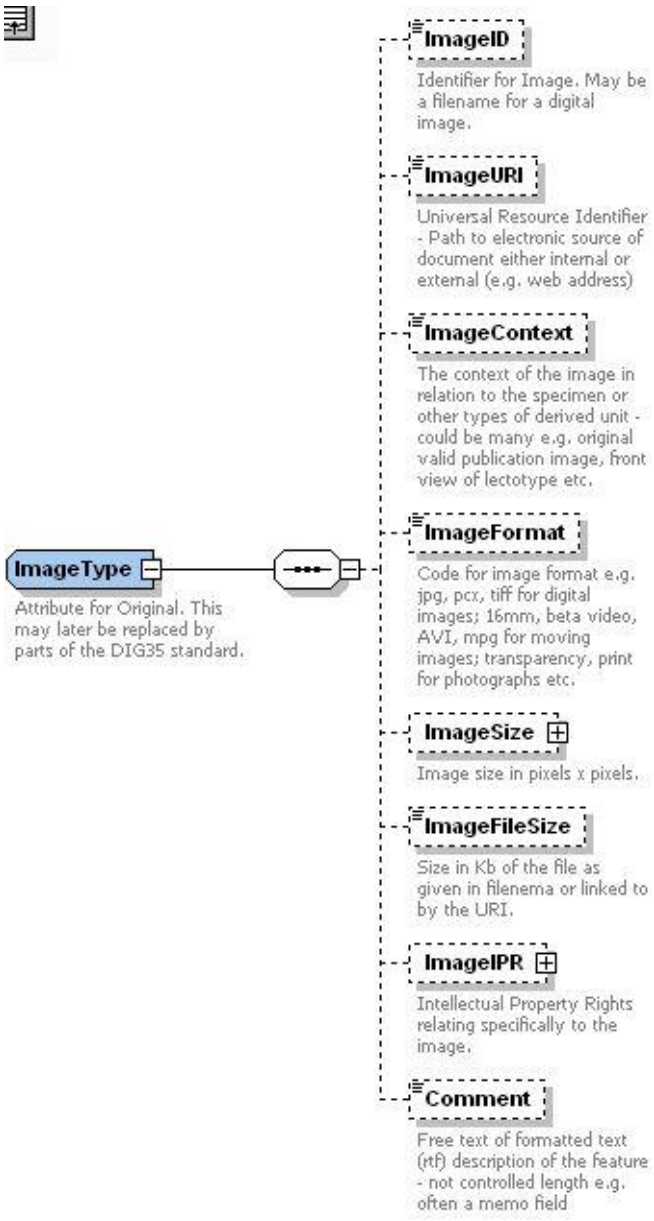
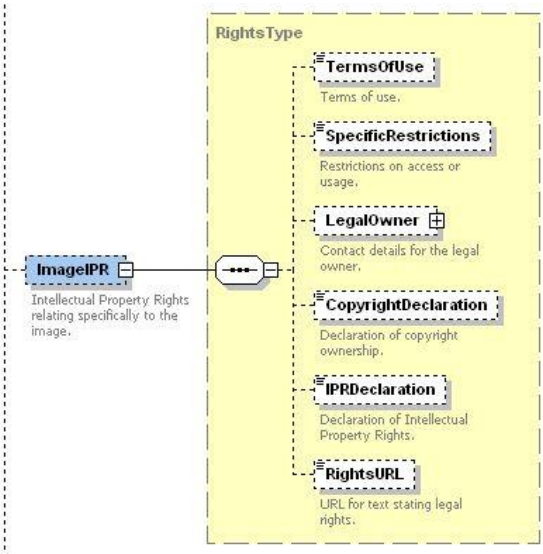


Figure 32: ImageType complex type element in the ABCD Schema

Figure 33: Expansion of ImageIPR complex type from ImageType in ABCD Schema



11. Collections Module – logical model

11.1 The relationship of the collections and survey modules

The original analysis that lead to the development of the NBN Data Model did not elaborate the information related to specimens in collections and their management but did indicate how the survey module related to them. In a project funded by the Luxembourg National Museum of Natural History a series of extensions to the NBN Data Model have been designed including, earth science and related occurrences, a new thesaurus module and an extensive collections and specimen management module. These models were used to design a number of extensions (collectively called the Collections Add-in) to the Recorder 6 Application which have been built by Dorset Software and which are currently in beta test versions (June 2004). Details of the Collections Add-In are given in Part 2 of this work, this section outlines the logical model on which the model extensions and application are based.

