# Extract data from the NBN Gateway into R

Stuart Ball, JNCC.

October 7, 2013

## 1 Introduction

The National Biodiversity Network (NBN) is an on-line repository for biodiversity data from the UK. At the time of writing, it contained over 85 million species records in over 800 datasets. Data can be accessed via web-services provided by the Gateway web-site (for documentation see `http://data.nbn.org.uk/Documentation/Web_Services/`).

This package provides methods to get species records and other supporting information from the NBN Gateway. The functions fall into three tiers:

**Low level functions** to prepare the ground

- **makenbnurl** constructs a URL to call a service from the supplied parameters (and check they are correct!).
- **nbnLogin** uses dialog boxes to allow the user to enter username and password for the NBN gateway. This function also manages cookies.run the URL and return the JSON object obtained in response
- **runnbnurl** run the URL and return the JSON object obtained in response to the web-service call in the form of an R list structure.

**Functions that access a particular service** and return a JSON object

- **getFeature** get information about a "feature" (a location at which occurrences have been recorded) given its featureID.
- **getGroupSpeciesTVKs** given the name of a group (see `listGroups`) this function returns the pTVKs (preferred taxon version keys) for all members of that group. This is currently restricted to returning up to 20 results..
- **getOccurrences** get occurrences for a particular taxon or list of taxa. Returns a data.frame containing the occurrences. Optionally, the datasets from which observations are to be extracted and a start and end year can also be specified.
- **getTaxon** get information about a particular taxon given its Taxon Version Key (TVK).

**getTaxonomy** given its TVK, get details of the taxonomical heirarchy of a taxon.

**listDatasets** returns a dataframe of the datasets available from the NBN Gateway for reference.

**listGroups** returns a dataframe of the group definitions from the NBN Gateway for reference.

**listOrganisations** returns a dataframe of the organisation definitions from the NBN Gateway for reference.

**listVCs** returns a dataframe of the Watsonian vice-counties and their keys for reference.

**High levels functions** that manipulate the returned data for a particular purpose

**getSDMdata** get the necessary occurrence information for one or more species to run a Species Distribution Model. This returns a data.frame containing the x,y coordinates of occurrences, which is the format required by various R modelling packages (such as dismo), but the information can also be saved to a CSV file for use with external modelling software such as maxent.

**getFrescaloData** get the necessary data to run Mark Hill's Frescalo method to estimate species trends.

Some other utility functions are provided which manipulate grid reference and date information returned by the NBN Gateway:

**gridRef** takes a grid reference string (OSGB or OSNI) and extracts grid references at other precisions. For example, extract 10km square grid refs from the grid references returned from the Gateway.

**gridCoords** takes a grid reference string (OSGB or OSNI) and calculates the x,y coordinates of the bottom, left-hand corner of the grid square.

**datePart** takes the vague date information, returned in three fields (startDate, endDate and dateTypeKey) from the NBN Gateway and extracts elements of the date like the year or week, whilst properly taking into account the type of vague date.

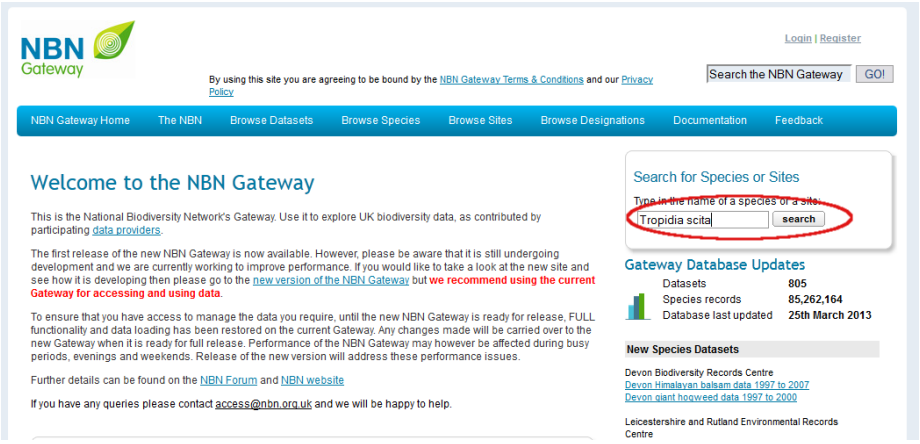
# 2 Getting species occurrence records

The `getOccurrences` function gets a data.frame of species occurrence records from the NBN Gateway. Columns include the name and TVK of the species and the date and location of the observation as a minimum, and may include other columns depending what has been submitted by the data providers and what access they allow.

