BSBI Scottish Recording Conference

Validation and verification of records on the DDb

14th April 2018 - RBGE (Jim McIntosh)
28th April 2018 - Strathpeffer Community Centre (Andy Amphlett)

What is a plant record? – WWWWV

What taxon (species, subsp., hybrid, var. etc)

Where was it recorded (site name and grid reference)

When was it recorded (date)

Who was it recorded by (and determined by if different to recorder)

and

Validated (and **V**erified) – confirmation that the record is considered correct (and who confirmed it)

Validation - checking that dates are valid / correct; sites, grid references and vice-counties are consistent / correct; and recorders and determiners are correct

Verification - that the taxon ID is considered to be accurate

Why do we need to validate / verify records?

- **Most datasets contain errors**: (*72 – 178 separate datasets per vc in Scotland on DDb*)
    - What % of records are incorrect in some way?
        (0.03%) – 0.5% - (10%) - >>>
    - Are these errors significant?
        See DDb > my county > data validation

- Are previous validations on DDb trustworthy / can they be accepted?
    Varies between vice-counties and datasets

- **It is impossible to eliminate all errors, in all datasets, so that is not the target of validation**

- Priority is to ensure that records are **good enough** for a purpose:
    - Good enough for Atlas 2020 (records correct at taxon x hectad & tetrad x date class precision). This is the priority.
    - Good enough for the VCR's needs, eg for a Checklist, RPR, or a Flora (needs even more thorough checking).

Verification (examples of taxon errors)

Misidentification

Correct ID but recorded wrong taxon, eg. crossed off wrong name on card or data entry error

Correct ID and record correctly input, but later taxonomic changes corrupt records

Improved knowledge of ID characters of a species may require revision or rejection of earlier records

Importing records from spreadsheet sidesteps many validation / verification checks

Importing of records from non-MapMate databases and spreadsheets may not handle agg. / s.l. taxa correctly

'Other species' section of recording card used to list species to look out for or not found (incorrectly assumed found)

Etc …

Validation (examples of sources of date, site, grid reference , vice-county and recorder name errors)
Grid ref estimated by eye from map
GPS grid ref inaccuracy
Spurious grid ref precision, eg. precise grid ref applied to pre-1945 record in GB; monad used for tetrad; monad used for quadrant
Grid ref = centroid (especially where 'site' extends to more than one hectad or tetrad), and centroid not specified
Grid ref = where parked car (but recorded over wider area)
100km grid ref error
Misreading 3 and 8 in grid ref on GPS screen
Transcription errors (including difficult to read handwriting)
Site grid ref not compatible with grid ref in Comments
Site name includes grid ref, but not compatible with site grid ref
Inaccurate vc boundary used
MapMate assigned wrong vc
Date of publication used as date of record
Incorrect or misspelt site name
Data entry by non-botanist / unfamiliar with local area
Importing records from spreadsheet, sidesteps many validation / verification checks
Excel date format errors
Compiler listed as Recorder
Etc, etc, etc, etc ....

NOW

However large the vc dataset, and however large the number of unchecked or incorrect records, it is a **finite** quantity.

Do not let the backlog of unchecked records get any bigger.

Prioritise record checking

Check, (correct where necessary), and **validate** records as soon as they arrive at the DDb

Check and validate **all** records

NOW

However large the vc dataset, and however large the number of unchecked or incorrect records, it is a **finite** quantity.

Do not let the backlog of unchecked records get any bigger.

Check, **(correct where necessary)**, **and validate** records as soon as they arrive at the DDb

Prioritise record checking

Check and validate all records

DEADLINE for recording, submission and validation of records for Atlas 2020 is
31/12/2019

Is it necessary to examine every record in detail when validating records?

NO!

When you Confirm a record does that mean that you have to be 100% certain the record is correct?

NO!

Confirmation of a record means that you are happy to accept it as likely to be correct.

If I Confirm or Reject a record can I change my mind later?

YES!

You can always assign a 'new' validation state to a record, which will over-write the previous version.

Ockam's razor or the "**law of parsimony**" - William of Ockham (vc17) (c. 1287–1347)

Historic plant records & records derived from herbarium specimens:

- **simpler interpretations are usually better** (one hectad not two)

Extraordinary claims require extraordinary evidence

Carl Sagan (1980)

The weight of evidence for an extraordinary claim must be proportioned to its strangeness

Pierre-Simon Laplace (1749–1827)

- There are an infinite number of potential plant records that were not, or will not be, collected.