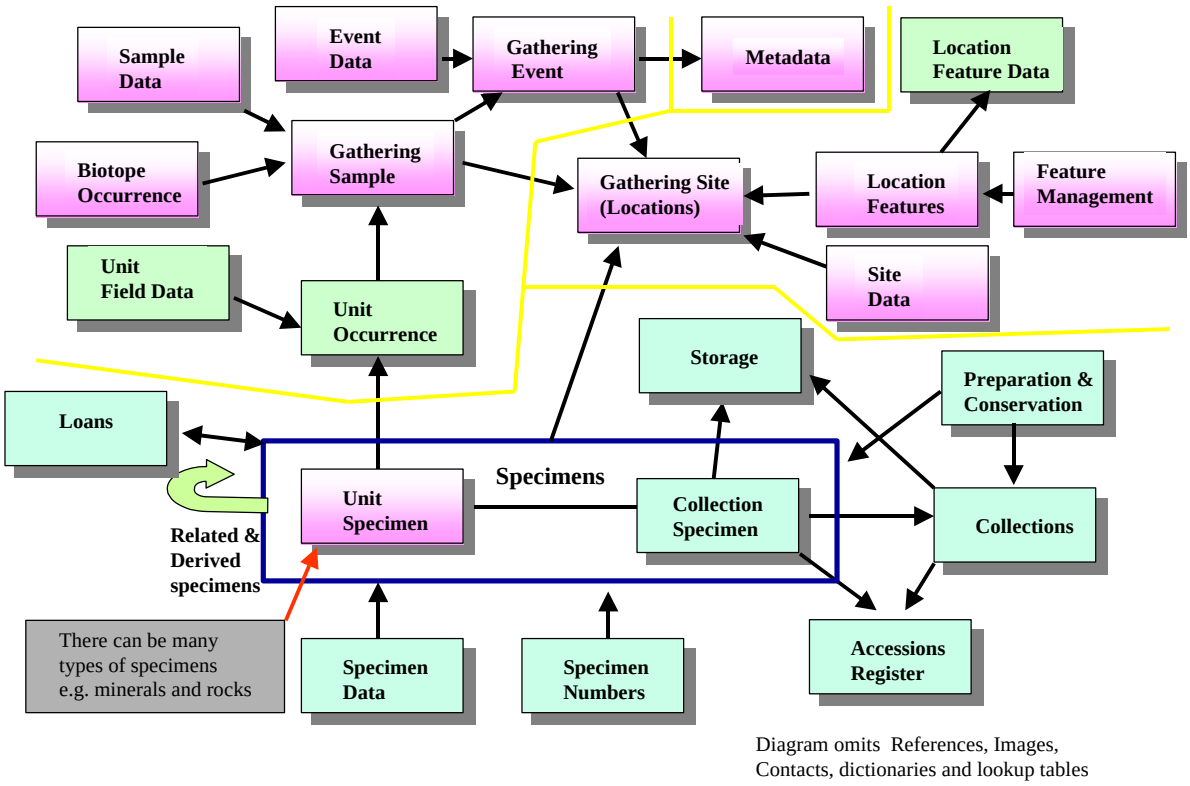


Figure 34: Conceptual diagram of the extensions to the NBN Data Model proposed in the Collections Add-in Outline specification (Copp 2002). (pink = existing entities, green new entities) nb. Gathering is a BioCISE/ABCD term for Survey in NBN model

The extensions to the existing NBN/Recorder model described here are extensive and include both additions to existing modules (e.g. the Survey module) and whole new modules (e.g. collections management). *Figure 34*) gives a map of how the main new entities link to existing Recorder modules and the scope of functional coverage. The entities labelled Unit Occurrence and Unit Field Data are placeholders for the existing Taxon and Biotope Occurrence and newly added occurrence types for soils, rocks, minerals and other earth science features.

11.2 Scope of the Collections Module

The collections module attempts to define the means for managing the following classes of information beyond the core NBN model as built into core Recorder.

1. **Acquisition and Accessions information** – acquisition of items, transfer of ownership and documentation of entry into museum/collection

2. **Collection details** – information related to groups of specimens identified by a common collector, location or sub-discipline (e.g. minerals)

3. **Specimen[38] details** – individual specimens or groups of specimens with common data. Specimens may be linked to Recorder taxon occurrences and survey/samples. New occurrence including types for minerals, fossils and rocks. Books and manuscripts may also be treated as specimens (e.g. have a store location and be subject to valuations and conservation). Specimens may have a wide range of information attached to them including descriptions, measurements, materials, determinations, taxonomic status (e.g. type), gathering localities, labels and inscriptions, associations with people or collections, preparation and mount details, loans and exhibitions

4. **Specimen/Collection History** The history of a collection unit in relation to other collections e.g. previous owners of a specimen. This can be important for historic specimens.

5. **Stores** – details of storage places and storage furniture e.g. building, room, cabinet, drawer. Also tracking of specimen movement (e.g. temporary removal from usual store place). Image maps of storage locations and images of cabinets.

6. **Loans** – tracking of temporary movement of specimens and collections into and out of the museum/collection

7. **Exchanges** – Documentation of specimen exchange with other collectors, museums and botanical gardens.

8. **Disposal** – documentation of specimens or collections that have been lost, destroyed or disposed of. Includes living collections (Exchange & death)

9. **Valuations** – documentation of monetary values placed against specimens or collections e.g. during acquisition or for loan purposes

10. **Condition Checks** – documentation of checks on storage areas, specimens and collections which might give rise to conservation actions such as fumigation or repair (Conservation Tasks)

11. **Conservation Jobs** – documentation of conservation or preparation carried out. A single job may consist of several linked tasks.

12. **Conservation Materials** – chemicals and other supplies used for conservation and preparation work - links to jobs. An extra development would be to document sources of materials, stock and cost.

[38] The term ‘Specimen’ is used here for convenience although the concept extends to groups and parts of specimens or objects. The terms ‘object’ or ‘Item’ could just as easily be used and in the BioCISE model the term ‘Unit’ is used to cover both field records and objects or specimens.

13. **Images** – Management of images and links to records. Also as information delivered to web pages. Images may also be specimens.

