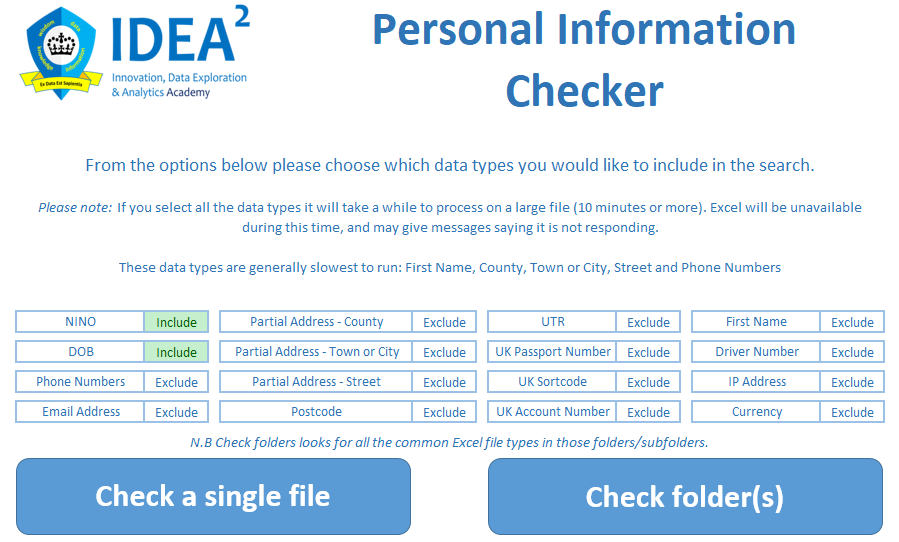
**Sensitive Data Checker**

On the front sheet, there is a set of option boxes for the different data types that can be included in the search.



If a data type is set to include, it will be searched for in the data. If it is set to exclude, it will not be searched for.

This is to offer options for speed of processing, especially when looking at large data files.

It also offers the options to run the tool in batches on large files. When you run a new search, if there are already results in the sheet the user is prompted to choose to replace previous results or not. If you choose not to replace, you can continue to build up the results.

The tool has two main elements; a regex search and a corpus search. The regex search is used when there is a pattern to look for in the data – e.g. National Insurance Number. The corpus search is for data types that do not follow a pattern, but where a list of pre-defined possible outputs can be used as a search criteria – e.g. a person’s name.

The tool currently identifies the following data types. (The notes section gives a bit of detail of which search type it uses, and how the search works)

|  |  |  |
| --- | --- | --- |
| **Data Type** | **Weighting** | **Pattern Notes** |
| National Insurance Number | 2 | Looks for the following pattern: A string that starts with two letters, 6 numbers, then an optional final letter. All letter matches are case-insensitive.  The first part excludes certain letters that are not used in NINOs - D,F,I,Q,U and V. There are also some combinations of 2 letters that are not used, and therefore excluded - BG, GB, NK, KN, TN, NT and ZZ.  The optional final letter can only be between A and D. Whitespace characters in between the pattern will still match the pattern overall. |
| Date of Birth | 1 | Searches for a day (number between 1-31), a month (number between 1-12), and a year (a 4 digit number starting with 19 or 20). These numbers can be seperated by either "-", " ", "/" or "."  Also searches for dates with the months as words e.g. 1st April 2001, 01 April 2001.  If there is either "-" or " to " between the dates, then the date is a range and less likely to be DOB, so the weight is divided by 4.  If the cell text contains the words DOB/D.O.B/birthday/Birth date then the weighting is doubled, indicating a stronger likelihood of the date being a DOB. |
| Postcode |  | Uses a REGEX pattern supplied by ONS. |
| IP Address | 2 | Looks for a pattern of 4 numbers, between 0-255, split by a ".". |
| Driver Number (Driving Licence) | 2 | Looks for a UK Driver Number, 16 consecutive alpha-numeric characters in the following pattern:  • Characters 1 – 5: Is the first five characters of the surname (padded with 9s if less than 5 characters long). • Character 6: Is the decade digit from the year of birth (for example, for 1987 it would be 8). • Characters 7 – 8: Is the month of birth. If the driver is female, the 7th character incremented by 5. • Characters 9 – 10: Is the date within the month of birth. • Character 11: Is the year digit from the year of birth (for example, for 1987 it would be 7). • Characters 12 – 13: Are the first two initials of the first names, padded with a 9 if there is no middle name. • Character 14: Is an arbitrary digit – usually 9, but decremented to differentiate drivers with the first 13 characters in common. • Characters 15–16: Are two computer check alpha characters. |
| UK Passport Number | 1 | Looks for a pattern of 9 consecutive numbers. Has the potential to find more matches than we are interested in, hence the reduced weighting.  If the word passport also appears in the text, this indicates a stronger likelihood and the weighting is doubled. |
| UK Sortcode | 1 | Looks for a pattern of 6 consecutive numbers, optionally split by a space or a "-". Has the potential to find more matches than we are interested in, hence the reduced weighting. |
| UK Account Number | 1 | Looks for a pattern of 8 consecutive numbers. Has the potential to find more matches than we are interested in, hence the reduced weighting.  If the word account/account number also appears in the text, this indicates a stronger likelihood and the weighting is doubled. |
| Email Address | 2 | Looks for any length/combination of alpha-numeric characters followed by @, followed by any length/combination of alpha-numeric characters interspersed with .'s. Needs testing with Twitter feeds. |
| First Name | 2 | Looks for first names, using a corpus of over 143,000 first names. The corpus contains all first names from England and Wales birth registers between 1838-2014.  Also looks for 47 common titles, if any of these appear in the text then the weighting is doubled. |
| Partial Address - County | 2 | Looks for counties, indicating a possible address, based on a corpus of 125 counties of the United Kingdom. |
| Partial Address - Town or City | 2 | Looks for cities or towns, indicating a possible address, based on a corpus of nearly 1000 towns and cities of the United Kingdom. |
| Partial Address - Street | 2 | Looks for 'street' suffixes, indicating a possible address, based on a corpus of 23 common suffixes. |
| Phone Numbers | 2 | Looks for a pattern of 11 consecutive numbers, with optional spaces after the first 4 or 5 numbers (area code), or half way between the last 6 numbers.  Also compares the start of the string against a corpus of over 650 known UK area codes, both geographical and non-geographical including business and mobiles.  2 is added to the weighting if an area code match is found. |
| Currency | 2 | Looks for either the £, $ or € symbol, indicating monetary values which might be linked to tax amounts. |

