Policy: Data Quality

1. POLICY STATEMENT

1.1. Cofnod (North Wales Environmental Service) is a not-for-profit company which provides environmental information products and services (particularly relating to biodiversity) to a range of users for the benefit of the environment of North Wales.

1.2. It is important for Cofnod to be able to ensure that Data used within its products and services are of the highest possible quality, and that the quality of each individual piece of Data is clearly identifiable to Data users. Therefore:

1.2.1. Data supplied to Cofnod is carefully managed. See Cofnod's policy on **Data Sourcing, Management and Use** for more details, available from Cofnod's website www.cofnod.org.uk/Library or on request.

1.2.2. Cofnod stores Metadata for each Dataset with which it has been supplied, and makes this Metadata available to potential users of the Data to enable them to assess its suitability for their proposed use.

1.2.3. A system of Dataset Categorisation has been implemented, and this information forms part of the Metadata for individual Datasets, which are made available to potential and actual Data Users.

1.2.4. Processing of biological records in electronic format, and entry of such records from paper sources, are both carried out in a consistent manner, using specifically designed Data Import and Data Entry systems and in accordance with set procedures. A Data Validation tool is used to screen all electronic Data being imported into Cofnod's database of biological records.

1.2.5. Individual biological records are assessed according to a Data Verification Process developed by Cofnod. Whether this process is applied to an individual record or not depends on the category of Dataset to which it belongs. Cofnod will actively seek assessment of records within relevant Datasets, subject to any restrictions imposed by Data Supplier(s).

1.2.6. The use of individual biological records in specific products and services is governed by the category of Dataset to which they belong, and their Verification Level if applicable. An indication of the quality of individual records will be provided to all data users.

2. DEFINITIONS

2.1. Within this document, the terms '**Data**' or '**Dataset'** mean a collection of biological records or other relevant environmental data, for example habitat maps, statutory or non-statutory site boundaries relating to biological or geological features. Datasets may be stored in a variety of electronic or paper based formats, often using a combination of several methods.

2.2. The term '**Metadata**' is used to mean information about the Data, such as collection methods used, purpose of Data collection and sources of individual records.

2.3. Datasets are owned, collated or managed by a **'Data Supplier'**.

2.4. **'Data Verification'** refers to the process of ensuring the accuracy of identification of whatever is being recorded (within this document it refers to identification of a species or other taxon). Records are assessed by relevant experts – these are usually County Recorders and are sometimes referred to as '**Verifiers'** - as part of the '**Cofnod Data Verification Process'.**

2.5. **'Data Validation'** refers to the carrying out of standardised, often automated checks on the 'completeness', accuracy of transmission and validity of the content of a biological record.

3. STATEMENT OF PROCEDURE

3.1. *Data Management*

3.1.1. Cofnod will comply with its policy on **Physical Security of Data** (available from Cofnod's website www.cofnod.org.uk/Library or on request).

3.1.2. Data supplied to Cofnod will be carefully managed using a range of custom built databases and tools.

3.1.3. Cofnod's Policy on **Data Sourcing, Management and Use** provides further details on its Data Management systems (available from Cofnod's website www.cofnod.org.uk/Library or on request).

3.2. *Metadata*

3.2.1. Cofnod stores comprehensive Metadata on datasets which it holds. Metadata is stored within a customised database known as CAS (Cofnod Administration System - further information is provided within Cofnod's Policy on **Data Sourcing, Management and Use,** available from Cofnod's website www.cofnod.org.uk/Library or on request), much of which is available online via Cofnod's website at www.cofnod.org.uk (select the 'Data We Hold' page).

3.2.2. Viewing Metadata which relates to an individual record may assist Data users in assessing its quality.

3.3. *Dataset Categorisation*

3.3.1. A system of Dataset Categorisation (see table 1) has been developed which forms part of the Metadata stored within the Cofnod Administration System.

3.3.2. Datasets categorised **1** to **3** are believed to be of high quality and all records from such datasets will be available for use within Cofnod's database of biological records. If data quality issues are identified with individual records from such datasets then these will be investigated and the dataset category might be changed to **4** in certain circumstances. Records from category **4** datasets are assessed within the Cofnod Data Verification Process (see below for more details) before they are made available for use within Cofnod's database.[1]

3.3.3. Datasets within category **5** are awaiting further consideration and will be moved into one of the other 8 categories over time.