14. **Living collections** – The management of data for living collections has not yet been fully incorporated into the model (e.g. growth conditions etc). Also need detailed location maps. Will need functionality for documenting stored seeds and testing germination. Will also need tracking of other forms of plant propagationand animal breeding. Best developed as a separate add-in module

15. **Enquiry Log** – System for logging enquiries from the public and external users including requests for loans, data relating to distribution of species/habitats/sites or specimens and objects in collection or requests for identification service.

16. **Visitors Log** - System for logging use of collections by non-staff. Includes names, times and purpose of visit.

11.3 Collection Units and associated data

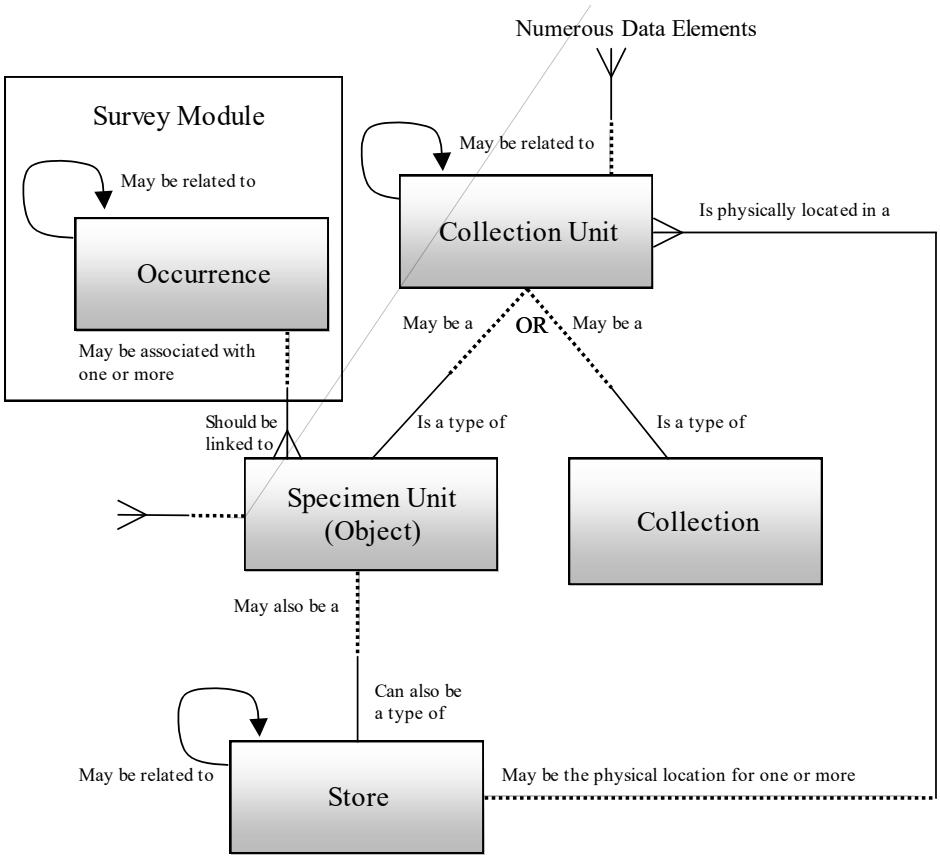


Figure 35: Specimens, collections and stores form part of the Collection Unit entity

Central to the Collection Module is the concept of a Collection Unit (see *Figure 35*). A Collection Unit represents either a specimen (object) or a collection of specimens. Collection Units are located in stores whuch may themselves be a type of specimen. For instance, a single pinned butterfly might belong to a named collection housed in an antique entomological cabinet. Each of these items (specimen, collection and store) can share many common attributes such as each may be accessioned, owned, loaned, conserved, valued, numbered, identified etc. Collection Units can have unlimited levels of hierarchy (e.g. sub-units of sub-units); collections contain specimens, specimens may be derived from other specimens, cabinets have drawers and so on.

The concept of a Unit was adopted as central to the BioCISE data model (Berendsohn et al. 1999) and is also a key feature of the ABCD schema. In the BioCISE/ABCD models field observations are also treated as units whereas in the NBN Data Model occurrences (and linked observations) are treated as part of the separate but linked Survey Module. The reason why the NBN model separates these concepts is that it makes it simpler to use the model for defining practical applications that focus either on field data or the management of physical objects. Modelling is always a balance between abstraction and application, the separation of these concepts helps modularise the data.

The Collections Module is complex both in its logical and physical build forms. For this reason its is described in parts, in the sections below. Full details of the physical model, tables and attributes are given in Part 2 of this work.