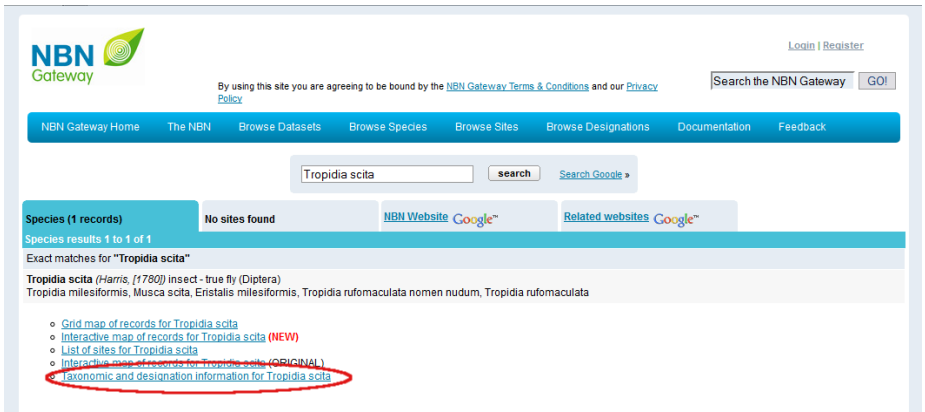
The minimum information required to request species occurrences from the NBN Gateway is the Taxon Version Key (TVK) of your target species. This is a 16-character string of (usually, upper-case) letters and numbers. For example, “NBNSYS0000007111”.

TVKs can be found by searching for a species on the NBN Gateway. At the time of writing, there is no web-service to do this, although it should be available soon. Consequently, this will have to be done manually.

On the NBN Gateway home page `http://data.nbn.org.uk/`, type the name of the species (“Tropidia scita” in the example) in the “Search for species or sites” box and press Return or click the “search” button.

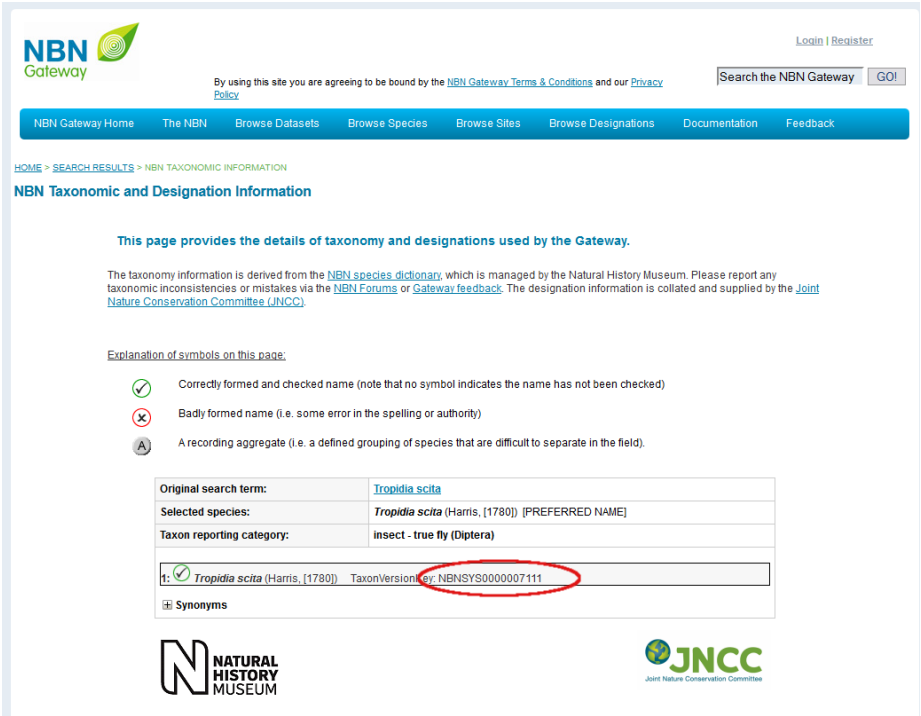



Hopefully, one or more matches will be found. Click the “Taxonomy and designation information for . . . ” link.




This will display the “NBN Taxonomic and Designation Information” page for

the species. The TVK is shown below the name of the species.




