Origins

SOFTWARE GUIDE

Operations manual

Version: 2.0.7

Last updated: April. 2022

STEP ONE: ISSUE OF LICENCE

Origins is a software application that can be accessed using a variety of access channels: online, in real-time, via an API or using desktop software.

This guide explains the operation of the desktop software.

This software is licenced on an annual basis. Each licencee is recognised by a unique user name and password. The user name and password are associated with the parameters applying to the use of the software by the software. These, among other things, include the country in which the software is to be used, the date the licence terminates and the identity of the distributor.

STEP TWO: OPERATION OF THE SOFTWARE
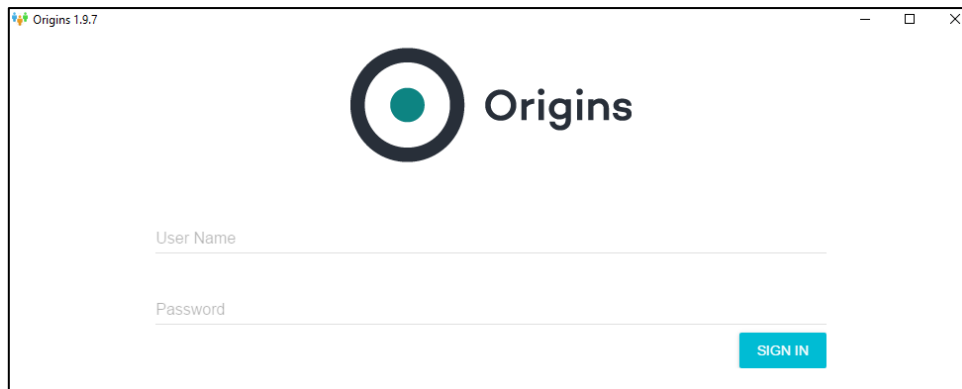
The current version number appears at the top left of the screen. The banner indicates the identity of the distributor.

The user enters a User Name and Password and signs in. Please note that both username and password are case sensitive. Please refer to your fulfilment email to ensure that the username is entered correctly.

From time to time, the application may indicate that a new version of the software is available. The new version downloads itself automatically and sufficiently rapidly not to cause the user material delay.
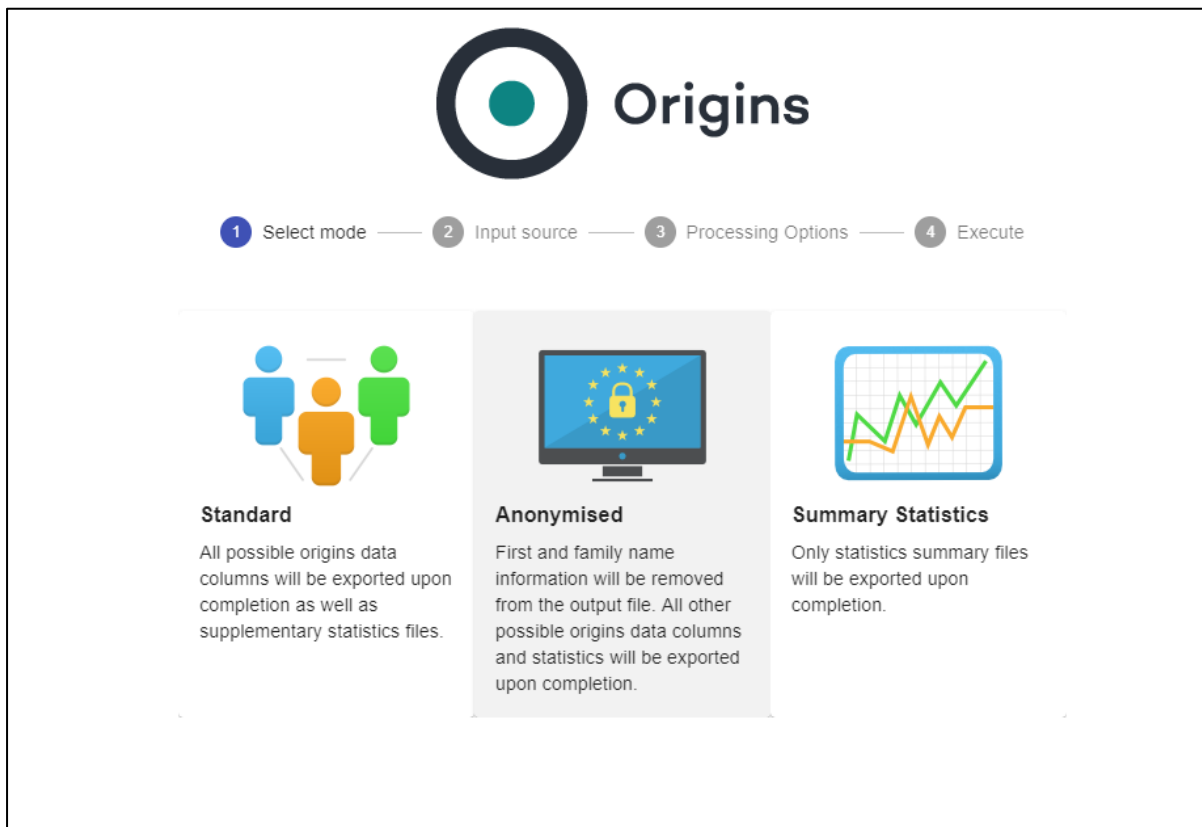
A logo indicating the identity of the distributor is shown at the top of the following and subsequent screens.