11.3.1 Main descriptive data elements associated with specimens

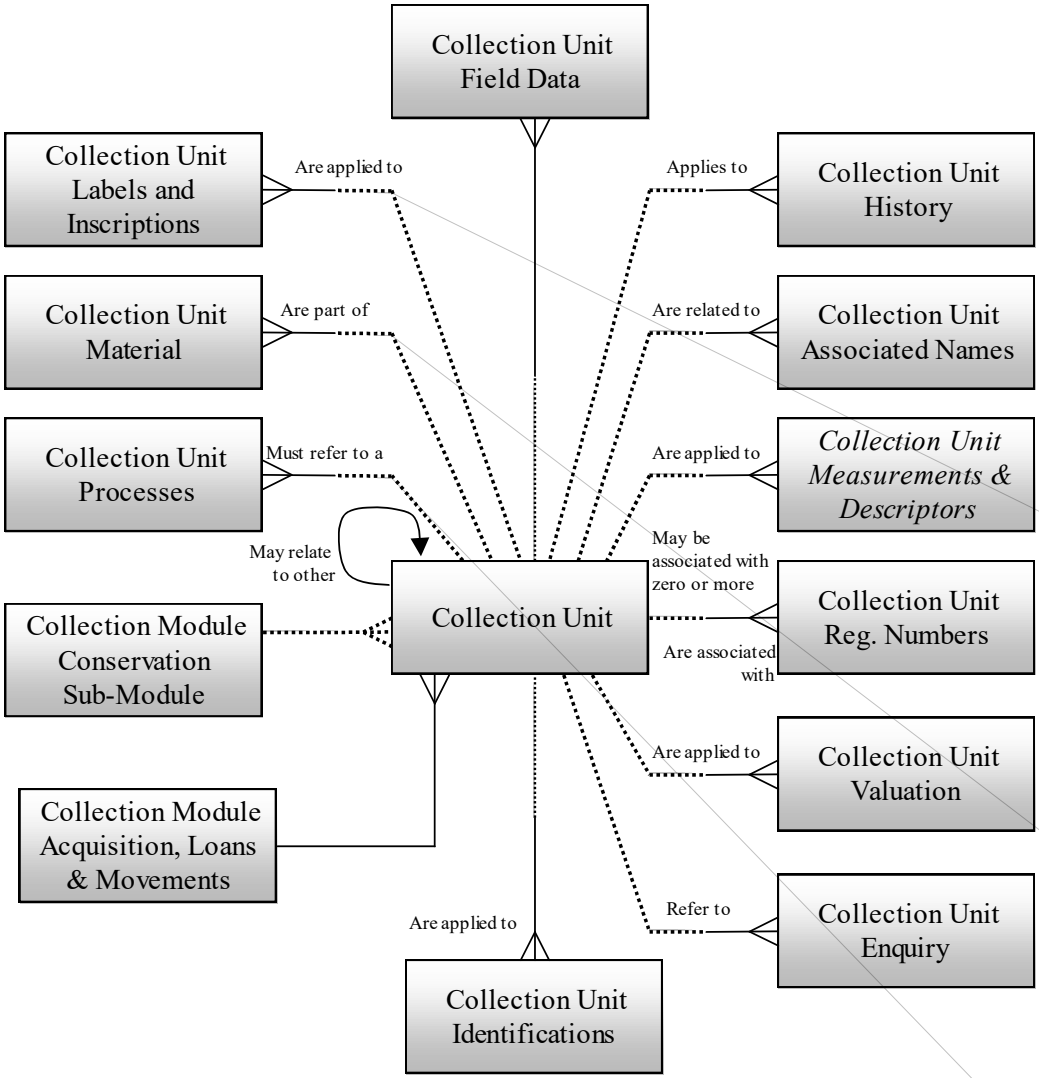


Figure 36: Descriptive data elements associated with collection units and links to conservation and management (movements) modules

The Collections extension to the NBN Model covers three main areas of information; descriptive data elements:

- *Descriptive data elements associated with specimens and collections*
- *Management information related to acquisition, accession, storage and movements*
- *Conservation and preparation activities related to specimens and collections*

Figure 36 outlines the chief items of descriptive data linked to Collection Units. Many of the individual characteristics of specimens (e.g. length, colour, texture etc.) will be covered by Measurements and Descriptors common entity linked to term-lists filtered by subject and domain. Generalisation of measurements and descriptors in this way makes it easy to extend the model for use in different domains, for instance palaeontologists may define numerous specialist measurements or descriptors relevant specific fossil taxa. *Figure 37* shows the General Information tab associated with a specimen in the Recorder Collections Add-In.