Once the document is searched, it shows a results screen:

This is sorted by the weighting, with the most likely sensitive data matches appearing at the top of the list.

It shows the string that has been identified as a possible match alongside the full cell content to show the string in context, which is useful to help decide what action needs taking.

It also contains a hyperlink that takes you to the identified cell within the document that allows the user to easily make any necessary amendments.

When the user closes the tool they are warned that all the data in the results page is about to be deleted, and offers them a chance to cancel the save and make any further actions they need on the results first. When the tool is allowed to close, it will delete all the results – thus saving the potential of sensitive data being located, and inadvertently saved in the tool which might not be stored in the correct location for sensitive data.

**Appendix 1: REGEX Patterns**

'Nino locator

Dim RE\_NINO As String

RE\_NINO = "(?!BG)(?!GB)(?!NK)(?!KN)(?!TN)(?!NT)(?!ZZ)(\s\*[A-CEGHJ-PR-TW-Z|a-ceghj-pr-tw-z][A-CEGHJ-NPR-TW-Z|a-ceghj-pr-tw-z])(\s\*-\*\d\s\*-\*){6}([A-D|a-d]|\s|.)"

Dim WE\_Nino As Integer

WE\_Nino = 2

'Date locator

Dim RE\_DOB As String

RE\_DOB = "(0[1-9]|[12][0-9]|3[01])[- \/.](0[1-9]|1[012])[- \/.](19|20)[0-9]{2}"

Dim WE\_DOB As Integer

WE\_DOB = 1

'Date locator 2

Dim RE\_DOB2 As String

RE\_DOB2 = "([1-9]|0[1-9]|1[0-9]|2[0-9]|3[0-1])(\s|th\s|st\s|nd\s|rd\s|-|/)(Jan|January|Feb|February|Mar|March|April|Apr|May|Jun|June|Jul|July|Aug|August|Sep|September|Oct|October|Nov|November|Dec|December)(\s|-|/)(19|20)[0-9]{2}"

Dim WE\_DOB2 As Integer

WE\_DOB2 = 1

'Postcode Locator

Dim RE\_PCDE As String

RE\_PCDE = "([Gg][Ii][Rr] 0[Aa]{2})|((([A-Za-z][0-9]{1,2})|(([A-Za-z][A-Ha-hJ-Yj-y][0-9]{1,2})|(([A-Za-z][0-9][A-Za-z])|([A-Za-z][A-Ha-hJ-Yj-y][0-9]?[A-Za-z])))) [0-9][A-Za-z]{2})"

'https://en.wikipedia.org/wiki/Postcodes\_in\_the\_United\_Kingdom

Dim WE\_PCDE As Integer

WE\_PCDE = 2

'IP Locator

Dim RE\_IP As String

RE\_IP = "(25[0-5]|2[0-4][0-9]|1[0-9][0-9]|[1-9]?[0-9])\.(25[0-5]|2[0-4][0-9]|1[0-9][0-9]|[1-9]?[0-9])\.(25[0-5]|2[0-4][0-9]|1[0-9][0-9]|[1-9]?[0-9])\.(25[0-5]|2[0-4][0-9]|1[0-9][0-9]|[1-9]?[0-9])"

Dim WE\_IP As Integer

WE\_IP = 2

'UK Driver Number

Dim RE\_DL As String

RE\_DL = "(\s[A-Za-z9]{5})([0-9])([0-9]{2})([0-9]{2})([0-9]{1})([A-Za-z9]{2})([0-9])([A-Za-z9]{2}\s)"

Dim WE\_DL As Integer

WE\_DL = 2

'Passport Number Locator

Dim RE\_Pass As String

RE\_Pass = "(\s\d{9}\s)"

Dim WE\_Pass As Integer

WE\_Pass = 1

'Sortcode Locator

Dim RE\_SC As String

RE\_SC = "(\s\d{2})(-|\s?)(\d{2})(-|\s?)(\d{2}\s)"

Dim WE\_SC As Integer

WE\_SC = 1

'Account Number Locator

Dim RE\_AC As String

RE\_AC = "(\s\d{8}\s)"

Dim WE\_AC As Integer

WE\_AC = 1

'Email Address Locator

Dim RE\_Email As String

RE\_Email = "(@)"

Dim WE\_Email As Integer

WE\_Email = 2

'Phone Number Locator

Dim RE\_Phone As String

RE\_Phone = "(\d){4,5}(\s)(\d){6,7}|(\d){4,5}(\s)(\d){3,4}(\s)(\d){3}"

Dim WE\_Phone As Integer

WE\_Phone = 2

'Currency Locator

Dim RE\_Currency As String

RE\_Currency = "(£|$|€)"

Dim WE\_Currency As Integer

WE\_Currency = 2