Below the logo is a travel bar showing the stage of use of the application that the user has reached.

The first stage is "Select Mode".

The "Select Mode" screen offers 3 modes of operation of the software.

- Standard mode

Under the "standard" mode of operation the application appends Origins and gender codes to an input file of names and outputs, a file containing one output record for each input record. This output file contains both the values contained against each record in the input file and the various categorisations relating to Origins and gender inferred by the software.

In addition the application outputs a series of "summary" files, each summary file containing the frequency distribution of input records by Origins Type, Origins Sub-group and Origins Group and by Origins Types grouped on the basis of various aggregations such as most common Language and Religion.

- Anonymised mode

Under the "anonymised" mode of operation the application operates in an identical manner to "standard" mode except that the personal and family name columns in the output file are suppressed. This mode is appropriate in situations where the user does not want to hold any file that identifies the inferred ethnic or religious heritage of any identifiable data subject.

- Summary statistics mode

Under the "Summary Statistics" mode of operation the application does not output any file containing Origins, gender or other codes for each input record. The only files that are output are the Summary Statistics files containing the frequency distribution of input records by Origins Type and derived groupings of the Origins Types.
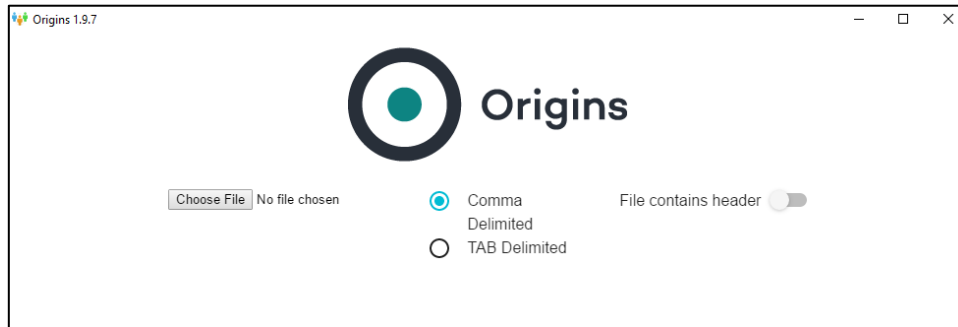
The second stage is the Input source stage.

In the Input source stage the user selects the input file and specifies whether it is in the form .csv (Comma Delimited) or .txt (TAB Delimited). The user also specifies whether the input file contains a header row. The default is that the input file does contain a header row.

Note that Excel files cannot be read by the application.

The user then clicks on "Choose file" to select the file which it is to be coded.

