Assumptions:

* An R-based tool, using RMarkdown, to analyse routine Water Quality data.

|  |  |  |  |
| --- | --- | --- | --- |
| Area | Essential | Desirable | Out of scope/To be confirmed |
| User interaction | Users should be able to apply settings for running the toolbox, such as specifying input file paths and sheets and start and end dates. |  |  |
| Import | A fixed-format Excel template will allow all the necessary data to be entered from:   * Automated probes * Manual field data * Lab data   The different sources will be assigned separate tabs in the template.  Data validation will be used to allow users to select the parameters for which data is available and ensure correct column names.  The toolbox will initially be limited to working with the following WFD parameters:   * Temp (°C) * DO (%) * pH * Biochemical Oxygen Demand, Total (mg/l) * Orthophosphate as P (mg/l) * Ammoniacal Nitrogen as N (mg/l) * Nitrate as N (mg/l)   A metadata tab will allow APEM data to be matched to EA data and water body type (eg, salmonid/cyprinid) to be identified.  Data validation/conditional formatting will be used to ensure site information matches between the different sheets.  The template will be locked down to prevent editing of column headers, parameter details etc.  EA data will be automatically imported using the R hetoolkit package, import\_wq function.  The system will perform automated checks and warn if, eg, the input data is not in the required format, or if any chemistry data appears to be missing.  The user can specify which determinands to process / analyse, and in which order. (Default = analyse all determinands in the input file). | Get location data (Easting/Northing) from EA WIMS database for running the tool using EA data only (hetoolkit::import\_wq does not import this).  Automatic conversion of chemistry units (users must ensure that concentrations are in correct units for each determinand or make necessary conversions manually).  Additional, non-WFD determinands (eg, metals). | TBC:   * Column names for template sheets. * Connection to Survey123. * Import of flow data and calculation of flow-weighted loads.   OOS:   * Import of continuous WQ data. * Nesting of locations within wider sites. |
| Merge | Each data tab will be individually processed so they have the same structure.  A “source” variable will be added to identify the source of each record (probe, field, lab, WIMS).  All available data sources will be merged to create a single dataset for further processing.  The merged dataset will have user-readable column names added and will be saved to a user-specified location. | Import template designed so that the data is structured correctly (eg, using hidden columns). |  |
| QC checks | Outliers beyond an agreed specification for each parameter will be tabulated.  Results outside an agreed expected range for each parameter will be tabulated.  A combined table summarising all out of specification (OOS) results will be exported to a location specified by the user so these results can be reviewed. | Outlier thresholds will be dynamically calculated or identified using statistical tests.  Out of specification export combined into main data export file.  Related parameters (eg, dissolved vs total, generally for metals) will be compared and unexpected results will be tabulated. | TBC:   * Thresholds * Current QC checks. * Any in-built QC checks on probe data. * Concentration vs saturation comparison? |
| Processing | User settings will be checked to confirm that the entries are valid.  Users will be able to apply a chosen multiplier (between 0 and 1) to LOD values.  Results outwith the LOD will be summarised and exported.  Summary statistics (mean, SD, min and max) will be calculated for each site/parameter combination and tabulated.  Percentile values will be calculated and tabulated as required for BOD and Ammonia.  The results will be compared to WFD standards. | Other statistical summaries as requested.  WFD standards can be calculated as required.  Linear regression, break point and/or Mann Kendall tests can be used to test for statistically significant trends. | TBC:   * Identifying LOD results in the data. May need to be standardised for manual entry.   OOS:   * Other methods of handling >LOD results (eg, include option of Kaplan-Meier estimator to adjust LOD values. |
| Reporting | The complete final dataset will be exported in excel (.xlsx) format to a location specified by the user with a date-stamped file name and agreed, user-readable column headers.  A summary report will be generated in .html format which will include:   * User specified settings. * Interactive versions of each chart. * Interactive table(s) showing OOS results. * Interactive table(s) showing the summary statistics. * An interactive map showing the location of each sampling site. * The functionality to manually download the tables and charts.   The charts will show WFD standards. | User controls will be available to control the formatting of the charts.  Time series charts for each site/parameter combination can be exported, in .png format, to a location specified by the user.  In the final export(s), results outwith the LOD will be highlighted using conditional formatting.  The user will be able to determine whether the charts are automatically saved to the output location. | TBC:   * Chart formats (examples, P11902 faceting, P11002). |
|  |  |  |  |

Error checks:

* Import file does not exist at specified path.
* Sheet names entered do not exist in the specified import file.
* Start/end dates in incorrect formats.
* End date before start date.
* Start date before 2000 (warning)
* Unknown parameter(s) specified.
* LOD multiplier is not numeric.
* LOD multiplier is outwith the range (0,1).
* Mandatory data fields are not filled in (date, time etc).
* Others as identified.

Current snags:

* Getting headers to EA to match APEM data (esp if not all determinands present. Parameters table?)
* If using additional determinands, identifying determinand ID code in WIMS - future work.
* Time formats
* Lat/long in lab data - add new columns/match from other tabs/leave blank.
* Location ID variable duplicating on merge - not in probe data tab. ~~- might not be needed at all~~.
  + ~~Needed to match EA WIMS IDs to APEM sites~~ - No, use EA\_ID, drop location\_ID
  + Need to add metadata to EA data to get location\_names
  + A match somewhere duplicates EA\_ID and location\_ID
* Confirm colname for DO%/RDO saturation (is RDO Sat in probe the same as DO?).
* Calculating statistics:
  + All data, including processed LOD/outliers.
  + Outliers removed.
* Outlier identification.
  + Boxplot.stats()$out being used for now. Other options may be available.