3.3.4. Datasets within categories **6** to **8** will not be incorporated into Cofnod's main database.

[1] NB: Records received in spreadsheet/database format which were imported into Cofnod's database before 16/12/2012 were not assessed through the Cofnod Data Verification process, but will be over time. Unless they have been verified by a local expert they are labelled as Unassessed – see more under 3.6.

To be incorporated into the Cofnod Database	1	Dataset originates from Biological Records Centre or National Scheme/Society
	2	Dataset originates from County Recorder
	3	Dataset originates from other trusted source
Individual records to be assessed by a relevant expert, according to the Cofnod Data Verification Process, before incorporation into the Cofnod Database	4	Dataset awaiting/undergoing assessment as part of the Cofnod Data Verification Process
Temporarily archived	5	Dataset of unknown quality/relevance pending further consideration
Not to be incorporated into the Cofnod Database, archived.	6	Dataset of poor quality.
	7	Dataset contains irrelevant data.
	8	Dataset contains records already stored in Cofnod's Database.

Table 1: System of Dataset Categorisation

3.4. ***Data Processing and Dilys***

3.4.1. Most Data on Cofnod's database of biological records have been submitted to us as electronic datasets which require varying amounts of Data Processing before they can be incorporated into Cofnod's database.

3.4.2. Data Processing is carried out by trained staff and volunteers. An internal procedure is referred to, especially while new staff and volunteers are being trained, to encourage consistency. The original data file is kept within Cofnod's Dataset Archive, together with all versions of the data saved during processing. Original IDs are always added to each record (if not already part of the dataset) to ensure that records can be traced back to the original version submitted to Cofnod. Comprehensive notes are also stored in Cofnod's Administration System (CAS).

3.4.3. A custom built Data Validation tool (Dilys) is used to screen records as part of the process of importing them into the Cofnod database of biological records.

3.4.4. Records are passed through Dilys in batches, in Excel format, once relevant pre-processing has been carried out, and all records are labelled with the appropriate Dataset ID (officially Dataset Module Version ID) so that they are linked to relevant Metadata held on the Cofnod Administration System (CAS).

3.4.5. Dilys carries out checks to ensure each record has the following information:

3.4.5.1. A valid date.

- Checks are carried out for inaccurate date formats or impossible dates, e.g. 30th February, or a date in the future.

3.4.5.2. A valid species name.

- Taxon names are checked against standard UK Species Inventory (maintained by the Natural History Museum) names.
- Taxon names imported can be either scientific, English or Welsh (where these are stored in the NBN Species Dictionary), although not a mixture of these, and may be stored either as a preferred name or as a synonym within the UK Species Inventory.
- Data can also be imported using NBN Keys (Taxon Version or Taxon List Item keys) which link directly back to the preferred name or synonym of a species (or other taxon) in the NBN Species Dictionary.
- Whichever option is used for import, once the species name/NBN key has been checked and accepted the database will store the NBN Taxon Version Key as part of each record, which means that the taxon names used in Cofnod's database are kept in sync with the UK Species Inventory.

3.4.5.3. A valid grid reference.

- Grid references are checked to see that they are in a valid format.
- Records with lower than a specified resolution of Grid reference, by default less than 4 figures, are highlighted.
- Grid references are also checked against various geographic boundary datasets – by default all Datasets are checked to see that records do not fall in the sea, and that they fall within the Cofnod area, but these settings can be amended for specific Datasets e.g. the boundary of a particular Vice-county might be used for screening.

3.4.6. In addition Dilys also checks for certain key words in the Comments field, sometimes in conjunction with the species group, and alerts the user to potentially problematic records.

3.4.7. Records which fail any of the above automatic checks are displayed for checking by the person importing the data, and each record must be accepted individually for import (depending on the type of rule failure, this may or may not be possible without editing the record, either within the Dilys program or within Excel followed by a re-loading of the Dataset into Dilys). Records may also be rejected at this stage e.g. because of a data quality issue or because they fall outside the Cofnod area, or they may be accepted for import but marked as needing to be checked by the Data Supplier/Verifier.

