National Biodiversity Network

Sharing Information about Wildlife

Improving wildlife data quality: guidance on data verification, validation and their application in biological recording

Compiled by:

Trevor James, NBN Development Officer for Societies & Schemes

These Guidance Notes are designed to help people involved in biological recording or the use of wildlife data to improve the quality of the data they collect or compile.

The Guidance consists of:

- Introduction: definitions of what "verification" and "validation" consist of.
- What wildlife records are, who makes them, and why:
  - What records consist of.
  - Processes carried out in making and compiling them.
  - Responsibilities for different parts of the process.
- What makes a good wildlife record, and factors underpinning quality.
- Responsibility for data quality, and some basic principles.
- How understanding data flow is important for improving data quality, and recommended approaches to promoting data quality at different points.
- Roles of people involved in recording:
  - Data collection.
  - Identifying and verifying records.
  - Quality control during data management, including validation.
  - Data quality and the data custodian, including data dissemination.
- Who could be doing what to support data quality:
  - National societies and recording schemes.
  - Local records centres and related bodies.
  - Non-governmental biodiversity organisations.
  - Statutory and other official biodiversity organisations.
  - Commercial and professional biodiversity organisations.
- Case studies.

1. Introduction

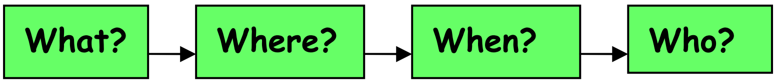
These guidance notes focus on wildlife **data verification** and **validation**, in the context of the overall collection, management and dissemination of information. They are intended for use by anyone involved in collecting or using wildlife data. They are not intended to be the last word. Different participants in biological recording will have more or less of a need to adopt particular methods. One solution will not suit everyone.

Definitions:
Data verification: ensuring the accuracy of the identification of the things being recorded.
Data validation: carrying out standardised, often automated checks on the "completeness", accuracy of transmission and validity of the content of a record.

Because the business of collecting, managing and disseminating wildlife data is a web of processes, supported by a complex network of organisations and individuals, guidance on quality control mechanisms must also be based on a good understanding of the way the business works. We therefore hope these notes will highlight key issues, and the recommendations be taken as a potential guide for particular organisations and individuals working as parts of this network.

2. What are wildlife records and who makes them?

A basic wildlife record is a documented occurrence of an organism at a location, at a point in time by a named person. It is an attempt to document an ephemeral event linking a representative example of a species with a place and possibly with other individuals and other species. This is often summed up as:

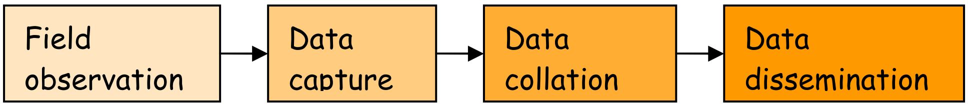

What? → Where? → When? → Who?

Underlying this, and often of over-riding importance, is the other question: This can be both "why are we making this record?", as well as "why is this organism here?".

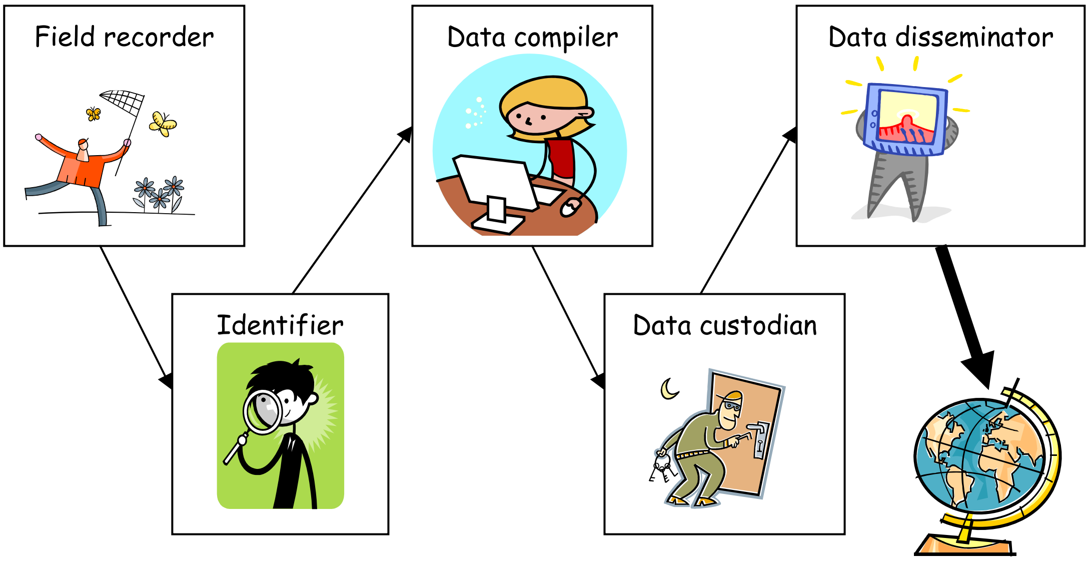
Why?

In order to understand data quality it is essential to appreciate the factors that can affect the accuracy and precision of information relating to each of these parts of a record. It is even more important to understand how the question "why?" can be of fundamental importance in both making an accurate and useful record, and in using these records effectively afterwards.

We therefore also need to understand the processes undergone in producing records, and in making use of them. A way of understanding this is by using a method of analysis called "data flow". This is dealt with in more detail below, but, in essence, it can be summed up, in relation to a wildlife record, as:



Finally, we need to recognise and understand the functions of the different roles of those involved in making and disseminating records:



These guidance notes are therefore intended to address not only the general questions of data quality, but also who should be doing what in the process.

3. What makes a good wildlife record?

If we are making a wildlife record, there is not much point in doing so unless it is as correct and complete as possible. It becomes increasingly important for wildlife records to be "correct" the more these are used by others in understanding or making crucial decisions about biodiversity. The creation of a wildlife record is therefore a means of creating a "true" statement about the occurrence of a species at a particular locality at a particular time. However, the number of variables involved is often considerable.

What is recorded will depend on the objective of the observer and of the organisation carrying out the survey. There will be questions about the likelihood of a particular species actually being found, either at all, or in a particular place. There will be issues of defining the locality and the "habitat", both in relation to the way a survey is designed, and physically on the ground. There are often questions about which species is being recorded (or whether the individual specimen observed actually represents a "species" at all!). Above all, the way single observations fit into surveys is important; and in addition, the way observations are put together for analysis impinges on the reality of what has been recorded and the way the data are subsequently used.

Wildlife data include not only “traditional” species records, in whatever way they are made, but also increasingly include structured observations on habitats or other physical features of the environment, either as the objects of recording themselves, or in relation to the presence of species. Standardised approaches to the way these are described also require accurate “identification” of what they represent.

