Ecological Informatics 24 (2014) 132–147

Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf

Vesper: Visualising species archives

Martin Graham [*], Jessie Kennedy

School of Computing, Edinburgh Napier University, 10 Colinton Road, Edinburgh EH10 5DT, UK

ARTICLE INFO

Article history:
Received 30 April 2014
Received in revised form 12 August 2014
Accepted 17 August 2014
Available online 30 August 2014

Keywords:
Information visualisation
Data quality
Darwin Core Archive
Open source
Biodiversity

ABSTRACT

Vesper (Visual Exploration of SPEcies-referenced Repositories) is a tool that visualises Darwin Core Archive (DwC-A) datasets, and is aimed at reducing the amount of time and effort expended by biologists to ascertain the quality of data they are generating or using. Currently, DwC-A quality checking is limited to table outputs of data 'existence' and compliance with DwC-A format guidelines via the online DwC-A archive validator and reader. Whilst these tools thoroughly examine the presence of data, and the correctness of data structure against the DwC-A schema, they do not give any insight into the underlying quality of the data itself.

Built on top of the D3 JavaScript library, Vesper analyses and displays DwC-A datasets in three fundamental dimensions—taxonomic, geographic and temporal—with a visualisation dedicated to each of these aspects of the data. By viewing a dataset's composition in these dimensions, a data consumer can judge whether it is suitable for the tasks or analyses they have in mind, whilst a data provider can identify where a dataset they've constructed may fall short in terms of data quality i.e. does it contain data that is obviously incorrect such as the classic longitude inversion that places North American specimens in China. A further visualisation of the taxonomic dimension can reveal the subtaxa distribution of reference taxonomies—whilst a simple table reveals the presence or not of certain data types for each record to give an overall data 'existence' profile for the dataset. Selections of parts of a dataset within one visualisation are linked to the other visualisation displays for that dataset, permitting the discovery of whether data quality issues are restricted to identifiable sub-portions of the dataset.

Vesper can handle client-side data sets of a million entities within a browser by judicious use of data filtering, as many of the data types within individual records are not necessary to judge the geographic, temporal or taxonomic distribution and extent of a dataset. Thus, many of the more verbose fields in the file can simply be passed over during an initial data decompression stage. Furthermore it can provide limited name and structure matching of a dataset against DwC-A packaged reference taxonomies to indicate data quality relative to sources outside the archive. A selection of annotated example scenarios shows how Vesper can reveal data quality issues in DwC-A archives.

1. Introduction

Data quality and coverage are major issues for biodiversity datasets. A recent large-scale survey (Ariño et al., 2013) and related analysis (Faith et al., 2013) carried out by GBIF revealed user concerns about quality and coverage for biodiversity data in general, especially at the pre-publishing stage. In parallel, the Darwin Core Archive (DwC-A) format has been proposed as the standard with which to publish and transfer self-contained species occurrence and checklist data, especially to the GBIF network. Hence, a tool that could help identify data quality issues and gaps in DwC-A files would help alleviate the current concerns voiced by publishers and users of such data.

Automated tools are useful for rigorous syntactical and lexical checking of data, such as checking whether an xml document verifies against a schema, or that a date field follows a certain convention, but understanding the range and coverage of data more often than not requires human assistance. To this end the techniques found in Information Visualisation are well placed, allowing users to visually assess data and form opinions and hypotheses as to the reasoning and suitability of such data for a given task. When the appropriate visualisations are used, such as tree displays for hierarchical information, it is especially powerful as it utilises the pattern-finding powers of human visual perception (Wertheimer, 1938): uncovering outliers and broad patterns and trends.

Vesper provides an interactive set of visualisations that allows biologists/ecologists to explore the range and data quality of DwC-A datasets. By showing linked, interactive visualisations of taxonomic, geographic and temporal coverage it helps both data publishers and users to assess the quality and coverage of the data contained within

* Corresponding author. Tel.: +44 131 455 2749; fax: +44 131 455 2727.
E-mail addresses: m.graham@napier.ac.uk (M. Graham), j.kennedy@napier.ac.uk (J. Kennedy).

http://dx.doi.org/10.1016/j.ecoinf.2014.08.004
1574-9541/© 2014 Elsevier B.V. All rights reserved.