The largest number of records so far processed is 50 million. We do not know what the upper limit is in the number of records that can be processed.

WARNING: problems can occur where Excel files are converted to .csv files for Origins coding and where commas are found in the data fields. The software is unable to distinguish between field values and delimiters. SOLUTION: either search for commas in the input file and replace/remove them or download files as .txt files instead.

Select and open the file containing the data you wish to process.
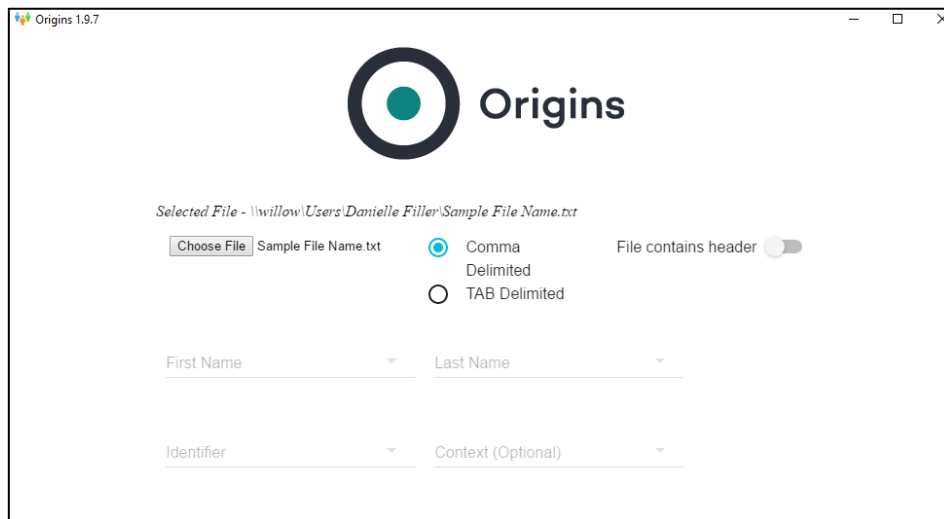
The application now lists the column headers on your input file (assuming you have them) and invites you to associate a minimum of three fields on your input file with three fields used in the application. If you do not have column headers it will list the values for the first record in each field.

The first is the "First Name" or personal name.

The second is the "Last Name" or family name.

The third is an "Identifier", typically a match key or a serial number (Required).

The fourth is a "Context Identifier", typically a postal code or, elsewhere, a unit of postcode geography. The text that appears here depends on the country in which the licensee is operating. For example in Australia it might refer to "Suburb" (Optional).

NOTE: It is not essential for the input file to contain a first name. If this field is not available on the input file then records will still be coded, but some other field needs to be entered against this position. Best is to point to a field containing numeric data only or a postcode.

The input file may contain more than four data fields. Fields not required by the system are discarded.

There is no need to assign a "Context Identifier".

Fields on the input file can be associated with more than one field in the software.

Click on any of these column headers to select from the list of input fields the one which matches the three or four data fields in the application that you wish to populate with data.