Key features underpinning the quality of biodiversity data are:

- **Accurate identification of the thing being recorded (species, habitat etc.)**
- **Precise recording of the geographical locality, depending on survey objectives.**
- **Careful documentation of other aspects of the record, such as time or date; the individuals that made the record; and the individuals that substantiated details of the record subsequently, where relevant.**
- **Transparency, robustness and appropriateness of the methods by which collected data are subsequently managed and made available to others.**

The way we verify the main elements of a record in the first place, and secondly the way we validate associated factors or the processes through which details of the record have been managed, are therefore two of a range of issues which directly influence the way data users can judge the quality of the final data.

This guidance paper focuses especially on these two functions: data verification and data validation. But they cannot be separated from the other elements of data quality, which are equally important:

- **Survey objectives and design.**
- **Organisational capacity to carry out the survey.**
- **Methods of data management and presentation.**

Therefore, these notes draw attention to the need for organisations and individuals involved in recording to be aware of and understand how all these factors come together to create reliable records.

4. Who should be responsible for data quality and how?

The simple answer is: everyone involved in the recording, data processing or data provision processes.

However, there are some **basic principles**:

- **Good quality data depend on collection of all relevant information as close to the point of observation as possible.**
- **Clear survey design and a statement of survey objectives are usually important, although casual recording may be useful, as long as the gathered data are structured in a useful way. In either case, having a clear policy from the outset on the level of accuracy required for a particular purpose, how this is to be achieved, and making this plain to participants is vital.**
- **Clarity from the outset over the role of individuals involved in the recording and data management processes is essential.**
- **Well thought-through processes of data management subsequent to field collection are vital.**

- **Clear documentation is needed of the way data are collected and processed so that others can judge what has been done.**

It is useful to recognise the different **potential sources of error** and unreliability of data. These can come from **people, processes and systems**:

- **Lack of relevant skills in field observers/collectors.**
- **Lack of appropriate reference to specialists or experts where these are needed.**
- **Lack of responsibility for or unmethodical processes of data collation, checking and presentation.**
- **Lack of technical skills in data management or lack of access to appropriate techniques or facilities.**
- **Mismatch between survey objectives and the application of recording methods, resulting in unevenness or inadequacy of survey coverage.**

The way data quality is assured therefore depends to a great extent on the role of individuals and organisations in the process. There could be very formal ways to achieve data quality through officially recognised training, qualifications, and accreditation, alongside technical solutions to the management of data, which would require imposed levels of acceptability of records. However, one of the outcomes of recent debates that have been undertaken through the NBN is the recognition that imposition of a one-size-fits-all solution would not only be impracticable, but also would be damaging to biological recording.

Most important has been the conclusion that a slightly more co-ordinated approach to the existing "peer review" process would be the most appropriate way forward, because it can be flexible, taking into account such things as an individual's altered capability in a subject over time.

Also, the capabilities and needs of different subject areas will be different. Therefore it is considered better that responsibility for identifying recorder capabilities, or assessing the way that a particular recording activity is carried out should be left to those involved in a particular organisation or activity. However, general recommendations can be given which, if adopted, would allow them to demonstrate that they have addressed the need to ensure the quality of the resulting data.

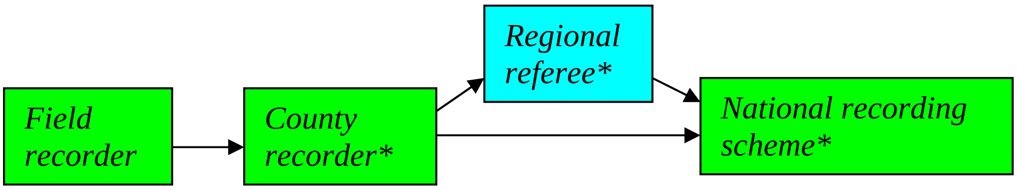
5. Data flow and data quality

Understanding the flow of data through the recording process is an essential first step in improving data quality. Advice on the application of data flows to the identification of levels of responsibility for data management in the context of the NBN has been developed by the NBN Trust and is available from its website.

However, understanding the flow of data is also especially important for improving the quality of data overall by:

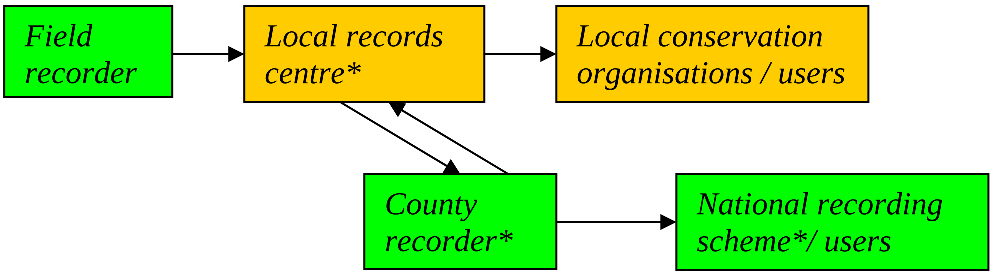
- **Reducing the amount of processing that records undergo, therefore reducing the likelihood of error.**
- **Defining responsibilities for and points where records should be checked at specific stages during the data management process.**
- **Establishing and promoting the most effective pathways for communicating data from and to other people or organisations.**

A traditional example of a data flow model for voluntary sector wildlife records might be:

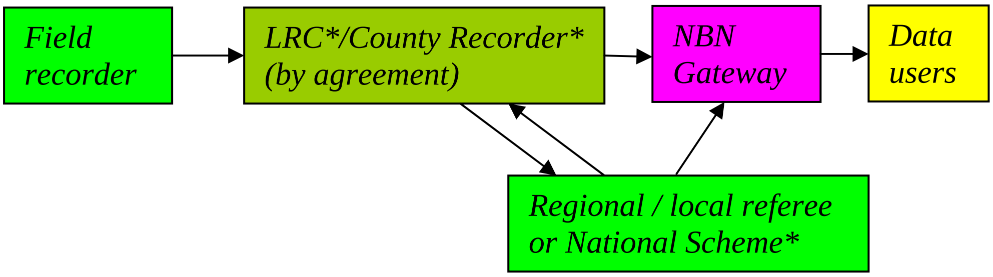


Points at which data validation might be carried out independent of the original supplier of records are marked: *

An alternative, if a local records centre is included in the process, might be the following, although this can lead to duplicated or different versions of data being made available to users from different sources:


Field recorder
Local records centre*
Local conservation organisations / users
County recorder*
National recording scheme*/ users

However, if the NBN Gateway, through its Validation level of access to a dataset, is used as a validation tool, and subsequently as a means of providing the data to users, an 'ideal' data flow might be:


Field recorder
LRC*/County Recorder* (by agreement)
NBN Gateway
Data users
Regional / local referee or National Scheme*

The last of these examples has the advantage that it integrates the process of verification and validation with the process of making data from whatever source available to users. In this model, agreement is needed as to who should be responsible for handling the data validation, and for making a dataset available to others. Such agreements can also be used as the basis for mutual data use and exchange arrangements.