the archive, and whether it is sufficient for their particular purposes, either for publication (upload) to a data aggregator or for use in analyses. Thus, Vesper can help in achieving several of the recommendations put forward in (Faith et al., 2013) as responses to the perceived weaknesses in current biodiversity data gathering and recording. These specific recommendations, with original numbering, are:

- *Recommendations relating to data gaps, data volume, and data quality*
  6. *Initiate the following steps to enhance the trust-worthiness of GBIF mobilised data:*
     g. *Improve pathways for data publishers to provide warnings about biases or errors in the data at an early stage of discovery and publishing process.*
     j. *Expedite efforts in improving taxonomic and geo-spatial quality of GBIF mobilised data. This task includes attention to geo-referencing.*
     K. *Improve fitness-for-use of data at the data producer and/or primary publisher stage.*

The following sections describe the DwC-A data format that Vesper utilises, and fleshes out the background to the information visualisation techniques that Vesper uses in the course of its operation.

1.1. Darwin Core Archives

Darwin Core Archives (DwC-A) (GBIF, 2011b) are self-contained datasets that contain either a set of species occurrences or a reference taxonomy, both of which can be packaged along with an array of associated data files. In essence, the DwC-A is a zip file, containing a collection of related files as shown in Fig. 1. Firstly, there is always an XML metadata file that outlines the relationships and data stored in the other files in the archive. All the other files, with the exception of a possible Ecological Metadata Language (EML) (Fegraus et al., 2005) file, are plain text Delimiter Separated Values (DSV) files that store data as tables where each row is a record, and each ‘column’ a data field. The metadata file declares one of these DSV files as the *core* file, and any other DSV files present must hold *extensions* to the data in that core file, linked by a common set of IDs across the files. The meta file also describes the contents of each file at a wide range of detail, ranging from issues as low-level as what delimiters each file uses, which columns are the keys that connect the different tables, to what Darwin Core (or extension thereof) field is described in each column, and optionally, the schemas and data dictionaries that were used to control the data vocabulary in a particular column.

One of the driving factors behind the development of the DwC-A standard was the reduction in transfer traffic when downloading a dataset (GBIF, 2011b). A DSV-based description of data may contain redundant data points (for instance some datasets record the Kingdom for every specimen, even when that Kingdom is always the same), but it is still more compact than the equivalent in XML. When the DSV files themselves are compressed they can be orders of magnitude smaller than the uncompressed XML equivalent. In addition to this the file can be transferred as a whole unit rather than requiring multiple network requests. This is illustrated (GBIF, 2011b) with the following example scenario:

> *“Sharing entire datasets as Darwin Core Archives instead of using pageable web services like DiGIR and TAPIR allows much simpler and more efficient data transfer. For example, retrieving 260,000 records via TAPIR takes about nine hours, and involves issuing 1,300 http requests to transfer 500 MB of XML-formatted data. The exact same dataset, when encoded as DwC-A and zipped becomes a 3 MB file.”*

GBIF already provides a checking tool for DwC-A datasets (GBIF, 2013); however it focuses predominantly on syntactic issues, such as making sure that the archive is valid according to the DwC-A schema and an analysis of whether record ids in extension files match to the ids in the core file. It has some sanity checking on synonyms, but doesn't give an indication of the coverage of the data dimensions contained within and only gives an indication of missing data for the first 100 records. Vesper's goal is to step beyond this (we assume syntactical checking on a data set with GBIF's tool has been carried out) and is to visualise the distribution of the data within DwC-A's in geographic, taxonomic and temporal terms. Next, we describe some of the background in visualising these three components of the data.

1.2. Information Visualisation

Information Visualisation is the field of graphically and interactively rendering data to make patterns, trends and outliers within the data readily apparent to users. The most common examples of visualisation within biodiversity research are geographic visualisations: displaying specimen individuals and aggregations in various forms (markers, heat maps, clusters, polygons) upon a background map, which then allows users to explore spatial distribution patterns. Indeed, with a plethora of open-source GIS tools (Steiniger and Hunter, 2013) and public APIs to mapping services such as Google Maps, OpenStreetMap etc., it is rare to find a biodiversity website that doesn't offer some mapping of the geographical spread of specimen records. Related work has explored incorporating further elements into a map visualisation to show data besides geographic spread, for example HerbariaViz (Auer et al., 2011) displays an aggregation of date records by month over localities. Interestingly enough they describe how they had to quality control their selected dataset beforehand to remove records with non-existent or partial geographic co-ordinates. Another example, BirdVis (Ferreira et al., 2011), displays tag clouds over a geovisualisation generated from data collected by a collaborative bird monitoring project.