3.4.8. Copies of the records accepted and/or rejected during each import of a Dataset Module Version are downloaded and saved in the Dataset Archive.

3.4.9. Rejected records can be corrected by Cofnod staff and then resubmitted through Dilys. However unless there is an obvious solution to the error in the data, all rejected records will be sent to Data Suppliers for checking. Once the Data Supplier has checked and changed the records they may be imported through Dilys as additional records belonging to the same Dataset.

3.5. ***Data Entry and the Cofnod Online Recording System (ORS)***

3.5.1. Data held in paper formats, or in non-suitable electronic formats, may require entry of individual biological records using Cofnod's Online Recording System (ORS). This is linked back to the Cofnod Administration System (CAS) so that each record is tagged with the appropriate Dataset ID (officially Dataset Module Version ID).

3.5.2. Internal ORS Data Entry is carried out by trained staff and volunteers. An internal Data Entry Guidance document is referred to, especially while new staff and volunteers are being trained, to encourage consistency. In addition, Quality Assurance checks are carried out regularly, whereby a member of staff/volunteer cross-checks a random sample of data (entered by another member of staff or volunteer) against the original paper files. This process identifies any inconsistencies in data entry methods as well as any potential data entry errors.

3.5.3. Records can also be entered into Cofnod's Online Recording System (ORS) by members of the public who register to use the system. At midnight of the day a record is entered it is automatically assigned to the appropriate Verifier (see 3.6 below).

3.5.4. Within the ORS, standard lists are used to populate fields where possible, to ensure consistent terminology. Various validation checks are automatically employed on the data entered. More details are given below:

3.5.4.1. **Species Name**: This is selected from a list populated from the UK Species Inventory (maintained by the Natural History Museum).

3.5.4.2. **Grid Reference**: This can be entered or checked using the integrated online mapping facility[2], which displays Ordnance Survey maps and aerial photographs. An appropriate grid reference qualifier may also be added.

3.5.4.3. **Dates**: It is not possible to enter invalid dates or dates in the future. Date ranges and approximate dates can be entered.

3.5.4.4. **Abundance and Record Type**: This information (where it exists) is entered from drop-down lists

3.4.7 Records can also be entered onto the Cofnod ORS by external parties who have been given access to customised data entry forms relating to specific Datasets on the Cofnod database. Using customised data entry forms provides many advantages in terms of ensuring high data quality:

- Records are labelled with the relevant Dataset ID which links back to appropriate Metadata.
- Additional drop-down lists can be used to enter data (to populate the sites name and/or recorder fields), grid references can be selected from a standard list linked to site names, and fields such as Abundance or Record Type can have customised drop-down lists.
- Additional custom fields can also be added, using drop-down lists where appropriate.
- Various fields can be set as mandatory or not depending on the requirements for the specific project/group.
- Viewing rights may be given to a number of different individuals involved with a particular project/members of a particular group, who are then able to see each other's records, reducing the likelihood of duplicate records being entered
- One individual can be assigned the task of checking all Data entered as part of a specific project

[2] Currently (February 2017) provided by Bing mapping

3.6. ***Data Verification and use of national rule sets***

3.6.1. A system of verification has been developed which was originally intended for application to ad hoc species records submitted directly through the Cofnod Online Recording System (ORS). However over time records within all datasets of category 4 or 5 will be assessed using this process. More details on this system are given within the document: **Cofnod ORS and Verification**, available from Cofnod's website http://www.cofnod.org.uk/Library or on request, which also includes detail of how nationally produced verification rule sets are being used to improve data quality within the Cofnod database.

3.6.2. Assessment of records using this process is carried out by relevant experts – these are usually County Recorders and are referred to below as 'Verifiers'.

3.6.3. Nationally produced verification rule sets are used to screen each record as it is entered/imported through Dilys and assign it a High/Medium/Low priority for verification. If a record is being entered directly onto the ORS and it fails one of these automated checks then the person entering the record will be alerted to check that they have entered the record correctly. Any records which are low priority for checking are set immediately to **Unassessed** (6), although their verification levels may still be altered by the Verifier if they wish to do so.