For example, the following example will get all publicly available observations of *Tropidia scita* from all datasets and for any date and lists selected columns for the first 10 rows:

```
> library(rnbn)
> dt <- getOccurrences(tvks="NBNSYS0000007111")
> dt[1:10,c("observationID","pTaxonName","location","startDate","endDate",
+           "dateTypekey")]
```

```
NULL
```

Occurrences for more than one species can be obtained by passing a list of TVKs, e.g.

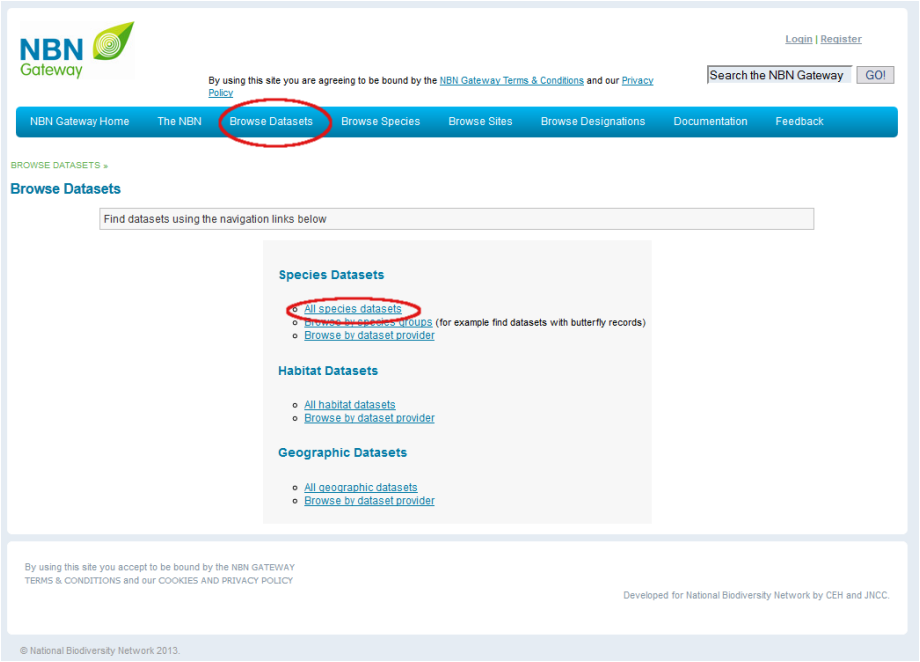
```
tvks=c("NBNSYS0000007111","NBNSYS0000007073")
```

Observations can be filtered so that they come only from datasets you trust by passing one or more dataset keys to the datasets parameter. A list of datasets can be passed in a way similar to a list of species keys. e.g.

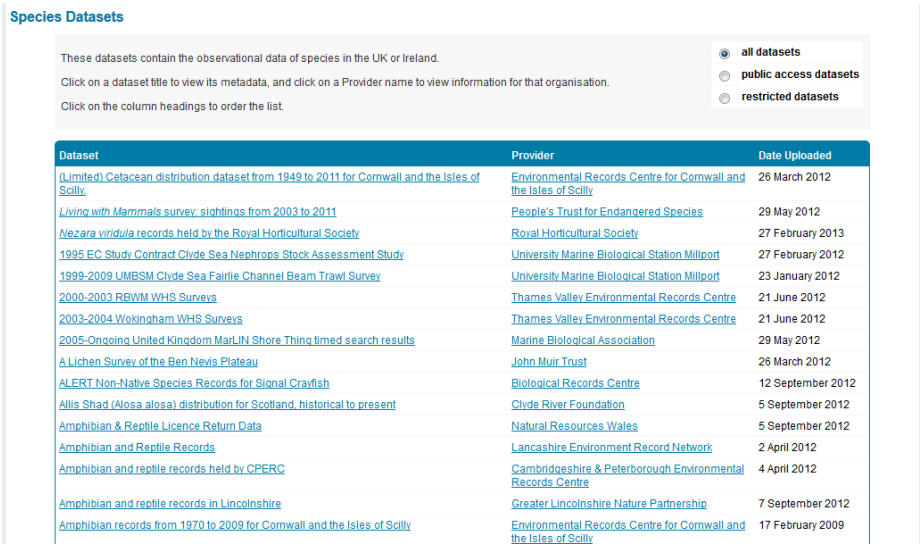
```
datasets= c("SGB00001","GA000483","GA000152","GA000306")
```

Dataset IDs are 8-character strings of upper-case letters and numbers. Like TVKs, at present, you will need to manually look them up from the NBN Gateway web-site as follows:

On the NBN Gateway home page `http://data.nbn.org.uk/`, click the "Browse Datasets" button in the menu bar at the top of the page and then click the "All species datasets" link.