Having populated the screen with data, click "CONTINUE".
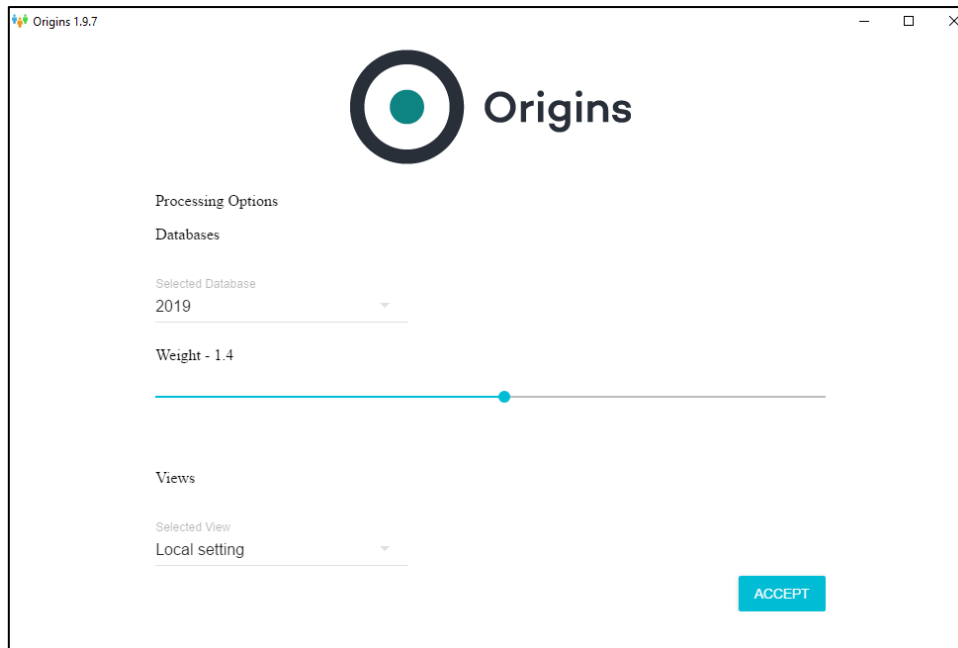
The Processing Options stage

The screen in the Processing Options stage offers the user a number of options.

The Origins software provides access to historic versions of the reference files associating names with their Origins classification. The benefit of storing historic versions of the reference files is that this makes it possible to undertake comparisons of customer files at different points in time using a consistent set of reference files. Were that not to be possible, and if only the most recent version of the reference files could be used, it would not be possible for a user to identify whether, for example, the reason for an apparent increase in the number of records coded as Lithuanian when coding the same file at different dates was the result of a real increase in their numbers or due to increases in Lithuanian names on the reference files.

Click on the arrow below "Databases" to select the version of the master reference file you wish to use. It is anticipated that these versions will be added to annually.

There may be reasons why some users believe that the coding by Origins would be improved by varying the relative weight given to first names and last names. The weighting bar enables them to do this. The numeric value (weights) that appears above the bar indicates the ratio of the weight given to the last name compared to the first name. The default is 1.4 to 1.

The box below, the "Views" option, allows the user to customise the Origins software application so that it works optimally when processing files of names originating from different countries. For example, the names "Thomas" and "Mann" are each common in both Germany and Great Britain. However, where the name "Thomas Mann" is found on a file of donors to a British charity he is likely to be of English origin. If the name appears on a list of