Visualisation of taxonomically-organised data is essentially the visualisation of tree structure, which has been comprehensively researched (Schulz, 2011). The basic choices for visualising tree structure can be summarised as either node-link, where relationships are denoted by

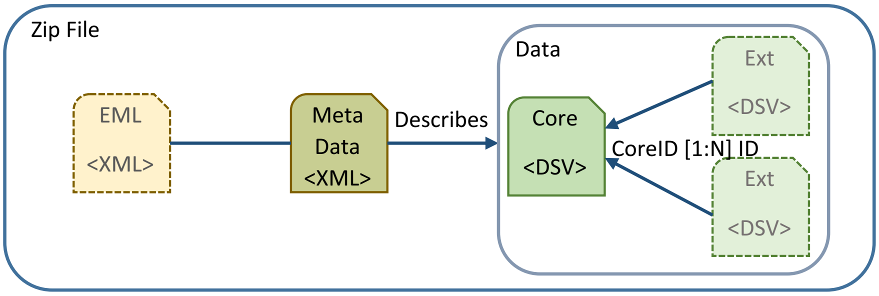


Fig. 1. DWCA file contents. A metadata file points to an optional EML file and describes the contents and syntax of associated text files. One of the text files is defined as the core file and is mandatory; the others are optional.

Table 1
JavaScript libraries used in the development of Vesper visualisations.

Library	Where it's used
D3	For binding data to visual elements in taxonomic and temporal views
JQuery-UI	Non-visualisation user Interface components (tabs, accordions)
Leaflet-JS	Used along with leaflet plug-ins to generate geographic view

Table 3
Visualisations and associated data fields required in the DwC-A.

Visualisation type	Fields required in DwC-A core/extension files
Normalised taxonomy	AcceptedNameUsageID, ParentNameUsageID
Denormalised taxonomy	Subset of {Kingdom, Class, Order, Family, Genus etc.}
Map	DecimalLongitude, DecimalLatitude, GeodeticDatum
Timeline	EventDate

drawing links (edges) that visually connect tree nodes, or one of a host of *implicit* styles, where relationships are indicated by adjacency or nesting of nodes. Schulz et al. (2011) give a wide-ranging survey of this latter type of tree visualisation. Picking a style of tree visualisation is often just a personal preference, but the data type also plays a part. Phylogenies for instance are always displayed using a node-link style display (Block et al., 2012; Munzner et al., 2003), because the length of the edge is a vital part of the information. In taxonomies, the edge merely indicates membership, so here the data can be visualised using one of the implicit tree visualisation styles (Graham and Kennedy, 2007b; Hong et al., 2003), which generally allow a greater data density.

Temporal data is the least frequently visualised type of the three data types we are concerned with—interaction with temporal data often consists of setting a filter or range on data to view in other visualisations (e.g. only show specimens on the map collected after 1980), though spatio-temporal visualisations (Andrienko et al., 2010) are becoming more prevalent. A suitable visualisation for presenting temporal data can be deduced by analysing the data and tasks concerned as detailed in Aigner et al. (2007) e.g. is the data time-points or ranges, are we looking for cyclical or linear trends, is the data to be shown in aggregated form etc. HerbariaViz provides visualisations that reveal annual patterns by aggregating data in Coxcomb charts, and BIDDSAT (Otegui and Ariño, 2012) shows temporal data in a polar coordinate scatterplot, the angle mapping to month and the radius to year. Where there are a large number of measures to display over time then stream graphs (Byron and Wattenberg, 2008) are a recent and popular alternative to stacked bar charts. Otherwise, when collating and aggregating a single or a few series of temporal data (e.g. when collections of specimens were collected) the resulting visualisation is often, and wisely, a fairly simple line plot or bar chart: BirdVis for example shows relative species occurrences over time through a line chart display.