The NBN Trust recognises that there will never be a single, agreed system for communicating all wildlife data and that data collected for specific uses may or may not need to be supplied to others. The NBN Gateway was established in order to provide a simplified mechanism for any participating body to communicate data to users, but it does not attempt to impose formal data flow mechanisms between field recorders, specific data collation bodies and data custodians.

As a basic set of principles in the context of data flows, verification and validation, the NBN Trust recommends that:

- **Organisations involved in operating field recording programmes should promote standard methods of capturing and processes for submitting records wherever possible.**
- **Field records collected by individuals should be collated, preferably using standard formats, by either a recognised local or national species recording scheme, or by a formally established local records centre, if there is one in a particular area.**
- **Records collated by local voluntary organisations should undergo validation and verification where necessary either by their own recognised experts, or through submission to external experts or a national recording scheme, according to published protocols.**
- **Records collated by a local records centre should be subjected to verification and validation by recognised local or national experts where relevant, according to agreed, published protocols.**
- **Records collected by professional or other official organisations should be subject to as rigorous quality checks as those recommended for voluntary sector or local recording organisations, and that they should consider making their data available for "peer review" by relevant experts where necessary before they are made available to others.**

6. Roles and responsibilities for data quality

Actions can be broadly split into different areas, relating to stages in the process of collecting, collating and disseminating data. However, each area is dependent on another, so it is not possible in practice, for example, to entirely separate identification/verification from either survey operation or from the process of managing data.

6.1 Data collection

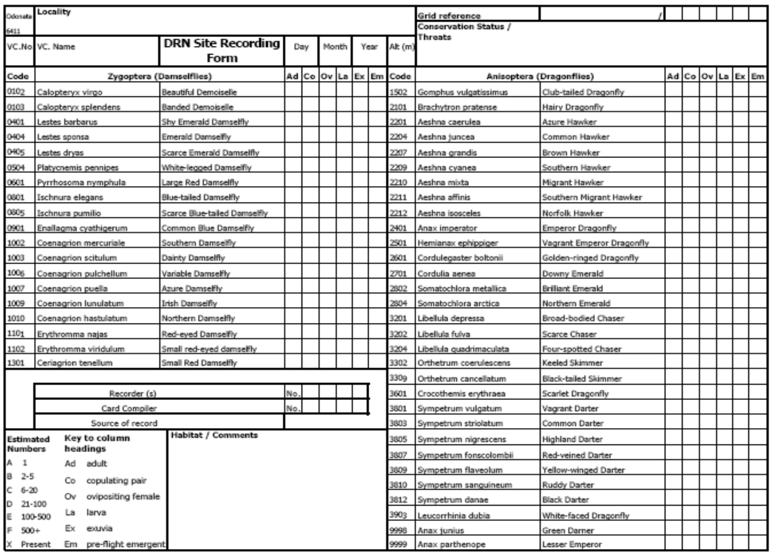
Bearing in mind that ensuring quality of data is best done as near to the point at which records are made as possible, the process of data collection becomes particularly important.

This might involve attention to:

- **Survey design and method (appropriateness to the subject being studied, appropriate or realistic timescale, capacity to deliver required information).**
- **Availability of appropriate skills in those carrying out recording, or capacity to train if necessary.**
- **Availability of necessary data capture equipment or materials and the knowledge of how to use them.**
- **The potential for collected data to be used flexibly (e.g.: ensuring data are in formats accessible by others).**
- **Ensuring that all the necessary facts are recorded appropriately at the point of observation.**

Making field records - ways to enhance data quality:

Standard record cards (or data logger entry screens), with clearly thought-out data entry boxes, relevant to survey objectives, appropriate for data handling processes; and incorporating accurate species checklists (e.g.: Dragonfly recording card):



Odonata 6411	Locality					Grid reference	/	
VC.No	VC. Name	**DRN Site Recording Form**	Day	Month	Year	Alt (m)	Conservation Status / Threats	

Code	Zygoptera (Damselflies)		Ad	Co	Ov	La	Ex	Em
0102	Calopteryx virgo	Beautiful Demoiselle						
0103	Calopteryx splendens	Banded Demoiselle						
0401	Lestes barbarus	Shy Emerald Damselfly						
0404	Lestes sponsa	Emerald Damselfly						
0405	Lestes dryas	Scarce Emerald Damselfly						
0504	Platycnemis pennipes	White-legged Damselfly						
0601	Pyrrhosoma nymphula	Large Red Damselfly						
0801	Ischnura elegans	Blue-tailed Damselfly						
0805	Ischnura pumilio	Scarce Blue-tailed Damselfly						
0901	Enallagma cyathigerum	Common Blue Damselfly						
1002	Coenagrion mercuriale	Southern Damselfly						
1003	Coenagrion scitulum	Dainty Damselfly						
1006	Coenagrion pulchellum	Variable Damselfly						
1007	Coenagrion puella	Azure Damselfly						
1009	Coenagrion lunulatum	Irish Damselfly						
1010	Coenagrion hastulatum	Northern Damselfly						
1101	Erythromma najas	Red-eyed Damselfly						
1102	Erythromma viridulum	Small red-eyed damselfly						
1301	Ceriagrion tenellum	Small Red Damselfly						

Recorder (s)	No.				
Card Compiler	No.				
Source of record					

Estimated Numbers		Key to column headings		Habitat / Comments
A	1	Ad	adult	
B	2-5	Co	copulating pair	
C	6-20	Ov	ovipositing female	
D	21-100	La	larva	
E	100-500	Ex	exuvia	
F	500+	Em	pre-flight emergent	
X	Present			

Code	Anisoptera (Dragonflies)		Ad	Co	Ov	La	Ex	Em
1502	Gomphus vulgatissimus	Club-tailed Dragonfly						
2101	Brachytron pratense	Hairy Dragonfly						
2201	Aeshna caerulea	Azure Hawker						
2204	Aeshna juncea	Common Hawker						
2207	Aeshna grandis	Brown Hawker						
2209	Aeshna cyanea	Southern Hawker						
2210	Aeshna mixta	Migrant Hawker						
2211	Aeshna affinis	Southern Migrant Hawker						
2212	Aeshna isosceles	Norfolk Hawker						
2401	Anax imperator	Emperor Dragonfly						
2501	Hemianax ephippiger	Vagrant Emperor Dragonfly						
2601	Cordulegaster boltonii	Golden-ringed Dragonfly						
2701	Cordulia aenea	Downy Emerald						
2802	Somatochlora metallica	Brilliant Emerald						
2804	Somatochlora arctica	Northern Emerald						
3201	Libellula depressa	Broad-bodied Chaser						
3202	Libellula fulva	Scarce Chaser						
3204	Libellula quadrimaculata	Four-spotted Chaser						
3302	Orthetrum coerulescens	Keeled Skimmer						
3309	Orthetrum cancellatum	Black-tailed Skimmer						
3601	Crocothemis erythraea	Scarlet Dragonfly						
3801	Sympetrum vulgatum	Vagrant Darter						
3803	Sympetrum striolatum	Common Darter						
3805	Sympetrum nigrescens	Highland Darter						
3807	Sympetrum fonscolombii	Red-veined Darter						
3809	Sympetrum flaveolum	Yellow-winged Darter						
3810	Sympetrum sanguineum	Ruddy Darter						
3812	Sympetrum danae	Black Darter						
3903	Leucorrhinia dubia	White-faced Dragonfly						
9998	Anax junius	Green Darner						
9999	Anax parthenope	Lesser Emperor						

These should ensure that all relevant data are captured at the point of observation wherever possible. They should be backed up by clear instructions on their use.

