GRU Workshop Conversation on Data Quality Considerations and Checks.

Data quality (dq) considerations and checks – an annotated list.

Workshop participants discussed data issues they look for and then generated a set of data quality checks to be considered when evaluating, cleaning, and improving data fitness for use. These were divided into three categories: time, geography, and taxonomy. Keep in mind this dq discussion focuses specifically on issues to look for in biocollections data. Addressing dq issues takes time and needs vary by research question. Each researcher will have to decide how much record cleaning (vs. record deletion) to do to best suit time constraints and the scientific questions. When evaluating and cleaning data – it is important to 1) save an untouched copy of the raw data – and 2) write down all steps taken when cleaning and standardizing the dataset (White 2013).

> (**annotation format:** issue is listed, followed by a brief explanation to clarify some the dq observations, suggested tests and salient issues. The prefix "dwc" indicates a reference to a term in the Darwin Core Standard (dwc)).

Time:

- **Problems with 9999 dates** (or other **placeholder values researchers use to represent no data**). In standardized data to be published and shared, it is best practice to leave a field blank when no date (or other information) is available, rather than a placeholder. These can be added in your data manipulation (if needed by the software). Putting them in public data to be shared / reused – can result in erroneous conclusions if these placeholder values accidentally get included in data analysis. See Nine simple ways to make it easier to (re)use your data for more on this issue. Look for (and remove) these placeholder issues in existing data to be (re)used.
    - Referencing first day of a month. Other common issues include referencing a point to the first day of the month, such as "01/01/XXXX" when the **record month is unknown,** resulting in spurious temporal precision. In other words, avoid using "01" as a placeholder when the day and or month are not known.
- **Order matters - ambiguous day/month values in date field**. In a spreadsheet program or with a tool like OpenRefine, sort the date column to look for this format issue. Use dwc:verbatimEventDate to store dates as written on labels or in notebooks. But avoid writing verbatim dates for future collecting events with these formats, i.e. dd-mm-yyyy or mm-dd-yyyy, because it can result in ambiguous day/month records. A researcher would have to delete any records like this when downloading datasets from other sources – if dates matter for the research or conservation question. Better practice to use dwc:eventDate with format yyyy-mm-dd (or other suggested formats in dwc:eventDate).
- **Single date** or **date range?** Some software does not provide the ability to capture an entire date range, creating situations where date ranges are ignored in the data. Check for this by looking at the **verbatim date** (dwc:verbatimEventDate) compared to the dwc:eventDate to determine if a date range exists, but is being ignored.

- **Extensive data ranges are not useful** (*e.g.*, 1600-9999 instead of 'date unknown'). Depending on the research or data visualization question, you may need to toss these out.
- **Years need 4 digits. Two-digit years are ambiguous.** Run tests like **sort**, on year values to look for outliers (like < > four digits, or unexpected year values outside the expected range for the data). Best practice for future data creation is to use four-digit years like 1800, 1900 or 2000. Records with two-digit years would need to be left out of any analysis or mapping requiring time values. (See dwc:year).
- **Months can only be 1 – 12.** Test month data for values > 12 or < 1. (See dwc:month)
- **Days can only be 1 – 31.** Again, run tests by month to look for outliers as months vary in number of days. (See dwc:day). There should be no values < 1 or > 31 in the day data.
- **Remember Leap Year changes days in month and days in year.** It is programmatically possible to identify leap years and adjust the expected number of days in February and days in the year, accordingly.
- **Collecting before birth, as an infant, or after death – use collector date ranges**. Use known ranges for when a collector collected (and where) to find date ranges that indicate potential biased activity or false collection date.
- **First day of month/last day of month**. Look for patterns that may be a forced correction if only the month and year are known. Use verbatim date and date ranges to confirm the date. Example: you might see a lot of specimen records having the first day of year, or first day of month as collecting date. Some systems do not allow blanks or non-standard text for dates with missing day and so many specimens have thus been recorded as have being collected on January 1st (or Feb 1st, or Mar 1st, etc.) when only the year (or year and month) of collection is known. Or, some software that evaluates date data (like Java) don't expect 00 for an unknown month or 00 for an unknown day because there is no 00 months or 00 days. Java will auto-increment those *place-holder* values to 1. (Don't use 00 as placeholder, leave blank).
- **Date parsers** such as those used by data aggregators may be a programmatic solution, but are not fool-proof. See above example for **first day of month**.
- The generation of a **histogram-by-year** will reveal the distribution of records over time and clustering record dates by collector so that links can be made between specific dates and locations.
- **Identify outliers** in a data set, such as occurrences that represent species presence at a location during the wrong time of the year.

Geography:

- **Higher geography** is often not provided, but can be provided by aggregators based on ontology/vocabulary/gazetteer. (See dwc:higherGeography).
- **Georeference is out-of-expected bounds.** Check the locality against the associated administrative or geographic boundary such as county boundaries or coast lines. Look for instances where topology of boundaries are not obeyed (political, bodies of water, geographical feature). Software like QGIS, or any simple mapping tool can be used to evaluate if points are in or out of expected higher geography boundaries. Data aggregators check for this to some degree – but this is not perfect. Knowing the domain of observations, such as a bounding box or the boundary of the state where the observation was made, can filter out outliers.