Viewing data and cross-linking selections across different views of the same data set produces a *multiple view visualisation* (Roberts, 2007), with the advantage of showing how data arranged within one view, representing a particular facet of the data, appears when shown in other views representing different data facets. This has proven to be a powerful mechanism for showing relationships and trends between data facets for selected sub-groups of data, e.g. Fyfe et al.'s (2009) analysis of historical hotel visits uses connected geographical and temporal visualisations of data to reveal patterns such as seasonal spikes in hotel occupation in county seats. A multiple view system when applied to DwC-A data could reveal whether a group of records having data quality issues in one dimension also have problems or not in the other data dimensions.

Probably the closest work in nature to Vesper is the BIDDSAT system and other related work by the same author (Otegui and Ariño, 2012; Otegui et al., 2013). They construct visualisations over the same three fundamental dimensions of what, where and when (taxonomic, geographic and temporal) to allow users to view the data quality of records stored in the GBIF database. The main technical differences are that VESPER focuses on Darwin Core Archives, which are data stored in files rather than databases, and uses client-side parsing and rendering technologies rather than server-side. The biggest conceptual difference is that focusing on Darwin Core Archives allows data providers to check data quality before it is uploaded to an aggregator site such as GBIF, and for a data user it allows the quality checking of a data set that sits outside of a database. Also, their server-based application does not work as a multiple view system and thus cannot leverage the associated advantages, such as viewing selections made in one view in different contexts.

Having described the problem that Vesper tackles and the related background work, the rest of the paper describes the technologies and techniques that underpin Vesper, and then describe a number of annotated scenarios that demonstrate how Vesper can reveal data quality issues within DwC-A archives. This is followed by a conclusion that outlines the situations in which the use of Vesper would be most advantageous.

2. Material and methods

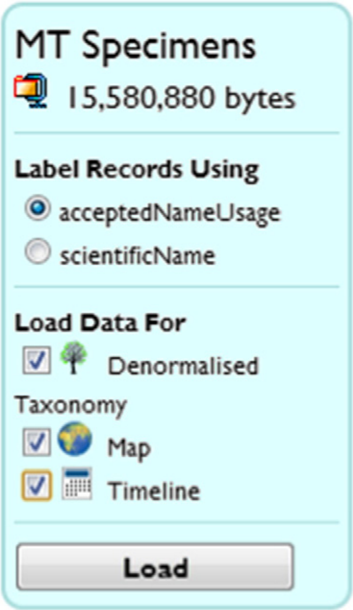
2.1. Technologies

Vesper has been developed to take advantage of the growing client-side abilities of modern web browsers. The latest HTML specification (HTML5) adopted by the World Wide Web Consortium has a range of associated technologies, including updated styling abilities and specifications (CSS3), widespread support of Scalable Vector Graphics (SVG), and the ability to retrieve and store client-generated data locally (File API). These are gradually being incorporated into the latest generation of web browsers, though completeness inevitably varies by browser and version (Deveria, 2014).

Vesper's base technologies require a modern HTML5-compliant web browser that supports JavaScript, SVG and CSS3, as it is primarily based on visualisations developed using the D3 data binding library (Bostock et al., 2011) that has sparked much web-based visualisation development, but for map rendering uses the Leaflet.js (2014) library and plugins on top of OpenStreetMap (Haklay and Weber, 2008) data. Table 1 summarises the main visualisation libraries used to develop Vesper, and on top of these work to construct and integrate the different views was carried out by the authors. This extra work mainly involved developing a cross-view selection object, and extending the existing visualisations to cope with the concept of selections i.e. that portions of the data could be marked as selected. The rapid pace of development in this area also means that new libraries now exist, such as C3.js that

Table 2
Datasets used during development and illustrated examples.

Data set	Type	Source	Record volume	DwC-A size (zipped)	Creation date
ENA reference	Reference taxonomy	EBI, Cambridge, UK	982,000	24 MB	9th April, 2014
GermanSL	Reference taxonomy	GBIF, Copenhagen, Denmark	29,000	1 MB	2012
VAScan	Reference taxonomy	Candensys, Canada	26,000	2 MB	2012
HIBG	Botanical garden specimen collection	Candensys, Canada	9000	0.5 MB	2012
MTSpecimens	Specimen collection	Candensys, Canada	140,000	15 MB	19th April, 2012



MT Specimens

15,580,880 bytes

Label Records Using

● acceptedNameUsage

○ scientificName

Load Data For

☑ Denormalised Taxonomy

☑ Map

☑ Timeline

Fig. 2. User dialog to choose the visualisations and label field which the chosen data set can support.