Training for field surveyors in the aims and methods of the survey.

Training for field surveyors in field identification and use of literature (e.g. how to use scientific names correctly, interpretation of a recording entity comprising an aggregation of species, how to recognise hybrids etc.).

Guidance in how to use GPS or map reading to ensure accurate map refs. (e.g.: know about the potential inaccuracies of GPS).

Clear procedures for and reasons why, when and how to collect voucher specimens, and how to handle them, where to send them, in what way (e.g. dried plants in absorbent paper, not in plastic bags).

This is not the place to issue detailed guidance on survey design or data capture methods. However, attention to details, such as providing guidance to field surveyors as to the way that recording is to be carried out, is an essential step in ensuring the quality of the resulting data. One example might be: the pre-definition of the way that "aggregate species" are to be treated in plant recording, so that the resulting data represent equivalent levels of definition from different field workers.

6.2 Identification and verification

Being sure of the identity of the thing recorded is obviously crucial. Responsibilities for and ways of ensuring accurate identification are not just the province of the field recorder, and fall into discrete areas:

Field identification

For species recording, the capacity of field observers to identify what they are recording is obviously a controlling factor, as outlined above. But how can an organisation be sure of species identifications, and to what level is it possible or necessary to go?

Attempts to set up standard qualifications to define individuals' capabilities have been proposed and attempted (e.g. the Natural History Museum's "IdQ" system). Although training is highly important, such qualifications often fall down on a number of counts: inability to impose a rigid framework on volunteers or staff across a wide range of organisations; change of a person's ability and experience over time; different levels of capacity of an individual with different taxonomic groups; or differences in a person's capability in different geographical areas or even in different habitat types.

"Peer review" (judgements about records made by experts in the relevant field) has tended to be the way individuals' capacity to record, and therefore the reliability or otherwise of their records, has been judged. It is open to abuse, though, such as through favouritism or personal bias, or merely through an organisation's lack of knowledge about the capacity of its recorders.

The questions of who judges the capabilities of field recorders and how are crucially important in the operation of any recording scheme or survey, and need particular attention from the outset. Some recording organisations use a somewhat formalised "checklist" approach:

- ***Beginner*: little experience, and with low levels of use of identification facilities or knowledge of methods; only common or easily identifiable species records acceptable without other evidence.**
- ***Experienced*: with good levels of field experience, possibly limited by geographical region or habitat types, but with access to adequate literature and facilities; records of most readily-identifiable species acceptable.**
- ***Expert*: with wide and deep understanding of their particular groups, good access to relevant literature and facilities, usually networking with others in their field; most records accepted, except some taxa needing critical determination.**
- ***Authority*: a nationally or internationally recognised expert in the determination and taxonomy of a particular group, operating alongside extensive reference material and other authorities; definitive judgement on identifications, except where taxonomic disagreements might occur.**

In practice, application of such a system can be difficult, for the reasons given above. However, an experienced survey organiser may find that such a checklist can be used as a guide to the way they form a judgement. In addition, they will need to consider the capability of an individual to learn and develop their knowledge, or whether a hitherto accepted level of expertise may be declining, through age of the individual for example, or new advances in the subject. There have been calls to make this process "transparent", with formal accreditation of recorders. However, others have pointed out the very human issues involved in formalising this kind of system, not least in the face of possible legal action for "defamation", and that focusing on the record rather than the recorder is best, with the willingness of the recorder to collect and submit a voucher specimen or other evidence being the test of whether or not their data are likely to be reliable. "Mentors" can also be used to help newly recruited recorders improve.

Whatever way a scheme or survey approaches this difficult issue, there can be an advantage in spreading the load of making such judgements. A regionalised, or partnership approach is one way of doing so. In doing this, it is advisable to put in place a more formal structure, with clearly defined roles and lines of communication, and to produce a protocol or "code of conduct" for how the system is intended to work, so that all those involved can see where they fit and how judgements are made.

Verifying records

To be clear: **verification** of a record is to do with the accuracy of the identification of the thing being recorded - either a species or other factors, such as habitats.

To augment a survey or recording scheme's assessment of recorder capabilities, there needs to be an agreed process of verifying incoming records, where necessary, so that any judgement about a particular record can be carried out without overtly calling into question the capabilities of the recorder. If a clearly publicised system is put in place, it can then be used to adjudicate objectively over records where needed.

For recording species, either as part of a recording scheme or alongside other survey work, this could involve a number of different activities:

- **Recording schemes or organisations setting up a survey have a responsibility to take the lead with setting standards for identification. They should define agreed levels of "difficulty" over the identification of the species being recorded. Checklists defining level of difficulty for each taxon should be produced, alongside a degree of competence (defined in terms of the skill level of the identifier) at which an identification would be acceptable. Geographical variation in these designations may need to be recognised.**
- **The scheme or organisation should define whether or not a voucher specimen or other evidence needs to be collected and determined by an expert or panel of referees at an appropriate level for particular species. This should include advice on how a voucher or other evidence should be collected and how they should be submitted for determination. It should also specify ways that such vouchers or documents are to be maintained for the future and who does this.**
- **Agreed panels of experts for particular taxonomic groups should be established where possible: in relation to particular species groups, geographical areas; or for use during the process of a particular survey;**

and the level at which these experts will operate should be defined (e.g. at a county, regional or national level).

- **Agreed protocols on the use and support of these functions need to be produced, including clear levels of responsibility for carrying them out at different points in the survey or data gathering process (see section on Data Flow above).**
- **Clear mechanisms should be established during data management for documenting decisions made over the verification of particular records, including details of by whom, when and why decisions were made.**

Identification: ways to ensure data quality

Focus on the accuracy of the record, not the recorder.

Use "checklists" of competence carefully, and as a guide, not a "last word".

Set out and publicise clear guidance on what are "critical" species/taxa for identification, and what are not.

Set out clear requirements as to when and how voucher specimens or other evidence need to be collected and submitted to named experts.

Have clear procedures in place and make sure databases have the capacity for documenting decisions on identifications: who did them, when, and how.