- **Borders change, water levels too.** Changes in boundaries through time (*e.g.,* changes in administrative boundaries, water level) mean a point may or may not be correct. Time plays a role here, as well as geopolitics.
- **Centroids.** Note that georeferences provided in existing datasets are sometimes to centroids of countries, states, counties, cities, parks, etc. This may be sufficient for some research purposes. If the research question requires a more specific georeference, then locality records georeferenced to centroid will need to be deleted or the georeferences refined if possible.
- **Standardized language for names of places** helps cluster these places for georeferencing. Be sure to look for use of colloquial local names and to standardize when trying to cluster by place. An example would be UCSB versus University of California, Santa Barbara. Generally, abbreviations should not be used.
- **It is not where you think it is.** With old data, or because of the practices of indigenous cultures, a place may not be where you expect. For example, in Papua New Guinea, when the villagers move, they call their new village the same name as their last location. So, if you have coordinates for years ago, for this place, compared to newer data for this same "place" they will not line up.
- **Check the coordinates.** Specific issues with coordinates can include incorrect signs, or even transposed values, both resulting in an erroneous georeference.
    - **Check the sign (+/-).** Visualize the data points to look for outliers. A common error is a **missing or additional negative (-) sign**, inverting E/W, N/S (often resolved by data aggregators).
    - **Latitude and Longitude values swapped.** Sometimes the latitude and longitude values have been confused for each other somewhere in the data sharing process. Visually, this is easy to spot and simple to fix. Also, these points will be out of the expected bounds.
- **Georeference, but no locality data.** Look for and evaluate instances of "Unknown" locality or no verbatim locality text information with a data point. It doesn't mean the data point is wrong, but one needs to be suspicious of these data points. Where did the geopoint come from if there is no locality data? There may be other hints such as a specimen label image indicating the record is not yet completely transcribed. The label may be the source of the locality data that made the georeference possible. Or you may need to toss out the record.
- Look at **dwc:locationRemarks**. Ambiguity in the location is often noted in the remarks (dwc:locationRemarks). These records can then be corrected or deleted from the dataset.
- **Look for clones –** same collector in two places at once**.** One can use the **time data** and **collector data** to find instances of a collector being in two places at once, or in two locations that are very far away from each other -- on consecutive days.
- Use **data layers for quality control** (*e.g.*, habitat, distribution models, boundaries, living collections, geologic maps). These can help to spot taxa in highly unlikely places, thereby indicating suspect records that require closer inspection and possible deletion from viable dataset.
- **Place of origin issues** (dwc:establishmentMeans): zoos, botanic gardens, ports of entry. Overlay boundaries may assist with recognizing such data points.
- Look for **too many decimal places** aka - **fake precision**."

    - i.e. a person clicks on a vague spot on a google map then copy paste coordinates to **umpteen decimal places**. It is important to look for information on how the georeference was done (see: dwc:georeferenceSources, dwc:georeferenceProtocol, dwc:georeferenceRemarks) and note the recorded GPS accuracy (dwc:coordinatePrecision). While these are best practice, existing data may or may not have this information. Personally I (participant) think these are more important than recoding the dwc:datum used (or assumed) which could only be out by a few hundred metres. Best practice is to include these data for any future collections or observations.
    - **too many 00000s.** Look for spurious precision from too many zeroes to the right of the decimal point. Sometimes databases do not allow NULL values in these fields. The result can be the addition of unnecessary 0's to fill decimal places or inclusion of placeholder coordinates (e.g., 0,0; 0.0000,0.0000). Rounding the digits during the cleaning process can get rid of this issue.
- **Garden or cultivated?** Be on the lookout for botanic garden coordinates (cultivated and/or reared specimens) or coordinates that point to the physical collection itself, rather than to the site of its original observation/collection. This can be confirmed through the inclusion of additional geospatial data layers that would support the discovery of these anomalous trends.
- **Georeferencing Best Practices.** See: Guide to Best Practices for Georeferencing. (Chapman 2006).

Taxonomy:

Data aggregators check taxon names against particular sources. They do this to facilitate searching and finding when looking through millions of records. For example, if the data provider shared Family and lower ranked names, the data aggregator would add the higher classification (names above Family) so that when searching on Phylum, Order, or Class, the researcher will find these records. Variations, typos, and homonyms (same name but in different organismal groups) sometimes result in aggregator algorithms putting taxa in unexpected company. For all these (and more) reasons, the taxonomic identification data needs to be evaluated before using it to draw research conclusions or make management or policy decisions.