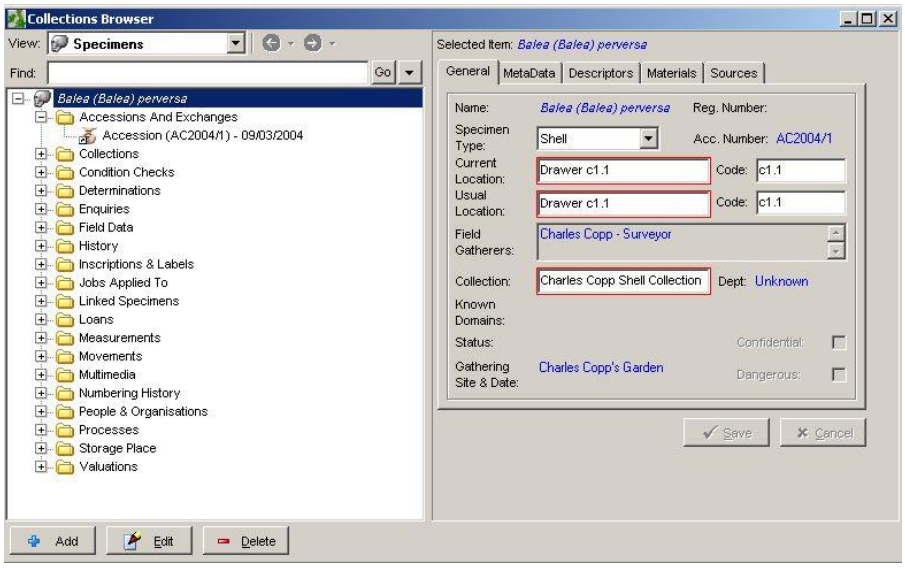


Figure 37: General data tab for specimens in the Recorder 6 Collections Add-in

11.3.2 Collection Management: Acquisition, Accession and Movements

Acquisition is regarded as the act of physically acquiring a collection unit (specimen or collection) and is associated with movement, usually movement into the museum or acquiring organisation. Accession is associated with the actual transfer of ownership, which may not happen at the same time as acquisition. Accession should be associated with accession documentation including the allocation of an accession number although individual registration of specimens in an accessioned 'lot' may take place piecemeal over a longer time span.

Accession and Acquisition can be regarded as types of movement, the first is physical movement and the second movement (or transfer) of ownership (or custodianship). Other activities involving collections Units may also be regarded as movements, such as loans into and out of a collection or placing objects on display. *Figure 38* shows a tradition way of relating these concepts to specimens. Study of the information content of the records associated with these kinds of movement shows that they can be generalised which allows for new ways of looking at the history of specimens and collections. In the work on the Collections Add-in this lead to the development of a Movements Module (see Figure 39) which tracks these items of information. Details of how the Movements Module works are given in part 2 of this work.

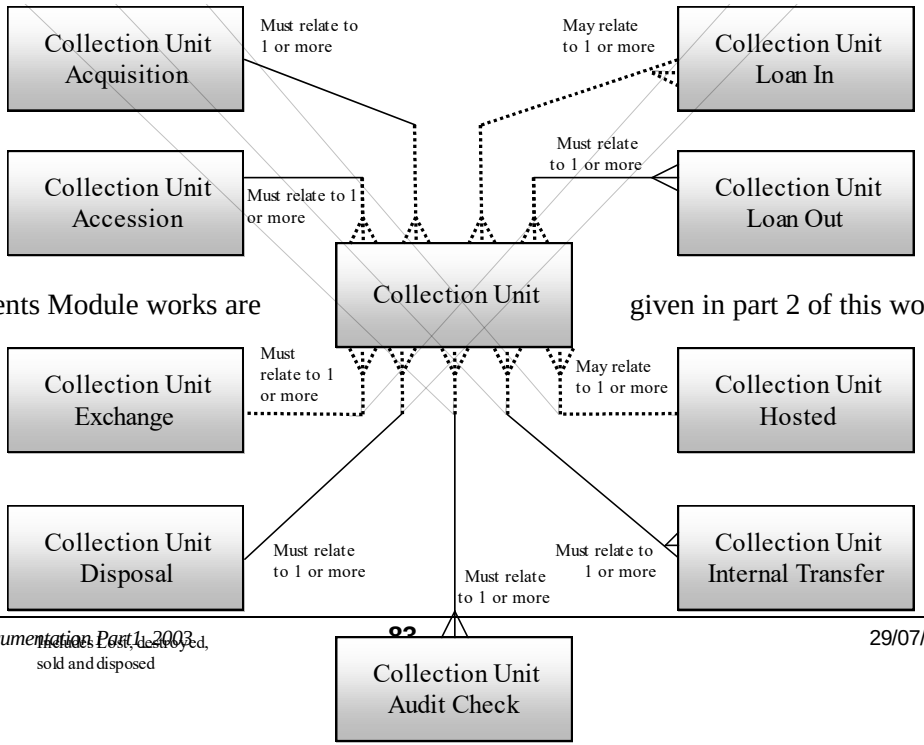


Figure 38: Data elements associated with the acquisition, accession and movement of specimens and collections