Produce guidance on the way species (or other things) should be identified, and set up training for field recorders.

Consider the use of "mentors" to help newly-recruited recorders.

Set up panels of referees or experts for referring "difficult" cases.

Publicise the way that a survey or recording scheme aims to handle the question of identification.

For recording habitat or physiographical features, the questions are rather different, because the entity being recorded is not definable in quite the same way as an individual organism. Checking recorded details against likely or expected features can be a basis of data verification in these cases. These might include:

- **Checklists of species used for defining habitat types, including proportions of populations.**
- **Checklists of attributes of habitats (e.g.: structure, water levels, humidity, pH).**
- **Mechanisms for comparing known occurrence of habitat features against new records.**

Having a well-publicised and transparent process of record verification in place from the outset safeguards a recording scheme or survey from doubts about its quality control and methods, as well as distancing the process of verification to some extent from problems of human relationships. Other aspects of records also need attention during the process of recording, and can to some extent be verified in a similar way. These include recording of geographical locality, date, sample sizes, etc.

6.3 Quality control during data management

The roles of the data compiler, and of the person carrying out subsequent data management, are also crucial in the process of ensuring data quality. While the basic facts of a record can be controlled to a great extent before or at the point of the record being made, "data capture" (entering a record into a computer system or database), and "data manipulation" subsequently are both potential sources of error, and can be improved by better ways of working, or by the use of automated tools. This is not the place for detailed guidance on data management itself. However, a few key issues need to be highlighted in relation to data quality control:

- **ensuring that data management processes do not over-ride or impair the integrity of captured data (e.g. through automated database processes);**
- **designing data management processes that deliver data in ways that are appropriate for the subject and of direct use to the data users;**
- **ensuring proper documentation of data management processes that have been carried out, and that this information remains with the data.**

At different stages in the data management process there will be different quality control issues that need to be considered:

- ***Data collation*** **(mechanisms used to do this need to maintain details of provenance, intellectual property etc., as well as maintaining the specific integrity of identifications, locality data etc. contained within original records). Archiving of original records is also needed.**
- ***Data manipulation*** **(the capacity for such activities to remove valuable parts of records through imposition of "standardised" formats etc., or for automated operations to "scramble" data need to be guarded against).**
- ***Data analysis*** **(the application of analytical tools needs to be appropriate for the kind of data being used to avoid spurious conclusions or summaries being produced).**

Use of tailored data management systems will help in this, especially those designed with wildlife data quality in mind from the outset, although none of the existing systems are perfect. The JNCC's 'Recorder' is a *de facto* standard for data capture and data management. It has some in-built data validation checks, but more are being considered. Proposals were made at a Local Records Centre seminar in Edinburgh in November 2005 that record categories should be standardised:

- Correct
- Considered likely to be correct
- Possibly correct [= unconfirmed]
- Considered likely to be incorrect
- Incorrect
- Not yet checked

Use of other proprietary databases may be satisfactory, but greater attention to details such as ensuring proper management and checking of dates, taxonomic names etc., may be needed when setting them up.

Compiling data: quality control checks and procedures

Aim to acquire raw data in standard formats (e.g. standard recording forms).

Ensure all necessary verification procedures have been carried out, preferably before computerisation of data.

Consider using quality-control checks on data entry (e.g. double-entry).

Use standardised data entry systems (e.g. purpose-built databases or adapted spreadsheets, with in-built taxon checklists, habitat codes etc.).

Use recognised standard term lists, taxon checklists, habitat codes etc. wherever possible (e.g. NBN Species Dictionary).

Ensure all relevant parts of records are retained during data capture, including details of determinations, locations of vouchers, sources of records etc. Arrange for original records to be archived as a back-up.

Aim for standardised data formats (e.g. dates, place-names, uniform formats of locality details, personal names).

Carry out data validation routines on data entry (grid refs, dates, sources).

Remember it is easier to correct a record at the start than it is to expunge a faulty record once it has been disseminated.

Data validation

Validation is the term applied to the process of carrying out standardised checks on the “completeness”, and “validity” of the content of a record. Working practices and mechanisms to ensure that species or other facts are properly recorded in the first place can be supplemented by automated validation during data management, e.g.:

- **Appropriate use of taxonomic names and authorities.**
- **Identifications validated against checklists.**
- **Statuses of taxa correct.**
- **Format of grid references correct.**
- **Grid references checked against counties/vice-counties or other defined geographic areas.**
- **Site names checked against standard gazetteers.**
- **Formats and contents of dates correct.**
- **Dates checked against survey periods.**
- **Observer/compiler/determiner names checked against standard lists.**
- **Validity of record sources checked.**

The NBN has focused a lot of effort in these areas through the promotion of the **NBN Data Standard**, and through developing methods and tools for handling data collation etc., in particular the **NBN Data Exchange Format** and an automated **Data Validation Tool** for carrying out basic routines on collated datasets. These are available from the NBN website. Alternatively, techniques for carrying these out may be available from existing institutions, such as the UK Biological Records Centre or local records centres, or can be developed in-house.

The NBN Exchange Format Validation Tool

The NBN Trust has developed a programme to validate datasets that providers send to the Gateway in the NBN Exchange Format. It can also be used to check datasets in this format for other data exchange purposes. It does the following automated checks:

- Ensures that all mandatory columns are present (e.g.: date, species code etc.).
- The correct combinations of columns are present. (e.g.: for grid reference: either 'gridreference' or 'Easting' and 'Northing' are present but not both).
- Each row of data has the correct number of fields.
- Dates are supplied in a standard format (dd/mm/yyyy).
- The end date is after the start date.
- Dates are valid in the calendar sense (e.g. 29th February is in a leap year).
- Grid references are in the correct format (either standard Ordnance Survey for Great Britain: TL207795; Ordnance Survey Ireland: T213392 etc.)
- Values in the 'Projection' field are correct (e.g. OSGB, OSNI, WGS84 etc.).
- Values for 'sensitive' records and 'zero abundance' are either 'true' or 'false'.
- Values in the 'Precision' field are correct for the grid reference precision given (in a 10m - 10000m resolution range).
- Each row has a unique 'RecordKey'.
- Fields that should contain numbers just have number values.
- Values in a field are no longer than the maximum length allowed (e.g. Site Names up to 80 characters).
- Taxon version keys are present in the NBN Species Dictionary.

The validator checks each row in turn and reports which rows in the dataset have errors and what the correct values should be. It can be set to stop to allow these to be corrected, or can tabulate them for later attention.

The validator also maps the dataset as a final check to ensure the distribution of points is what the data provider expected (no nasty surprises when they see it for the first time on the Gateway). Suspicious points on the map can be selected and the record details viewed to identify which records may be wrong.