3.6.4. At midnight of the day a record is entered it is automatically assigned to the relevant Verifier, after which point the Recorder is no longer able to edit their record but can contact Cofnod if they decide they need to make any changes. The status of the record will either show as **Assigned** (1) or **Unassessed** (6).

3.6.5. Once a record has been assigned to a Verifier, he/she is able to view it and change its verification level, and also attach verification notes to it if desired. The Verifier can decide to assign one of five levels to the record (details below).

- **Known Incorrect** (2) - The expert knows that this record is wrong, usually based on correspondence between them and the recorder.
- **Probably Incorrect** (3) - The expert has decided that the record is likely to be incorrect, based on their knowledge of a number of factors, including the species likely locations, date of sighting, difficulty of identification etc.
- **Unconfirmed** (8) - The expert has decided that they do not have enough information to make a decision on the status of the record, perhaps because the date is too far in the past.
- **Probably Correct** (4) - The expert has no reason to doubt the validity of the record, but cannot be certain without proof.
- **Known Correct** (5) - The record is definitely correct. This is usually only selected if a photograph has been provided, or the expert has seen a voucher specimen.

3.6.6. The **Unassessed** (6) status is also applied to any Medium priority records which have been assigned to a Verifier for 30 days but which have not been assessed during that time. Or in cases where there is no expert available to assess a record then its status may be manually set to Unassessed by Cofnod staff. This category is also used for all records within datasets of category 4 or 5 imported before 16/12/2012, unless their verification levels have since been changed by a Verifier (these records will gradually be checked over time as and when Verifiers have available time). No Verifier Name is attached to such records.

3.6.7. Verifiers can add notes to individual records elaborating on their decision or pointing out future action. They can also contact the Recorder to request further details if applicable. Not only is contact information for the Recorder available through the ORS but also information that he/she has entered about their **recording interests** and **experience.** Recorders also apply **Confidence Levels** to each record that they enter which may help Verifiers in deciding on which Verification Level is appropriate. In some cases the Recorder may have attached supporting evidence to their record, usually a photograph.

3.6.8. Verifiers are also able to view the location of individual records using the integrated online mapping facility.

3.6.9. The Verification Level of a record can be changed at any time, for example if further information becomes available.

3.6.10. Records from Datasets of Categories 1-3 are given a Verification Level of **Not Applicable** (9) (unless they relate to species which aren't covered by the Data Supplier's area of expertise e.g. a butterfly record within a County Recorder's dataset of moth records.

3.7. Data use

3.7.1. In producing most general reports, all relevant records from datasets falling within categories 1-3 will be used. However only those records with Verification Levels of Known Correct, Probably Correct, Unconfirmed or Unassessed will be used from datasets in categories 4 and 5, with a clear explanation of their Verification Level.

3.7.2. In any general reports which Cofnod produces, each record is given a summarised Verification Level as follows:

- **Unassessed (1):** Any records from Category 4 Datasets which have not been assessed by an expert.
- **Unconfirmed (2):** Any records where a Verifier has applied a Verification Level of Unconfirmed.
- **Considered Correct By Cofnod (3):** Any records from Datasets of Categories 1, 2 or 3, except in occasional circumstance where the Data Supplier has chosen to apply our Verification Levels within his/her dataset because they recognise it is not all of a consistent quality.
- **Considered Correct By Expert (4):** Any records where a Verifier has applied a Verification Level of Probably or Known Correct.

3.7.3. Records flagged as Assigned, Probably Incorrect or Known Incorrect (detailed Verification Levels 1, 2 or 3) will not be used in standard searches, but may be made available in certain circumstances (e.g. to the relevant County Recorder), with a clear explanation of their Verification Level.

3.7.4. A record may also be temporarily flagged to be ignored during the reporting process, for example because there is a possible data quality issue or it is a known duplicate.

3.7.5. Each record supplied will include at least the following fields:

- Name of original recorder
- Species name
- Date (the format may vary)
- Grid reference (the detail may vary)
- Dataset ID
- Verification Level (if applicable, see section 3.6)

3.7.6. The Dataset ID can be used to check Metadata for individual records e.g. Dataset Category (see section 3.3). See section 3.2 for more details on the Metadata held by Cofnod.