- **Ensure internal consistency** for analyses. The dq checks listed here offer steps needed to evaluate and validate taxonomic data before using them to draw scientific conclusions. All of these dq checks require decisions be made and/or assumptions tested and that the same protocol is then followed faithfully for the entire dataset.
- **Taxa out-of-expected spatial bounds** - in an unexpected place. If a taxon has a well-known range, or expected habitat, then appropriate data layers (e.g., habitat, range maps) can be used to look for records indicating taxa in unexpected places. These will need closer inspection to decide if the data are erroneous or plausible.
- **Taxa out-of-time bounds -** in an unexpected time. Using time as a scale, taxonomic information and taxa trait data (e.g., emergence, flowering, fruiting) one can look for taxa that are present in an unexpected season. These will need closer inspection to decide if the data are erroneous or plausible.

- **Level of taxonomic identification varies.** Here are some places to look and things to look for when evaluating the level of identification.
    - **Infraspecific designations are often not provided**, and so level of taxonomic precision needed must be decided that suits the research question or data model. And then records kept or deleted as fits research needs.
    - **Undetermined, Indet or empty fields.** Look for records that have *indet* or *undetermined* in the dwc:scientificName field. Perhaps the dwc:family is provided, or even dwc:genus but if the lower rank name like species is necessary for your research, these records need to be deleted. See also next dq check of taxon rank.
    - **Taxon rank.** This is the rank of the lowest taxonomic identification provided in dwc:scientificName. You can use dwc:taxonRank (if provided) to sort the taxonomic data by rank (Genus, Species, Subgenus, etc.). If your research requires specimens to be identified at least to a certain rank, you can use this field to find and then remove records that do not meet your needs.
    - **Note dwc:scientificName field includes the lowest determination for that specimen**. Someone might misunderstand the dwc:scientificName field thinking it refers to the *species* name, but this is not true. It is the name string for the lowest level determination. For example, you could find a genus name only in dwc:scientificName, or you might find the subspecies determination (Genus+species specific ephithet+infraspecific epithet) and the author name may be included in this field as well.
- **Automated taxonomic updates** using authoritative sources. Taxonomic backbone data normalization provided by aggregators for some taxonomic groups may not be the most accurate for some research purposes. Most aggregators today provide the raw (original) and enhanced data in downloads. Raw taxonomic data may be more useful – for example, see dwc:genus, dwc:specificEpithet, or dwc:scientificName (which sometimes includes author) and compare these values in the raw and enhanced downloaded datasets. See explanation at the beginning of this taxonomy section.
- **Gender agreement**: standardize the taxon name for the purpose of analyses. Sort to see taxonomic spelling variations. For example, in raw data you may find different spellings for the same taxa (e.g. what is really the same specific epithet ending in both –us and –is). Standardize these using caution. Check to make sure they really are the same taxa. If you are using raw data, you will need to check other issues too, like synonymy.
- **Synonymy** (name indices). Names may be synonymized in the enhanced data downloaded from aggregators, but will not necessarily be in the raw data. As a first hint, you might check, the distinct genus names in the raw data, looking for potential synonyms. Then, be sure to evaluate the dwc:scientificName data looking for further evidence of synonyms. You will need to decide which name/s to use – for consistent results in any research analysis and mapping.
- **Higher taxonomy often missing** or **erroneous.** Data providers sometimes share higher classification data (Kingdom, Phylum, Class, Order, etc.) and sometimes not. This information facilitates searching aggregated data. Sometimes, a data aggregator provides the higher classification from a supplied taxonomic backbone (an outside source). Automated routines run to look for taxonomic name matches (or near matches) to retrieve the higher classification data for a given taxon name and rank. This can result in errors.

Be aware of this when looking for outliers and instances of names that don't seem to make sense or be in the expected Family, Order, Class, etc. Again, the raw data helps because one sees what the data provider shared – not the automated higher classification.

- **Check endings of taxonomic names** in case of incorrect dwc field mapping. Nomenclature spelling conventions can be used to find errors where a name has ended up in the wrong field. For example, family names end in "eae" for plants and "idea" in animals. If you find these name endings in a field other than family – a mapping error is likely. And if you find other endings besides those, in the dwc:family field, you may have found another mapping error. These and other taxon names can be checked for in the dwc:genus, dwc:specificEpithet fields for incorrect field mapping, and can be checked for in dwc:scientificName for specimens not determined to a significant resolution.
- **Taxonomic Identification changes made over time.** The **determination history** for a given specimen (if present) can be useful (See dwc:identification). A specimen may be re-identified over time. If data providers share this information, it may help with deciding which names to use, and which records to keep or not. It may also be useful to check the dwc:identifiedBy field to find out which taxonomic expert made the determination.

References.

Chapman, A.D. and J. Wieczorek (eds). 2006. Guide to Best Practices for Georeferencing. Copenhagen: Global Biodiversity Information Facility. Available online at http://www.gbif.org/orc/?doc_id=1288

White, E., E. Baldridge, Z. Brym, K. Locey, D. McGlinn, S. Supp. 2013. Nine simple ways to make it easier to (re)use your data. http://dx.doi.org/10.4033/iee.2013.6b.6.f