6.4 Data quality and the data custodian

The role of a data custodian in maintaining and promoting data quality is especially important at the dataset level. Their role is to ensure that proper processes are carried out in maintaining data, and in such a way that the data can be communicated readily to others. Providing data to third parties therefore also includes the need to address data quality issues. A key aim here is that the communication of information or data should be as transparent to the user as possible, enabling them to be as sure as they can be that the data they are using are fit for the purpose for which they intend to use them. The NBN Trust, through the setting up of the Gateway, has attempted to address many of these issues, but other bodies handling datasets and passing data to users, either in pre-digested form or as raw data, should ensure that quality control measures are being addressed.

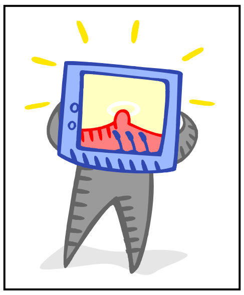
Actions could include:

- **Maintain adequate documentation about the accuracy, within definable limits, of identifications, including:**
  - **re-determinations or levels of taxonomic application where these are important to the way the data are to be used;**
  - **use of standard definitions of habitats/biotopes;**
  - **ensuring standard documentation of other attributes, such as dates, sampling methods etc.**
- **Make sure that the appropriate level of detail to which the data may be interpreted is clear to users (such as the level of resolution of the original survey, or the extent of coverage of a survey, temporally or geographically).**
- **Ensure the retention and communication of quality information from data providers or third parties.**
- **Document clearly information on the provenance of data, so that users can make their own judgements about its authenticity, as well as allowing them to make appropriate acknowledgments.**

Some aspects of this need attention to the requirements of things like the **Data Protection Act**, or **Copyright** legislation, which may limit what can be done with important information relating to the quality of data. Detailed guidance on these has already been issued by the NBN Trust.

A data custodian may or may not be the original compiler of the data. If they are not, then a data custodian needs to ensure that their practices in data management are agreed with the data provider, and any data quality processes carried out are appropriate to their needs.

6.5 Data dissemination and data quality



The business of disseminating data itself is beyond the remit of this guidance. However, the process of dissemination needs to reflect and uphold the quality issues that have been addressed during the data capture and data management processes. There are many ways to communicate data between a custodian and a user, and some of these will be specific to particular situations, while others are more general. In any case attention to some basic principles is important in maintaining overall data quality and confidence in the use of the data.

The most important tool for describing and communicating information about data quality is "**metadata**". Metadata is a mechanism for documenting the source and characteristics of datasets of any sort, but especially electronic data. It aims to produce a standardised description of the data, with details of what the dataset consists of; why it was made and by whom; who owns it; and its reliability. This metadata description should be retained alongside datasets to ensure that future users can understand the origin of the data, and therefore understand restrictions on and purposes for which they can be used.

The NBN Trust was set up to enable better data communication, and its Gateway is a prime mechanism developed to do this. For dissemination of data through the Gateway, the Trust has focused on the concept of making data of "known quality" available, and has promoted the use of standard metadata to address at least the basics of this. NBN metadata follows minimum requirements to conform to the "UK GEMINI" standard. This enables holders of data that relate to geographical areas to standardise the way data are described.

The standard NBN metadata format records information on:

- **Name of the dataset.**
- **Name of the dataset provider.**
- **Subject of the data.**
- **Methods of data capture.**
- **Purpose of survey or data capture.**
- **Geographical extent of survey.**
- **Time span of survey.**
- **Outline of ways in which data were checked.**
- **External sources of information about the data.**
- **Access and use constraints.**

One prime aim of the metadata is to enable a dataset that is provided through the NBN Gateway to be judged for its reliability. However, standard metadata of this type can be used in other situations, and is recommended as good practice generally.

Guidance on compiling NBN standard metadata has been issued by the NBN Trust separately, and is available through its website.

7. Who should be doing what to support data quality?

In Section 4, it was suggested that everyone involved in the recording and wildlife data process should have at least some responsibility for ensuring data quality.

However, it is possible to identify some kinds of organisations that are best placed to carry out some of the specific roles and tasks that have been identified above.

National Societies and Recording Schemes

These organisations (and individuals) have a key role to play in underpinning species data quality in the UK. They are usually the focal point of taxonomic understanding of their subject, and are in a pivotal position to be able to influence the quality of records and recording. However, their resources are often not enough to sustain some of the work this might entail, and this is an area that needs further support and strengthening in many of them if they are to take on these roles more formally.

Recommended actions

Bearing this caveat in mind, Societies and Schemes should be in a position to:

- ***Develop and clarify survey objectives and needs for a particular taxonomic group, and identify recommended sampling and field survey methods.***

[Such guidelines should be promoted not only through the society or recording scheme concerned, but more widely, so that other potentially interested bodies can tailor their methods and activities to suit accordingly.]

- ***Draw up standard lists of species for groups, which define those that are "critical", requiring expert determination at respective levels; those that are acceptable from "competent" recorders; and those (if any) that are acceptable from other sources.***

[These checklists should be made available both to volunteers and others in the recording schemes themselves, as well as to third parties to improve processes of recording elsewhere.]

- ***Formulate and keep up to date potential panels of referees or experts to whom records requiring validation might be referred.***

[This may be an impossible task for many groups, owing to a lack of people with the relevant expertise, and the potential for an overload, so that such referees may only be available to members or upon payment of a fee. However, in some groups it may be possible for local or regional panels of referees to be established, in collaboration with local groups or local records centres, to share the load.]

- ***Produce guidance on the collection, processing and housing of voucher material for a group.***

  [This should include advice on preparation and curatorial techniques, as well as on the potential housing of accumulated collections for reference. There is much scope for collaborative work on this between societies and with external institutions, such as museums and local records centres (see below).]

- ***Produce protocols for the documentation of records to assure data quality.***

  [Such protocols should not only relate to the way the Society or Recording Scheme carries out its own data management, but also give advice to others handling data in these groups.]

- ***Publish general guidance on recording in their taxonomic groups, including field recording methods, roles and responsibilities for identifications, training etc.***

- ***Aim to rationalise the processes by which data from other bodies, such as local records centres, might be verified.***

  [For example, data from a local records centre could be validated remotely by Society referees or vice-county recorders, using the NBN Gateway. In exchange, a local records centre could come into an agreement to handle automated data processing and validation checks for relevant Societies and Schemes at the local/regional level.]

Several national societies are either in the process of drawing up such guidance, or have already done so to some extent. Co-ordinated promotion of such guidance is needed for the benefit of a wider community.

♣

Local Records Centres
(and related organisations, e.g. local natural history departments of museums)

Local records centres, where they are fully-functional, may already have a strong role in promoting data quality among their own volunteer recording community. However, this is often carried out independently of other organisations, and integration of their efforts with those of the national societies and recording schemes would be particularly beneficial. However, while data quality may be important internally for the operation of a particular centre, the centre may not be supported adequately to underpin a wider remit, and this may be an area which requires strengthening and further support, particularly through encouraging its primary sponsors to recognise these roles as central to its operation.