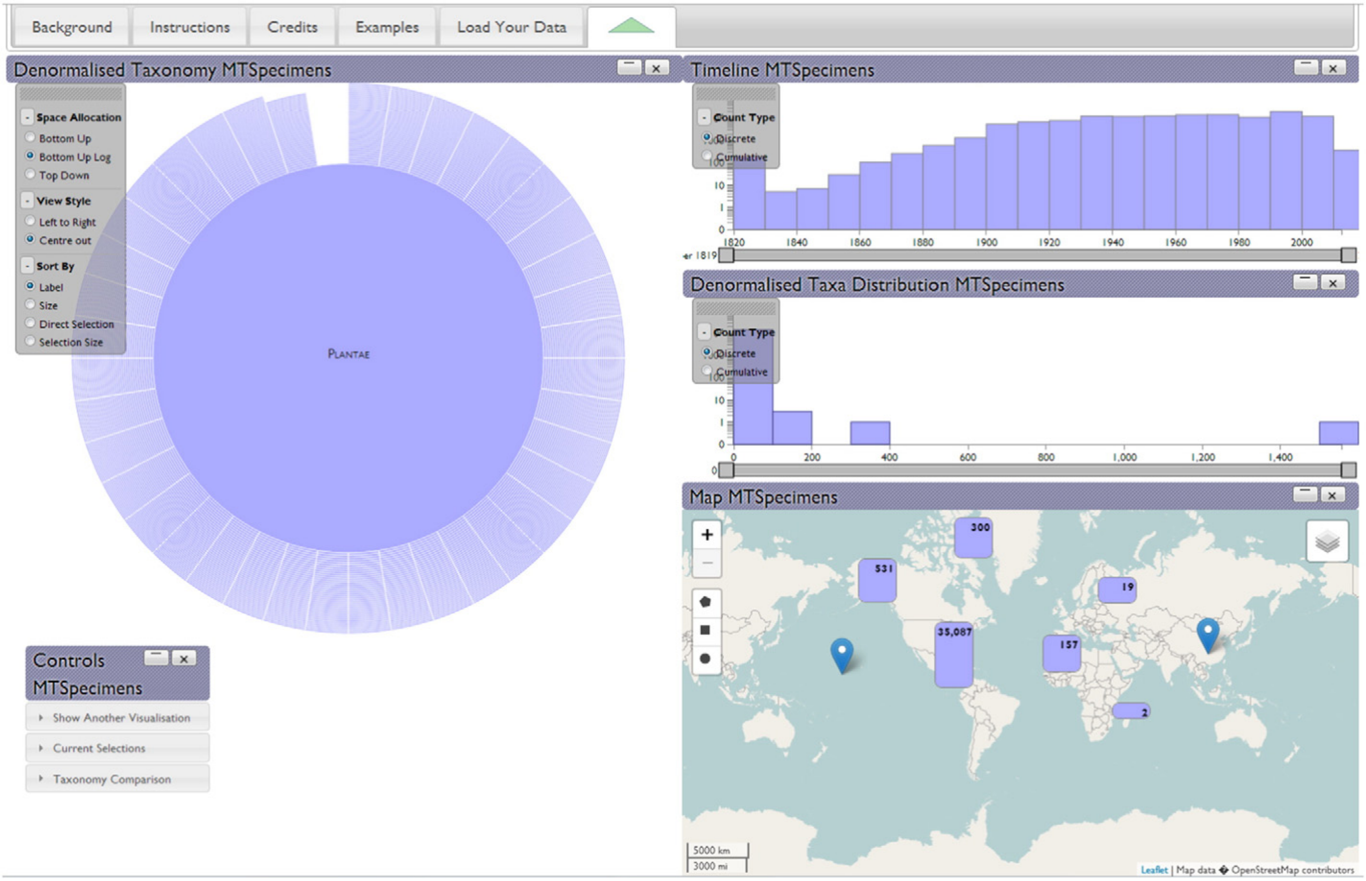
could be used to replace the bespoke bar charts developed for Vesper, and the external libraries Vesper does use keep improving in quality and capability with each version.