Find the dataset you are interested in by scrolling through the alphabetic list of datasets that appears. When you find the one you want, click on its name.



**Species Datasets**

These datasets contain the observational data of species in the UK or Ireland.

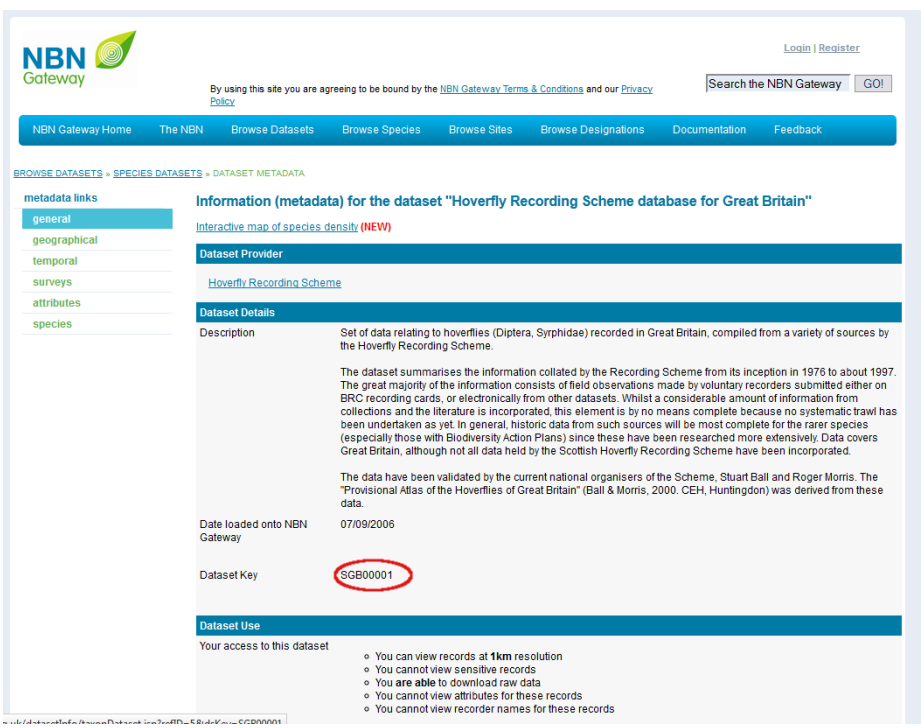
Click on a dataset title to view its metadata, and click on a Provider name to view information for that organisation.

Click on the column headings to order the list.

all datasets
public access datasets
restricted datasets

| Dataset | Provider | Date Uploaded |
|---|---|---|
| (Limited) Cetacean distribution dataset from 1949 to 2011 for Cornwall and the Isles of Scilly | Environmental Records Centre for Cornwall and the Isles of Scilly | 26 March 2012 |
| *Living with Mammals* survey: sightings from 2003 to 2011 | People's Trust for Endangered Species | 29 May 2012 |
| *Nezara viridula* records held by the Royal Horticultural Society | Royal Horticultural Society | 27 February 2013 |
| 1995 EC Study Contract Clyde Sea Nephrops Stock Assessment Study | University Marine Biological Station Millport | 27 February 2012 |
| 1999-2009 UMBSM Clyde Sea Fairlie Channel Beam Trawl Survey | University Marine Biological Station Millport | 23 January 2012 |
| 2000-2003 RBWM WHS Surveys | Thames Valley Environmental Records Centre | 21 June 2012 |
| 2003-2004 Wokingham WHS Surveys | Thames Valley Environmental Records Centre | 21 June 2012 |
| 2005-Ongoing United Kingdom MarLIN Shore Thing timed search results | Marine Biological Association | 29 May 2012 |
| A Lichen Survey of the Ben Nevis Plateau | John Muir Trust | 26 March 2012 |
| ALERT Non-Native Species Records for Signal Crayfish | Biological Records Centre | 12 September 2012 |
| Allis Shad (Alosa alosa) distribution for Scotland, historical to present | Clyde River Foundation | 5 September 2012 |
| Amphibian & Reptile Licence Return Data | Natural Resources Wales | 5 September 2012 |
| Amphibian and Reptile Records | Lancashire Environment Record Network | 2 April 2012 |
| Amphibian and reptile records held by CPERC | Cambridgeshire & Peterborough Environmental Records Centre | 4 April 2012 |
| Amphibian and reptile records in Lincolnshire | Greater Lincolnshire Nature Partnership | 7 September 2012 |
| Amphibian records from 1970 to 2009 for Cornwall and the Isles of Scilly | Environmental Records Centre for Cornwall and the Isles of Scilly | 17 February 2009 |

This will display a page of metadata about the selected dataset. The dataset key is near the bottom of the first section "Dataset Details" and just before the "Dataset Use" section starts.