Recommended actions

Local records centres especially could:

- ***Re-examine their data quality and data management methods to see if improvements can be made.***

  [Many records centres will already be carrying out many of the processes highlighted in this guidance. However, moving towards the standards that allow easy data exchange through the NBN Gateway may need improvements in some areas].

- ***Establish local panels of referees, in partnership with local specialists.***

  [Many LRCs already have these. They can oversee records from their areas, according to agreed criteria, but in some cases may need to be integrated with the relevant national society or recording schemes so that levels of capabilities and acceptability of records can be agreed, and processes can be standardised].

- ***Enter into data capture, data management and quality assurance agreements for data from other organisations or individuals*.**

  [These roles could be especially useful in ensuring that data from other local sources are brought in to agreed processes of data verification and validation].

- ***Carry out data capture and other automated data validation processes on behalf of local individuals or groups*.**

  [This could include handling feedback from NBN Gateway validation routines on behalf of local groups].

- ***Instigate training in recording at the local level.***

  [Again, for this to work most effectively, collaborative work with the relevant local or regional representatives of the national societies would be beneficial].

- ***Enter into partnerships with relevant organisations to maintain local or regional facilities for receiving and managing necessary voucher material in support of records.***

  [This requires partnership development with, in particular, local or regional museums and the development of agreed criteria for identifying the need to maintain vouchers].

- ***Provide proper documentation and metadata to users alongside their own and third party data supplied to others, e.g. through the NBN Gateway.***

♣

Non-governmental biodiversity organisations

There are a wide range of non-governmental biodiversity or conservation organisations that collect data, both at national and local levels, such as wildlife trusts, the National Trust, Woodland Trust, RSPB and so on. Many, especially larger ones, already have sophisticated survey and data management practices in place, but some of the smaller ones may not. Even if they have, they may not have addressed some of the data verification and validation issues outlined in these guidance notes.

NGOs also may or may not communicate effectively with existing networks of information, at the national or local levels. It would be especially beneficial for their recording to be more fully integrated with those of both the national societies and recording schemes on the one hand, and with local records centres on the other. Putting in place mechanisms to make use of these networks to verify and validate their data might be one way of doing this.

NGOs might also need to integrate their approach with other activities underpinning data quality, such as identification training, issuing guidance on survey methods, collection of voucher specimens, etc.

Recommended actions

- ***Establish data management and data validation agreements with relevant national societies and recording schemes and local records centres.***

- ***Work with appropriate national societies and local records centres to develop identification training for their staff and volunteers.***

- ***Work with relevant societies and schemes to develop agreed methods for and guidance on surveys and recording for use within their organisations.***

- ***Develop and publish protocols for the dissemination of their own data, e.g. through the NBN.***

♣

Statutory and other official biodiversity organisations (including academic departments, research institutions etc.)

These organisations have a range of roles in relation to the maintenance of data quality, including data verification and validation. These include:

- Providing support for existing networks of organisations carrying out survey, data verification and validation roles.
- Collecting and managing their own data.
- Making use of data for strategic, research and management purposes.
- Making their data available for third parties.

Recommended actions

It is not possible to produce detailed recommendations here for the verification and validation of data collected or held by this wide range of bodies. However, it is worth reiterating the points made in Section 5:

- ***Data collected by professional or other official organisations should be subject to as rigorous quality checks as those recommended for voluntary sector or local recording organisations, especially if the data are to be disseminated to others.***

- ***These organisations should consider making their data available for "peer review" by relevant experts where necessary before they are made available to others.***

It would be desirable if these organisations could integrate their data verification as much as possible with the existing specialist networks that underpin data quality, especially with the relevant national societies and recording schemes. However, it would be unreasonable to expect voluntary bodies to undertake substantial data verification processes for official organisations without material support. Potential actions might therefore also include:

- ***Work with national societies and schemes to integrate recommended verification and validation practices relating to specific subject areas into their internal systems.***

- ***Examine further ways to support the role of the key specialist organisations in carrying out this work.***

The NBN Gateway's Data Validation function and the NBN Data Validation Tool are two mechanisms that have been developed to help organisations with these activities, especially if data agreements with societies and schemes include use of the former as a means of carrying out validation of datasets by relevant experts.

Another role of some official organisations is often overlooked, and that is the vital role that museums, some university departments, botanic gardens and their key staff play in verifying data through identification of specimens and provision of access to reference collections and libraries. **The recently accelerating tendency for these facilities and expertise not to be retained or replaced needs to be reversed if data quality overall is not to suffer**. An action for statutory and other official organisations involved with biodiversity data in support of this role might be:

- ***Promote partnership arrangements between biological recording organisations and relevant institutions for the maintenance and use of biological reference collections and research facilities.***

Finally, Conservation Agencies in particular, and especially the Joint Nature Conservation Committee, have a special responsibility for assisting the biological recording communities to improve the standard of their data, particularly bearing in mind their capacity to influence the systems of data management currently available. A potential specific recommended action could therefore be:

- ***Improve the capabilities of data capture and data management software, e.g. by allowing individual records to be "tagged" with agreed levels of acceptability; enhancing biotope recording functions etc.***

♣

Commercial and professional biodiversity organisations

Commercial ecological consultancies and other professional bodies have roles in collecting, managing and using biodiversity data that need to be recognised. Questions of data quality will exist with all their data, just as it does in other bodies. Enabling them to tap into the data verification and validation network available to the voluntary and official sectors may present difficulties, but the benefits would be considerable, enabling their data to contribute to the pool.

Recommended actions

Some potential actions might include:

- ***Set up formal agreements over access to data with and sponsorship of voluntary organisations responsible for data verification.***

- ***Establish partnerships with local records centres or other biodiversity organisations to enable commercially acquired data to be managed, validated and made more widely available, e.g. through the NBN Gateway.***

- ***Professional institutions supporting the commercial biodiversity sector (notably the Institute for Ecology and Environmental Management) could issue codes of conduct and professional guidance in support of data quality for use by commercial bodies.***

Case Studies: 1

(An example of survey design and metadata upholding data quality)

The Survey of Bryophytes of Arable Land (SBAL)

SBAL was set up in 2001 by the British Bryological Society to get baseline data on the distribution and ecology of bryophytes in tilled land in the UK.

The field survey

Clear project aims and a sampling strategy were defined:

- To survey single fields with crops or fallow soils.
- To survey in autumn, winter or early spring.
- Two fields each to be selected from 100 random tetrads in areas with at least 15% arable land use.
- If suitable fields in random squares were not found, nearby suitable fields were substituted. In addition "ordinary" fields were visited by field workers not able to visit random ones, as well as "special" fields with rare species.

Occurrence of species was augmented with DOMIN abundance data.

Field surveyors were issued with a pack containing guidance notes, identification aids and standard record cards. Training in field survey was set up, with specially run field days.

Progress reports on the survey were put on the BBS website and in the Society's newsletter, and in later stages of the survey participants were additionally encouraged individually to complete the survey.