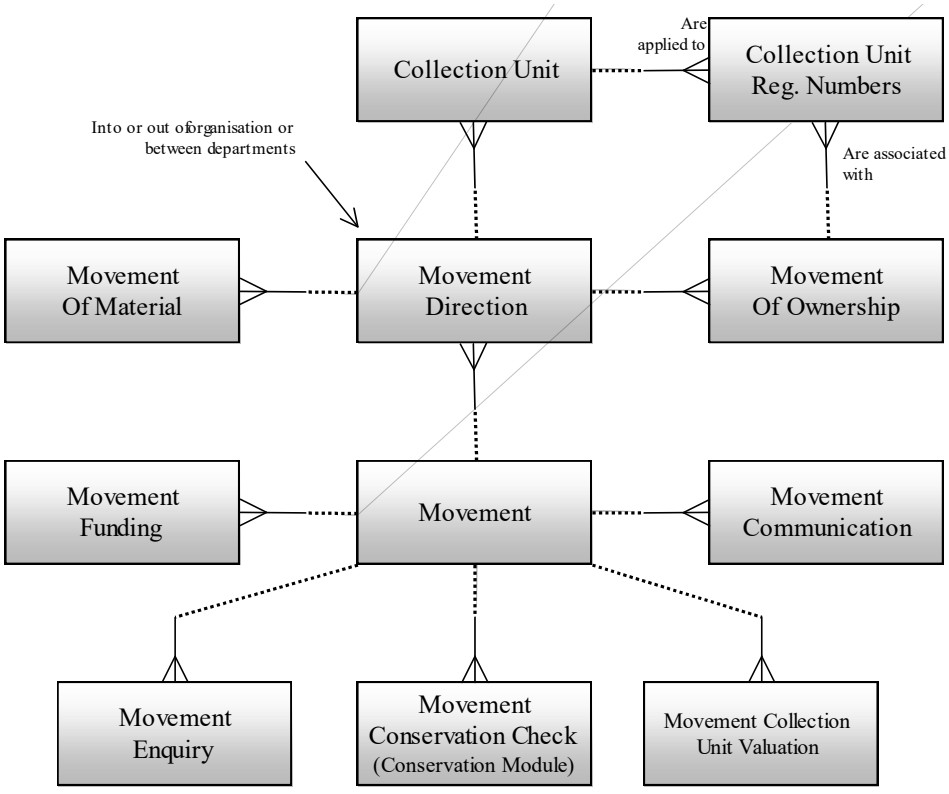


Figure 39: Logical Data Model for the Movements Module, which encapsulates all of the different elements associated with movement of objects and transfers of ownership.

11.3.3 Collections Module – Conservation and Preparation

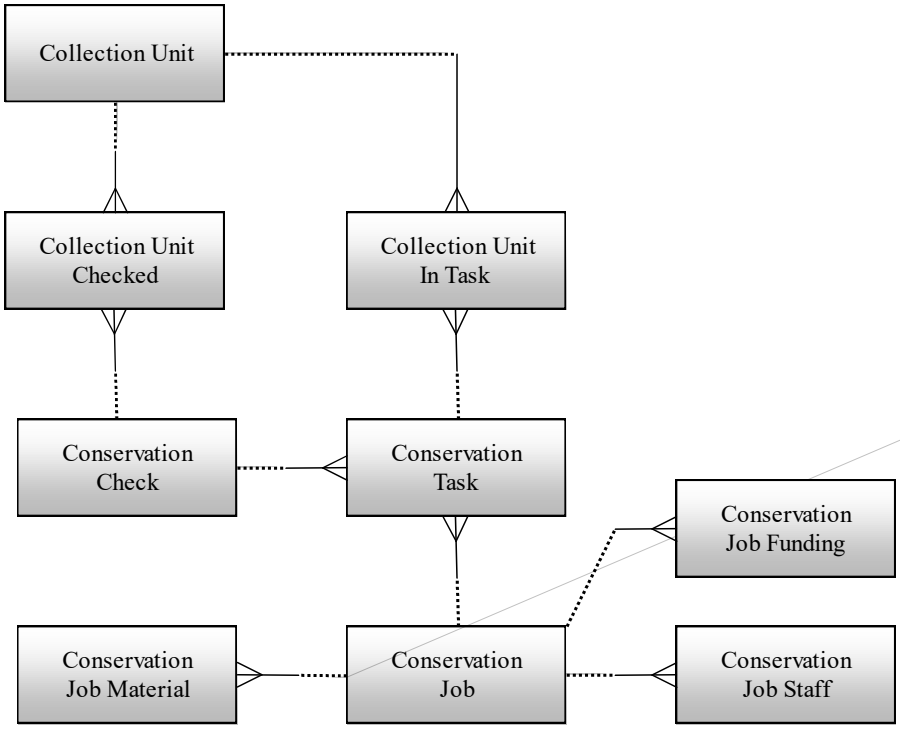


Figure 40: The conservation and preparation sub-module for the collections add-in

Figure 40 shows the logical structure of the Conservation Module. Activities and information are divided into three areas. Conservation checks may be carried out in relation to any combination of specimens, collections and stores. Checks may be part of a regular monitoring programme, associated with acquisitions, loans or exhibitions or random. Each conservation check may result in the identification of a number of tasks where each task may be related to one or more items. For instance, a check of the mounted bird collection might reveal that all the specimens need fumigation against mites whilst one might need mending and several need cleaning. These tasks can accumulate until a specific job is identified to be carried out, for instance the fumigation of a store which would coincidentally include all the specimens in the task list. A later job of cleaning birds could be carried out by a different conservator and include specimens from several conservation checks. Coonservation jobs may be funded from various sources and a log of materials used can be made.

12. Other Implementations of the NBN Data Model

Since the launch of Recorder 2000 a number of add-ins and variations have been developed for it for specialist purposes. In most cases the original data model has met the needs of the developers but there are examples where simple extensions have been made.

Of more note is the development of the database for the NBN Gateway which represents a physical simplification of the database model and some may argue does not use the model at all. This is not essentially true because the Gateway does utilise data from the taxon dictionary and its table structure can be readily mapped back to the NBN structure. It is important to remember that the NBN model is a tool for understanding the relationships of items of information relating to biodiversity, earth sciences and collections. It is not a required physical database model, some applications, like Recorder, are very close in their structure to the model but others, like the BioCASE Thesaurus have physical builds that look very different from the logical model. The NBN database is a derived collation of selected data from other databases and its structure is optimised for its purpose, which is essentially reporting. The more that a database needs to deal with the management of complex biodiversity data relating to different collection methods or integrating specimen and field data or work in more than one subject domain, the more likely that its structure will need to match the NBN model.

12.1 Marine Recorder

The following extract on changes to the NBN Model for the implementation of Marine Recorder are taken from the Technical Documentation for Marine Recorded by Dr. James Perrins. (ExeGesIs 2002)[39].