donors to a German charity, it is likely that the donor will be of German ancestry. This option determines how names which are given multiple Origins codes on the master reference files (such as "Thomas" and "Mann") are coded based on the source of the data file. French Canadian is the default view in Canada.
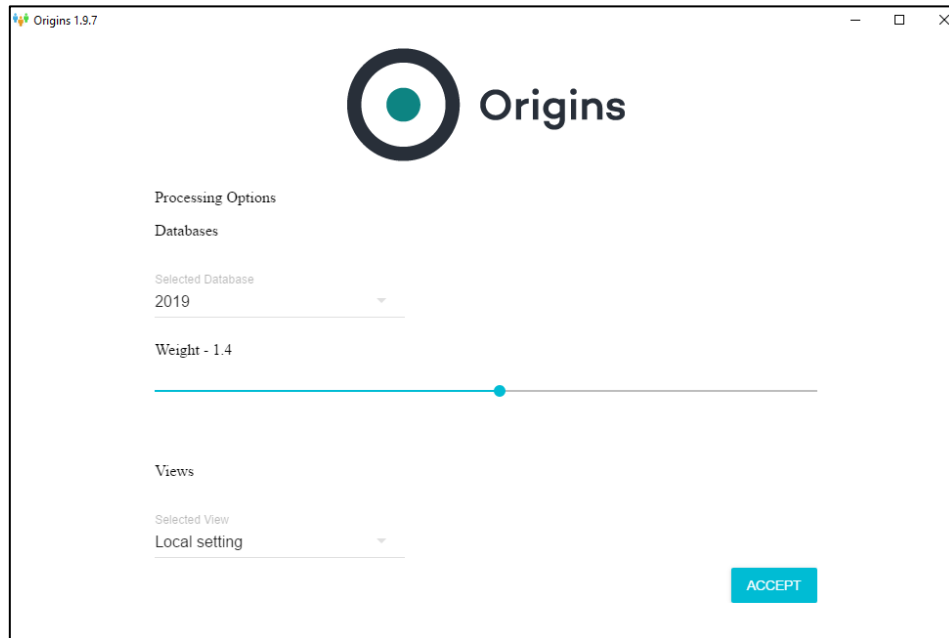
This facility is activated by the drop-down menu box "Local Setting" which denotes the country where the licensee operates. The default is configured by the token. Press "Accept", the application will prepare data files and the coding will begin.
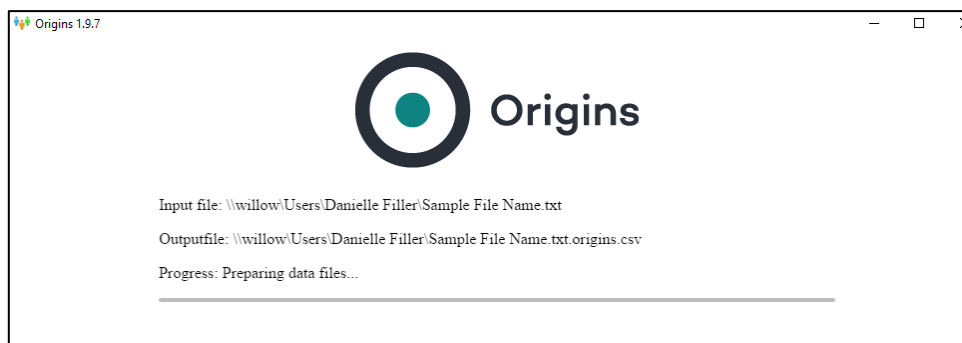
Once the processing has been completed, "Progress" reaches "100" and the user is given the option to check that the results are broadly as expected.

Click on "SHOW STATISTICS" to obtain the number and percentage of names that could be coded by Origins, the number and percentage of other names which could be recognised but not classified and the number and percentage that could not be recognised.

Typically, you would expect around 98.5% of names to be coded by Origins. If the percentage is significantly lower, it is likely that there is a problem with your input data.

Additionally, users can click on "PROCESS NEW FILE" to code a second or subsequent file. Once the processing is completed, three files will appear in the folder containing the input file. The file with the extension ".txt.origins.csv" will contain the output from the run. Files such as "origins.csv" and "groups.csv" contain the counts of records in each of the Origins types and groups. Counts are also available for a variety of other groupings of the Origins types, typically sub-groups, religion and language.

Files such as "origins.csv" and "groups.csv" contain the counts of records in each of the Origins types and groups. Counts are also available for a variety of other groupings of the Origins types, typically sub-groups, religion and language.

Sample File Name.txt.origins	5/7/2019 2:28 PM	Microsoft Excel C...	47 KB
Sample File Name.txt.origins.csv.groups	5/7/2019 2:28 PM	Microsoft Excel C...	1 KB
Sample File Name.txt.origins.csv.language	5/7/2019 2:28 PM	Microsoft Excel C...	2 KB
Sample File Name.txt.origins.csv.origins	5/7/2019 2:28 PM	Microsoft Excel C...	6 KB
Sample File Name.txt.origins.csv.religion	5/7/2019 2:28 PM	Microsoft Excel C...	1 KB
Sample File Name.txt.origins.csv.subgrou...	5/7/2019 2:28 PM	Microsoft Excel C...	2 KB

If you are processing more than one data set in a single session, for instance by clicking on "PROCESS NEW FILE", new sets of counts files will be created for each file processed. These counts files can be distinguished via the name of the input file which precedes the description of the category whose frequency is summarised. Thus File1.language.csv will contain language frequency counts for file 1, File2.language.csv will contain the language frequency counts for file 2 and so on.

Records are output in the same sequences as they appear on the input file.