NBN Gateway

Login | Register

By using this site you are agreeing to be bound by the NBN Gateway Terms & Conditions and our Privacy Policy

Search the NBN Gateway GO!

NBN Gateway Home | The NBN | Browse Datasets | Browse Species | Browse Sites | Browse Designations | Documentation | Feedback

BROWSE DATASETS » SPECIES DATASETS » DATASET METADATA

metadata links
general
geographical
temporal
surveys
attributes
species

**Information (metadata) for the dataset "Hoverfly Recording Scheme database for Great Britain"**

Interactive map of species density (NEW)

**Dataset Provider**

Hoverfly Recording Scheme

**Dataset Details**

| | |
|---|---|
| Description | Set of data relating to hoverflies (Diptera, Syrphidae) recorded in Great Britain, compiled from a variety of sources by the Hoverfly Recording Scheme.<br><br>The dataset summarises the information collated by the Recording Scheme from its inception in 1976 to about 1997. The great majority of the information consists of field observations made by voluntary recorders submitted either on BRC recording cards, or electronically from other datasets. Whilst a considerable amount of information from collections and the literature is incorporated, this element is by no means complete because no systematic trawl has been undertaken as yet. In general, historic data from such sources will be most complete for the rarer species (especially those with Biodiversity Action Plans) since these have been researched more extensively. Data covers Great Britain, although not all data held by the Scottish Hoverfly Recording Scheme have been incorporated.<br><br>The data have been validated by the current national organisers of the Scheme, Stuart Ball and Roger Morris. The "Provisional Atlas of the Hoverflies of Great Britain" (Ball & Morris, 2000. CEH, Huntingdon) was derived from these data. |
| Date loaded onto NBN Gateway | 07/09/2006 |
| Dataset Key | SGB00001 |

**Dataset Use**

Your access to this dataset

- You can view records at **1km** resolution
- You cannot view sensitive records
- You **are able** to download raw data
- You cannot view attributes for these records
- You cannot view recorder names for these records

If you do not require the metadata for each dataset, but instead want to look up the dataset ID by name you can use the listDataset function which returns the datasets currently held on the NBN.

```
> datasets <- listDatasets()
> # Preview some rows with short titles
> head(datasets[45:50,])
```

```
                                                title      key
46        Bedfordshire Bumblebees (BNHS) -  2006-2012 GA000700
47         Bedfordshire Coleoptera (BNHS) - 1986-2012 GA000674
48          Bedfordshire Diplopoda (BNHS) - 1975-1985 GA000675
49       Bedfordshire Dormice (BNHS/BDG) -  2000-2012 GA000703
50               Bedfordshire Fish (BNHS) - 1800-2011 GA000704
51 Bedfordshire Herpetofauna (BNHS/BRAG) -  1973-2013 GA000458
```

The range of dates for which you want to extract data can also be specified using the `startYear` and/or `endYear` parameters:

```
> dt <- getOccurrences(tvks="NBNSYS0000007111", datasets="SGB00001",
+                      startYear=1990, endYear=2006)
> dt[1:10,c("observationID","pTaxonName","location","startDate","endDate",
+           "dateTypekey")]
```

```
NULL
```

# 3 Getting data for Species Distribution Models and Frescalo

## 3.1 getSDMdata

The function `getSDMdata` gets the data required to fit a Species Distribution Mode for one or more species from the NBN Gateway. One element required by many modelling methods consists of the coordinates of the locations at which a species has been observed. This is often supplied in the form of a CSV file with either the x,y coordinates for a particular species in two columns or, if the method can fit a sequence of species in one run, then species name,x,y in three columns. If modelling is being done via an R package such as dismo (Hijmans *et al.* (2013)), then the coordinates are generally supplied as x,y columns in a matrix or data.frame. These functions assumes that you are fitting a model Using coordinates of the National Grid for either GB or Ireland.