- So don't worry too much about rejecting the relatively few records over which you have serious doubts, or which contain contradictory and un-resolvable details.

- Always explain your decisions to reject or doubt records (add a comment to your validations)

From now on: (MapMate example)

• Ensure you make use of MapMate's inbuilt **validation tools**.

• **Submit sync files after every data entry session** (currently averages 10mins from sending sync to records appearing on DDb; occasionally longer).

• Don't hoard records on your copy of MM; sync regularly.

• **Check records as soon as they arrive at the DDb, or as soon as is practical**.

• **Adopt a minimum sufficient workflow practice to detect any unusual or significant records**. No need to use all the validation tools on the DDb (use the most appropriate), but be aware of the range of available tools.

• **Once the more unusual or notable records have been checked, confirm ALL records in the set**, eg.

The backlog:

- **Prioritise checking**, using DDb > my county > data validation tab, as **starting point**.

- **SINGLETONS are easiest type of record to tackle**. Start with hectad singletons.

- **For non-singletons need to look at all records for taxon x hectad, including confirmed records to make an informed judgement.**

- **Don't get stuck on difficult, often old, records.** Tackle more straightforward records first.

- Hieracium and Taraxacum records are lower priority, tackle at end if you have time. (Key datasets for those genera have been verified but not validated)

- (Pete Stroh has been checking records for NR and NSc species on the DDb, but only from hectads where there are no dots in the New Atlas). Still LOTS to do!

Records to target for validation for Atlas2020

Suggestions for starting points for detailed checking of records from your county.

Singletons and anomalies

One-off or isolated occurrences records should have their status positively confirmed or rejected.

	current validation state		
	unchecked	confirmed	needs verification
county singletons	82	55	4
100km radius singletons	13	14	0
hectad singletons	1604	2265	33
first county record	368	824	11
records with no Atlas hectad status assigned	3315	1076	23

☑Exclude neophytes and unchecked taxa.

Rarities

Records of taxa that appear on lists of rarities (either regionally or nationally) may be worth closer scrutiny for Atlas2020.

	current validation state		
	unchecked	confirmed	needs verification
GB National Status	135	70	11
UK Bap Species	40	27	4
GB Red list	63	56	7
Scottish Biodiversity List	80	92	10

☑Exclude occurrences that already have hectad-level status assigned (e.g. fitting the distribution of Atlas 2000 or later revision).

Other possible oddities

Please check occurrence records that the database system has highlighted (only a small proportion of these are likely to be erroneous).

	current validation state		
	unchecked	confirmed	needs verification
records near the maximum of their (inferred) altitudinal range	6224	4518	2
records near the minimum of their (inferred) altitudinal range	29	142	0

618 records used site names that may be inconsistently applied to grid-references (i.e. the site name is associated with two or more unrelated locations).

Based on known or inferred dates of death of recorders or determiners, 6 records may have incorrect dates (in the wrong BSBI date class).

Records to target for validation for Atlas2020

Suggestions for starting points for detailed checking of records from your county.

Singletons and anomalies

One-off or isolated occurrences records should have their status positively confirmed or rejected.

	current validation state		
	unchecked	confirmed	needs verification
county singletons	148	151	4
100km radius singletons	27	22	0
hectad singletons	2423	3008	46
first county record	535	1257	12
records with no Atlas hectad status assigned	5050	1583	23

☐Exclude neophytes and unchecked taxa.

Rarities

Records of taxa that appear on lists of rarities (either regionally or nationally) may be worth closer scrutiny for Atlas2020.

	current validation state		
	unchecked	confirmed	needs verification
GB National Status	8398	7193	18
UK Bap Species	2368	1838	11
GB Red list	4408	3051	8
Scottish Biodiversity List	5348	3982	20

☐Exclude occurrences that already have hectad-level status assigned (e.g. fitting the distribution of Atlas 2000 or later revision).

Other possible oddities

Please check occurrence records that the database system has highlighted (only a small proportion of these are likely to be erroneous).

	current validation state		
	unchecked	confirmed	needs verification
records near the maximum of their (inferred) altitudinal range	6224	4518	2
records near the minimum of their (inferred) altitudinal range	29	142	0

618 records used site names that may be inconsistently applied to grid-references (i.e. the site name is associated with two or more unrelated locations).

Based on known or inferred dates of death of recorders or determiners, 6 records may have incorrect dates (in the wrong BSBI date class).