COMPLIANCE WITH THE PROVISIONS OF THE GENERAL DATA PROTECTION REGULATION ACT

Webber Phillips receives many requests for guidance regarding the compliance of the Origins software with the provisions of the General Data Protection Regulation Act (GDPR) enacted in May 2018. This note aims to provide practical assistance in ensuring use of our data which is compliant with GDPR legislation.

A: Context
It goes without saying that users of the Origins software need to satisfy themselves that they have adequately considered regulations in relation to both:
•	Personal data
•	Sensitive personal data
Because the requirements of GDPR vary depending both on the status of the organisation and the uses to which data is put Webber Phillips are not in a position to give legal advice to individual users. However the software provides a number of different methods for accessing and using Origins data to meet the needs of different users to ensure that their use of personal data is compliant with the Act.

B: Public Interest Exemption
The provisions of the Act include what is referred to as a "public interest exemption". This exemption is likely to apply to an organisation which is under a statutory obligation to be

cognisant of variations in the use that members of different communities make of its services. This exemption is relevant to the level of use of services, channels of communication and to outcomes.

This statutory obligation is particularly relevant to public sector organisations and is frequently used to legitimate the appending of Origins to databases of service users. Public Interest Exemption is also relevant in so far as there is a statutory requirement to monitor the level of diversity of senior employees and to introduce policies to improve levels of diversity in the workforce.

This use of Public Interest Exemption does not relax obligations regarding the security of the data supplied using Origins. Standard practice is for our public-sector users to restrict access to any records that have been coded by Origins to members of their research and analysis teams. In most cases the data are held in a physically secure environment with restricted access.

C: The choice of modes for analysing personal data

When using the Origins software application the user is first required to select one from three different modes of analysis. These are:

- Standard
- Anonymised
- Summary statistics
- Postcode Origins

The Standard mode is appropriate for situations for which a public interest exemption has been agreed.

This mode outputs a series of summary files containing counts of input records falling into each of the various Origins classifications. In addition it outputs a file containing the personal name and family name of each input record together with various additional fields, such as Origins, obtained using reference files embedded in the application that associate personal and family names with the various Origins classifications and gender codes.

The outputs of the Anonymised mode are similar to those of the Standard mode other than that in this case the personal and family names on the input file are not included in the output file. This mode of analysis is designed to protect the identity of each input name but providing the input file contains a valid Unique Reference Number (URN) it allows Origins codes to be matched to data held anonymously against each data subject on databases and which are used exclusively for statistical analysis, not for communications.

The Summary Statistics mode has been designed to meet the needs of those users whose interest is restricted to obtaining statistics on the distribution of input records by Origins codes and who do not want to access the Origins code of any individual subject. This mode differs from the Standard mode by not outputting a list of each individual input record with its Origins code.

The Postcode Origins mode is used by organisations whose want to know the ethnic mix of the types of neighbourhood in which customers or client live. To use the Postcode Origins mode it is necessary to include the postcode of the data subject on the input file.

Postcode Origins can be used in a manner analogous to the postcode versions of Acorn and Mosaic. Using the postcode of the data subject Origins technology will:

- Flag whether, for each data subject, there are a sufficient number of adult names in his or her postcode for it to be possible to append to his or her record information on the cultural make-up of the postcode without compromising the personal data pertaining to any adult living there.
- If it is to identify whether the postcode has a significant non-white British population or whether it is almost exclusively white British.

- In the event that the non-white British population is significant, indicate which is the dominant minority grouping, subject to a frequency threshold.

OriginsCanada is a system that classifies consumers according to the part of the world from which their forebears are most likely to have originated. OriginsCanada uses the same information to code customers on the basis of their most likely language and religion. An age estimate and gender can also be appended. The output represents derived modelled data - estimate heritage tags. The estimate is not actual information regarding heritage, but a prediction based on generally available generic name dictionaries, genealogy records, etc. Our view is that the heritage tag when assigned to a name is not PII for which consent is required. Once modelled data are appended to actual customer records the resulting combined record could be considered PII and sensitive information depending on the organizations internal privacy policies.