Data collation

Field record cards were submitted to the Biological Records Centre for processing as the survey progressed.

Initial cards returned were checked by the scheme organisers for compliance with field methodology, as well as for identification.

Data were captured in yearly batches by experienced data processing staff, using standard data inputting software for entry into an Oracle database.

Compiled data were subsequently checked using an Access database, with locality data checked visually once, and species lists for each locality separately, using the BRC species numbers used for data inputting as an auto-generator for species names.

Data analysis and reporting

11,061 records were generated from the data received. Data were analysed to produce a classification of arable field assemblages.

The survey report noted limitations of the survey, especially lack of associated information on habitat management, heterogeneity of habitat within the sampled fields, and differences in evenness of recording.

Distribution data were amalgamated with other BBS data and disseminated through the NBN Gateway, although the dataset metadata do not highlight the SBAL data.

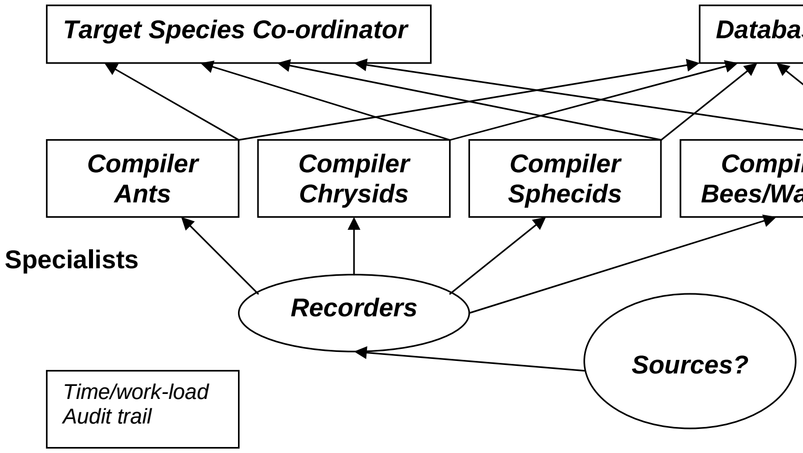
Case Studies: 2

(An example of a data flow system and data verification in a smaller scheme)

The Bees Wasps & Ants Recording Society (BWARS) and their data management and verification system

BWARS concentrates its recording on producing national distribution data for atlases of species. Recording is focused through short to medium term "projects", focusing on groups of species, which are then used to produce atlases. The production of an atlas is seen as a primary spur to encouraging recording.

The Society has an agreed, integrated process of managing data, including processes for data verification and validation:



Data are mainly received electronically, in various formats, and are integrated into a standard database.

All records are checked by recorders, and by the species group compilers. Doubtful records may be followed up by visits to the site from more experienced field workers.

Species identification is a particular concern, as some groups of species lack accessible identification literature, although this is improving. Recorders' competence is largely measured by a "peer review" process.

Requirements for the submission of voucher specimens for "critical" species are defined, although the species concerned depend on the level of experience of the recorder.

Training in identification is carried out, and new recorders are encouraged to focus on small groups first.

Data from outside sources, e.g. local records centres, may not be acceptable, unless they have in place a process of collecting vouchers.

Case Studies: 3

(Data verification and validation in a larger local records centre)

Sussex Biodiversity Record Centre (SBRC): verifying and validating species data.

SBRC regards data verification to be one of the most important, but also one of the more difficult tasks it undertakes, particularly because data it uses may be used for important land-use decisions.

The Centre can receive up to 100,000 **new** records a month, mostly in digital form, so a targeted approach has to be taken to quality checking.

Automated data validation is carried out during the data import process to the Centre's Recorder 6 database, relying on its in-built date, grid reference and name checking capability.

Data verification involves partnership working. Because of the quantity of records being received, and because most essential use is focused on them, a formal policy decision has been taken to focus effort on rare (at the Sussex level) and protected species.

Criteria for defining locally rare and threatened species have been developed, in collaboration with local specialists. These are combined with national designations to form a list of some 3,000 "critical" species.

All data received each week are filtered against this checklist of species, and records for species meeting these criteria are manually reviewed.

The following questions are used as a basis for the review:

- Has the species been recorded here before?
- Is this location a likely one for the species?
- Who has recorded it?
- Are there special problems with the identification of this species?
- Is this record already known to local specialists?
- Do these experts need to verify the record further before it is used?

Data that may need further verification are submitted to local specialists by electronic spreadsheets. If these specialists require further checks, they follow this up with the Centre and/or with the original recorder.

Case Studies: 4

(Data validation and verification in the UK Biological Records Centre)

Processing datasets submitted to the UK Biological Records Centre

Automated validation routines are applied:

Species identifications

- Valid BRC species code used.
- BRC species code is for the appropriate taxonomic group.
- Any code used to flag species identification issues is valid.
- Any code used to explain record status is valid (e.g. native or introduced).

Location information

- Grid reference is in a valid format (e.g. TL22; 52/22).
- Any assigned 10km square value matches the grid reference provided.
- Any tetrad value provided is a valid 'DINTY' letter.
- Any 'DINTY' tetrad value given is correct for the grid reference given.
- Any code used to denote the quadrant of a 10km square is valid.
- Any quadrant value of a 10km square corresponds with the grid reference provided.
- Any code used to flag particular spatial data issues is valid.
- 10km square is on land (applicable to squares in Britain as well as Channel Islands and Ireland), and for 2km or 1km square grid reference (in Britain only).
- Valid Vice-county code.
- 10km square is in its corresponding Vice-county (applicable to squares in Britain as well as Channel Islands and Ireland), and for 2km or 1km square grid reference (in Britain only).
- Trim any extra spaces from locality name.

Date information

- Year is in a valid, four-digit format.
- Valid day and month used.
- Where values for day are provided, values for month are also provided.
- Where a year range is given the second year is after the first; all data in form 'before NNNN' (including publication dates); 'after NNNN' to be converted to ranges.
- A code used to explain dates given is valid.

Other information

- Name for recorders, determiners and compilers are in standard canonical form (e.g. Hill, M.O.); conversion to this form may be done at least partly algorithmically.
- Source of the record is validly coded (for field, museum etc).
- Where the record is from literature, the literature reference is stored.
- Altitude is within a valid range for measurements in metres.
- Any code denoting the type of recording card is valid.
- Any code denoting a particular type of record (e.g. droppings, tooth marks) is valid.
- Where habitat coding systems are used, any code denoting a habitat is valid.

Metadata are generated for each dataset, including: a brief description; name of data supplier; why the data were collected and how; what geographical area the data cover; what time-period they cover; and notes on the quality of the data, how they have been checked etc.

Automated processes to assist in the data verification process are also carried out: new Vice-county records; new 10km square records.

Reports and formatted copies of the checked dataset may be sent to the data supplier, identifying any necessary corrections to be made, before incorporation in the BRC database.