Validation state options

'Validation' state	Mapped?	Returned by default queries?
Unchecked (i.e. newly added to the DDb)	✓	✓
Accepted / Confirmed	✓	✓
Needs verification / May need verification	X	✓
Doubtful	X	X
Rejected	X	X
Duplicate	X	X
Deleted externally (in the origin database)	X	X

- There is no functional difference between **Doubtful** and **Rejected**. Rejected means that you are virtually certain the record is incorrect in some way and can not be salvaged; Doubtful means that it is highly likely to be incorrect.
- **Needs verification** should be used when you plan to, or want someone else, to follow the record up, and there is sufficient doubt about the record that you don't want it mapped (yet).
- Detecting **Duplicate** records is **not** part of validating records. **There is always something better to do with your time than searching for duplicates!** But avoid creating new duplicate records as far as is practical.

Confirmation options

Reason	DDb explanation	Suggested use
Positive identification	Record was seen and checked by someone with the necessary expertise (either in vivo or as a voucher specimen)	**Use for VCR's own records, those confirmed by a taxon referee and those by other trusted recorders.**
Considered correct	The verifier has not seen photo/s or specimen/s but has a high degree of confidence that the record is likely to be correct, based on difficulty of ID, date, location and recorder skills/experience etc.	**Almost as trusted as 'Positive identification'. Use for most records by other reliable recorders.**
Conforms with literature report	Record is from or related to a credible literature report.	**Where record has been subsequently confirmed by a taxon referee, use 'Positive identification'.**
Readily identifiable & uncontroversial	Common or otherwise uncontroversial occurrence that can be presumed to be correct	**Use for relatively unremarkable occurrences from recorders who are unknown to the VCR, ie. common taxa from locations where already recorded, or to be expected. Only use if records cannot be assigned to one of the first two categories.**
Consistent with accepted distribution	Occurrence matches the distribution reported in a national or local atlas or is from a known site for the taxon.	**Use for records of less common taxa from locations where already recorded. Only use if records cannot be assigned to one of the first two categories.**

Confirmation options – the new 'do not map' validation option

The <u>default option</u> is 'include on maps'

A validation-related option for use with records of plants that are probably **cultivated in a non-wild situation**, eg. within a garden, and that (though correctly identified and located etc.) should not be included on maps intended to show the **wild distribution**, eg.

	record	taxon	recorder	vc	locality	grid ref	date	status
☐	📄✏️	_Wahlenbergia hederacea_ *	McCallum Webster, Mary	VC95	Rose Cottage, Dyke	NH9858	1975	⊙ planted Sco GB:NT

1 record matched your query.

(The most northerly record of this species in GB. It was planted by the Recorder in the lawn of her cottage, and is included in her published Flora).

Validation state change

Please confirm records that you are happy to accept as likely to be correct. Prioritize validation of records for rare taxa and other more interesting or unusual occurrences. **There is no need to validate all data, most noncontroversial records can be left as unchecked.**

Use the 'reasons' below to distinguish between data that you have high confidence in versus records that are assumed to be valid.

Set 39 records as 'confirmed'?

reason: ▼
Atlas inclusion: include on maps ▼
comment:
- include on maps
- do not map
- leave unchanged

Please state th... the validation status.

Mark as confirmed   Cancel

The 'do not map' option <u>must **not**</u> be applied to all planted occurrences or to all aliens in wild situations. Many planted occurrences and most alien escapes are absolutely valid for atlas inclusion, and should be mapped.

- **This option _cannot_ be used as a way to hide sensitive or private records.**
- **BSBI _will_ share 'do not map' occurrences in the same way as other data.**

How to help VCRs when supplying records

- Check with the VCR what format they would like to receive records in.

- Ideally use either:
  - A proper database, ie. MapMate (and supply records via sync file), or
  - A pre-formatted spreadsheet (eg. as available from BSBI website)

- Do not record a taxon unless confident of the ID. (If in doubt take a photo).

- Double check grid references, vice-counties and locality spellings.

- Do not be put-out when the VCR queries your records (it is their job to do this)!

<u>TIPS for VCRs</u>

- Use the DDb regularly

- Look at the DDb guidance document for ideas

- Expand the range of queries you are comfortable using
    - Want some guidance / advice? Ask.
- Be inquisitive! Explore the records for your vc
    - Want some suggestions? Ask.
- Enjoy it! - validation is satisfying and useful

- Stuck? Ask.

- For advice please post a question on the DDb message board, ask Jim (or your Country Officer) or Andy.