## 3.2 getFrescaloData

Mark Hill's Frescalo method (Hill (2012)) calculates trends in the frequency of a species over time. It attempts to correct the frequency with which a species has

been recorded in a series of time periods using the total amount of recording. This is done by identifying the commonest species in a neighbourhood around a given location and then quantifying the recording effort in terms of the proportion of the commonest species that were recorded in the neighbourhood. The basic assumption is that the more recording, the greater the proportion of commoner species that will be discovered. One of the inputs required is a set of observations consisting of unique combinations of location, species and time period. `getFrescaloData` extracts this information from observations obtained from the NBN Gateway and writes them to a file in a suitable format for Frescalo. Grid squares are used to provide the locations.

### 3.3 Specifying species

The format of the species parameter is the same for `getSDMdata`, and `getFrescaloData`, i.e. a data frame with columns `tvk` and `name`. In general, you will probably specify one or a few species for `getSDMdata`, but a list covering a large number of species for `getFrescaloData`. This is because the Frescalo method corrects for recording effort using the amount of recording of common species in the same group. You will therefore need to list all the species covered, e.g. all the species in a family. This is most conveniently done as an external file with a line for each TVK. In preparing this file, you will also need to consider how the species should be aggregated. There may be subspecies or named forms where you wish to combine the observations under a single name or groups of species which need to be aggregated, for example, a recently split group of sibling species. This can be achieved by assigning the same entry in the `name` column to two or more entries in the `tvk` column.

Example (extracts from a CSV file):

```
name,tvk
Anasimyia contracta,NBNSYS0000007039
Anasimyia interpuncta,NBNSYS0000007040
Anasimyia lineata,NBNSYS0000007041
...
Platycheirus peltatus agg,NBNSYS0000006879
Platycheirus peltatus agg,NBNSYS0000006886
Platycheirus peltatus agg,NBNSYS0000033188
...
Volucella bombylans,NBNSYS0000007094
Volucella bombylans,NBNSYS0000172195
```

Here *Platycheirus peltatus* was split into a group of sibling species recently, so they are combined as an aggregate named *Platycheirus peltatus agg.* Also a form of *Volucella bombylans* is combined under the one species name.

If this is stored as `syrphidae.csv`, it can be loaded as follows:

```
sp <- read.csv("/path/syrphidae.csv", as.is=TRUE)
```

Notice the use of the `as.is` parameter to prevent strings (in this case, both the name of species and the TVK are character strings) from being loaded as factors.

### 3.4 Constructing periods for Frescalo

Periods are specified as a list with two items: `breakYear` and `plabel`.

`breakYear` should contain a list of year numbers starting at the earliest year you want to include and finishing with the latest. The number and size of steps is up to you and periods do not have to be of equal sizes.

`plabel` provides the labels which will be used to identify the periods in the output file. The labels should be in character format and there should be one less label than the number of breaks.

For example:

```
periods <- list()
periods$breakYear <- seq(from=1980, to=2012, by=2)
periods$plabel <- as.character(seq(from=1980, to=2010, by=2))
```

It is believed to be good practice to choose your break points so that roughly equal numbers of observations fall in each period. Since the amount of recording (or at least the number of records that have been captured in databases!) has tended to increase over time for many recording schemes, this implies that the earlier periods will most likely need to be longer than the more recent ones.

### 3.5 Parallelism

If the number of species to be covered is substantial, this function will take some time (many minutes) to run. The way that the function works is to find all unique values in the `name` column of the `species` parameter, then process each of these separately - using the corresponding `tvk` entry(s) to call `getOccurrences` and get observations for that species. The unique location/species name/period combinations are then extracted and appended to the growing output file. This is "embarrassingly parallel" and scales almost linearly with the number of CPUs available to run it, i.e. a dual core machine will take only slightly more than half as long as a single core machine. The `foreach` package `%dopar%` operator is used. Therefore you can use any available methods to register a parallel backend for `foreach` before calling this function. If no backend is registered and multiple CPUs are detected, they will be used automatically.

## References

Hijmans, R., Phillips, S., Leathwick, J. & Elith, J., 2013. dismo: Species distribution modeling. r package version 0.8-11.
URL `http://CRAN.R-project.org/package=dismo`

Hill, M.O., 2012. Local frequency as a key to interpreting species occurrence data when recording effort is not known. *Methods in Ecology and Evolution*, **3**, 195–205.
URL http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2011.00146.x/pdf