12.1.1 History

MarineRecorder was originally produced to be a marine equivalent of Recorder 2000. At the time it was felt important that the data structure of Recorder 2000 be maintained and so MarineRecorder was constrained by that. Also in addition (and separately from the development of MarineRecorder) all the data in the MNCR AREV database was extracted and re-structured so that it would fit into the NBN data model implemented by Recorder 2000.

MarineRecorder was intended to be a collect and collate piece of software, aimed primarily at allowing end users (especially contractors who it was anticipated would be entering more of the data) to get data into a system and do some preliminary validation on it.

Version 2.0 of Marine Recorder was produced under contract to JNCC in August 2002. This incorporated a large number of enhancements to the original version, but the original data structures remained more or less unchanged. Data entered in the previous version would automatically upgrade.

12.1.2 Changes to the NBN Data Model

A lot of effort has been spent in making the marine data fit the existing NBN data model as implemented by Recorder 2000 in the NBNDATA.MDB. In various places it would have been considerably easier to alter the structure than fit with the existing one, but this was avoided for the sake of standardisation. This was basically made possible by extensive use of tables such as the SAMPLE_DATA table to hold additional related sample information

[39] See http://www.esdm.co.uk/MarineRecorder/New/

There were however a couple of changes that had to be made to the data structure in order to accommodate the marine data. In all cases these were additions to the data model, so there should be no problem incorporating pre-existing data.

Two new tables were created (SURVEY_DATA, SURVEY_EVENT_DATA). These have similar structure and function to other DATA tables (e.g. LOCATION_DATA).

The DATA field in all DATA tables (except SAMPLE_DATA) was increased to 20 characters in length. This was necessary in order to store user entered Lat/Long positions in the field so that bounding boxes could be used for LOCATION and SURVEY_EVENT records.

SORT_ORDER was added to the new MEASUREMENT_QUALIFIER table, as there is a requirement for qualifier terms to appear in a specified order (e.g. depth ranges).

SAMPLE_REF field in the SAMPLES table was indexed to improve performance as this field is integral to the way MarineRecorder groups samples.

Referential Integrity within the database was altered to allow for cascading updates and deletes. This made certain deletions easier to undertake, and made the potential merging of data sets and re-keying of information easier at some point in the future.

A very simple front end was placed in the NBNDATA.MDB which denies the casual user access to the database (so reduces the risk of someone doing something silly with cascading deletes in place).

12.2 Lowland Parks and Woodlands Information System (LPWIS)

The Lowland Parks and Woodlands Information System[40] is primarily a metadata gathering project but development work included the development of an add-in to Recorder which allowed for the recording of Site Feature Data. The addition of site feature data and the relation of features to survey occurrences and to specimens is now covered in the extensions to the model associated with the Collections add-in.

[40] See http://www.ukwildlife.com/metadata/parks/index.htm

13. Glossary

Attribute	An attribute is the term used for what is commonly called a 'field' in a database record.
BioCASE	The BioCASE[41] project is due to run until January 2005 and is charged with delivering a working network linking national nodes for as many as 31 countries and providing access to hundreds of biodiversity databases in those countries. The BioCASE project has developed software for collecting and storing collections metadata, backed by an XML metadata schema, portal and wrapper software, an XML data transfer standard (ABCD)[42] , user interface software and a thesaurus to provide controlled terminology (See Figure X). All of this work is in the public domain and can be used to build local networks as well as national and international ones.
CASE	Computer Aided Systems Engineering – generally applied to software that aids information system development through the use of modelling tools and functions for generating databases.
Conceptual Model	A conceptual model is a representation of how the designer views the system that is being analysed and from which the detailed system models are developed.
Data Flow Diagram (DFD)	A modelling technique that depicts how data moves within a system. DFDs are drawn at various levels from a high level over-view (conceptual view) to low-level details of specific tasks.
Data model	A means of depicting the data content within an information system. Data models can be created at various levels from high level overviews to details of actual data stored. Data models include various types of model including Logical Models and Relational Data Models
DEFRA	The UK Department for Environment, Food and Rural Affairs
Element	An element is an item of data. In XML (Extensible Markup Language) an element is a data structure which consists of surrounding tags, a name (which may relate to a specific naming convention called a namespace), attributes, data and closing tag (e.g. (2001-01-30)). Complex elements are constructed by the inclusion of other elements.
Entity	An entity is an item of interest i.e. something about which the system stores data.
Entity Relationship Diagram (ERD)	An Entity Relationship Diagram is an illustration of the relationships between items of interest. A number of diagramming techniques exist which formalise the symbols used for different types of entities and relationships.
Function Model	A model that depicts the actual functionality of an application or information system
GBIF	The purpose of the Global Biodiversity Information Facility (GBIF)[43] is to make the world's biodiversity data freely and universally available via the Internet. GBIF works cooperatively with and in support of several other international organisations concerned with biodiversity. These include (but are not limited to) the Clearing

[41] BioCASE http://www.biocase.org
[42] ABCD http://www.bgbm.org/TDWG/CODATA/Schema/default.htm
[43] GBIF http://www.gbif.org/GBIF_org/what_is_gbif

	House Mechanism and the Global Taxonomic Initiative of the Convention on Biological Diversity and regional biodiversity information networks. Participants in GBIF have signed the Memorandum of Understanding and support network nodes through which they provide data. Functionally, GBIF encourages, coordinates and supports the development of worldwide capacity to access the vast amount of biodiversity data held in natural history museum collections, libraries and databanks. Near term GBIF developments will focus on species and specimen-level data.
JNCC	Joint Nature Conservation Committee
Logical Model	A model, usually diagrammatic, that shows the relationships between entities (things of interest to the system being modelled).
LRC	Local Record Centre also Biological or Environmental Records Centre
NBN	National Biodiversity Network
Object Oriented Model	An object is a data structure (incorporating data and methods) whose instance is unique and separate from other objects, although it can "communicate" with other objects. Objects are most frequently used in application programming. An object model can convey information about entities (items of interest) that includes not only their structure but any properties or processing associated with them.
Object Relational Model (ORM)	The Object Relation Model extends traditional relational database structures of tables with rows of data to include complex 'objects' such as polygon boundaries from geographical information systems or images or sounds
Ontological Model	An ontology is a representation of the nature and relationships of a chosen subject for instance an ontology of animal behaviour seeks to represent the different types of behaviour, their characteristics and relationships. This type of modelling is becoming important for knowledge capture or representation in relation to data retrieval from large databases and the web.
Physical Model	The actual way that the logical model will be represented in whatever data management system is chosen for its implementation. This may look very different from the logical model because of constraints in the data management system and to meet performance requirements
Process Model	The graphical representation of processes, including a hierarchy of relationships between them. Process models capture the essence of the system being configured and developed. A given process may be used in the delivery of more that one function - what the system does with the data
Recorder (2000/2002/6)	Biological Records collection and collation software see: http://www.nbn.org.uk/information/info.asp?Level1ID=1&Level2ID=1
Relational Data Model	A model that uses 'normalisation techniques' to render the data required by an information system into a number of related tables with the minimum degree of data redundancy (repetition) compatible with the performance requirements of proposed application(s) that will access the data
Semantic Model	A model that records the meanings of links between entities (items of interest)
SSADM	Structured Systems Analysis and Design Method, a set of standards developed in the early 1980s for systems analysis and application design widely used for government computing projects in the United

	Kingdom. SSADM uses a combination of text and diagrams throughout the whole life cycle of a system design, from the initial design idea to the actual physical design of the application.
Systems Analysis	Work that involves applying analytical processes to the planning, design and implementation of new and improved information systems to meet the business requirements of customer organizations.
Table	A database structure that stores data as a number of rows (originally called tuples) in which each row may be divided into a number of attributes or fields.
UML	Unified Modeling Language, an object-oriented design language. UML is a standard notation and modeling technique for analyzing real-world objects, developing systems, designing software modules in object-oriented approach. UML has been fostered and now is accepted as a standard by the group for creating standard architecture for object technology, OMG

14. References

ABCD	2003	*Content definition: the ABCD Schema*. CODATA/TDWG Task Group on Access to Biological Collection Data. http://www.bgbm.org/TDWG/CODATA/Schema/default.htm
Berendsohn, W.G.	1997	*A taxonomic information model for botanical databases: the IOPI model.* Taxon 46 pp. 283-309
Berendsohn, W.G. et alia.	1999	*A comprehensive reference model for biological collections and surveys.* Taxon 48 August 1999 pp511-562 http://www.bgbm.org/biodivinf/docs/CollectionModel/
CIDOC	1995	*International Guidelines for Museum Object Information: The CIDOC Information Categories* , International Committee for Documentation (CIDOC). Printed version, ISBN 92-9012-124-6 *http://www.willpowerinfo.myby.co.uk/cidoc/guide/guide.htm*
Copp, C.J.T.	1998	*The Recorder Project: Systems Analysis.* Unpublished Report. JNCC Peterborough. Jan. 1998. http://www.bgbm.org/biodivinf/docs/archive/Copp_C_1998_-_The_Recorder_Project.pdf
Copp, C.J.T.	1998b	*Description of the NBN Data Model: Logical and Physical Models* Unpublished Report. Environmental Information Management
Copp, C.J.T.	2000	*The NBN data model and its implementation in Recorder 2000.* Environmental Information Management, October 2000 http://www.bgbm.org/biodivinf/docs/archive/Copp_C_2000_-_NBN_Data_Model.pdf
Copp C.J.T.	2002	*The BioCASE Thesaurus: Logical and Physical Models* Deliverable 9. Workpackage 4. The BioCASE Project. http://www.biocase.org/Doc/Results/results.shtml
Copp C.J.T.	2003	*The BioCASE Thesaurus: Proposed Final Logical and Physical Models* Deliverable 25. Workpackage 4 The BioCASE Project http://www.biocase.org/Doc/Results/results.shtml
Copp, C.J.T.	2003	*ENHSIN*
Dorset Software	2003	*Technical System Design: Recorder 6.* Unpublished technical specification, Dorset Software, Poole, Dorset
Dorset Software	2003	Technical System Design: Collections Module for Luxembourg musée national d'histoire naturelle. Unpublished technical specification. Dorset Software, Poole, Dorset, UK.
MDA	1994	*SPECTRUM: the UK Museum Documentation Standard* (1994). Edited by A. Grant. Cambridge: Museum Documentation Association. Available in print and electronic editions from the Museum Documentation Association.