2.2. Data

To aid development, we identified suitable DwC-A datasets that held the different information types such archives can store and that we then aimed to cover with Vesper. These were a large reference taxonomy, a smaller taxonomy with many synonyms, a reference taxonomy of vascular plants from Canada, a botanic gardens specimen collection, and a moderately large specimen collection that had the full set of geographic, taxonomic and temporal data available. The details of each data set are shown in Table 2.

As stated, one of the reasons behind the development of the DwC-A standard was a reduction in transfer traffic as the datasets are zipped text files. However, whilst this accelerates data transfer across networks, they must obviously at some point be uncompressed to access the data within. This is problematic for browsers and JavaScript, as development efforts have focussed on compatibility and speed efficiency, hence the popularity of micro-optimisation sites such as jsPerf (JSPerf, 2014), rather than memory efficiency. A 25 MB DwC-A such as the ENA reference taxonomy can contain 250 MB of uncompressed text, which in turn expands to occupy over 500 MB of browser memory (in JavaScript, string characters are stored in UTF-16 format), which typically results in the browser crashing before any data visualisation can be attempted.

However, since we have decided to show the what, where and when of the datasets, we can unzip and query the meta.xml part of the DwC-A, establish if the necessary fields are present for each aspect of the data, and then offer the appropriate visualisations as choices before unzipping the remaining data files in the archive: Table 3 gives a summary of the fields required for the rendering of each visualisation type. In addition to this the core file's fields are searched for



Space Allocation
○ Bottom Up
● Bottom Up Log
○ Top Down

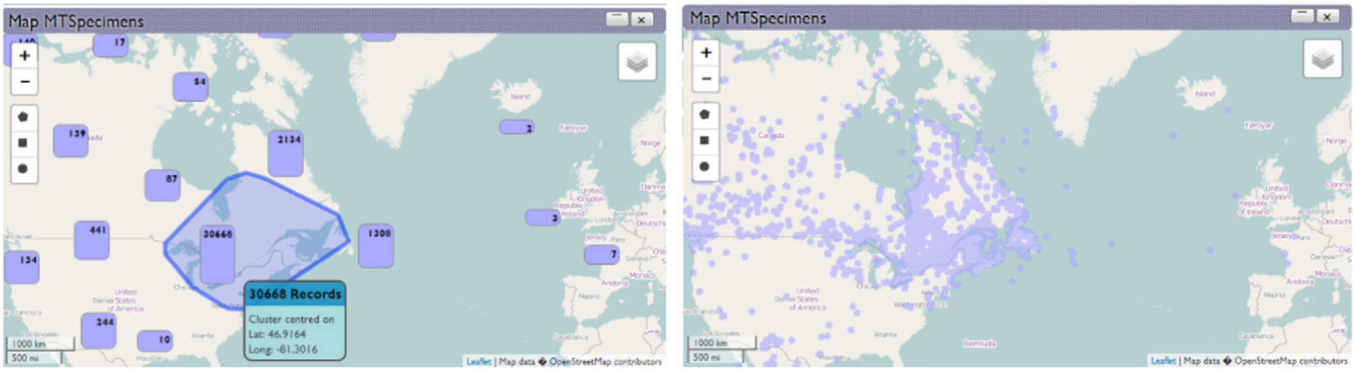
View Style
○ Left to Right
● Centre out

Sort By
● Label
○ Size
○ Direct Selection
○ Selection Size

Count Type
● Discrete
○ Cumulative

Count Type
● Discrete
○ Cumulative

Fig. 3. Full screen shot of typical VESPER output. On the left is the taxonomic view, top right is the temporal view, middle right is the taxonomic distribution view and bottom right is the geographical view.


Map MTSpecimens
17
54
139
2134
87
441
30668
1308
134
244
10
2
3
7
1000 km
500 mi
Leaflet | Map data © OpenStreetMap contributors
Map MTSpecimens
1000 km
500 mi
Leaflet | Map data © OpenStreetMap contributors

30668 Records
Cluster centred on
Lat: 46.9164
Long: -81.3016

Fig. 4. Clustered and unclustered views of geo-referenced species data. Both layers can be activated simultaneously.

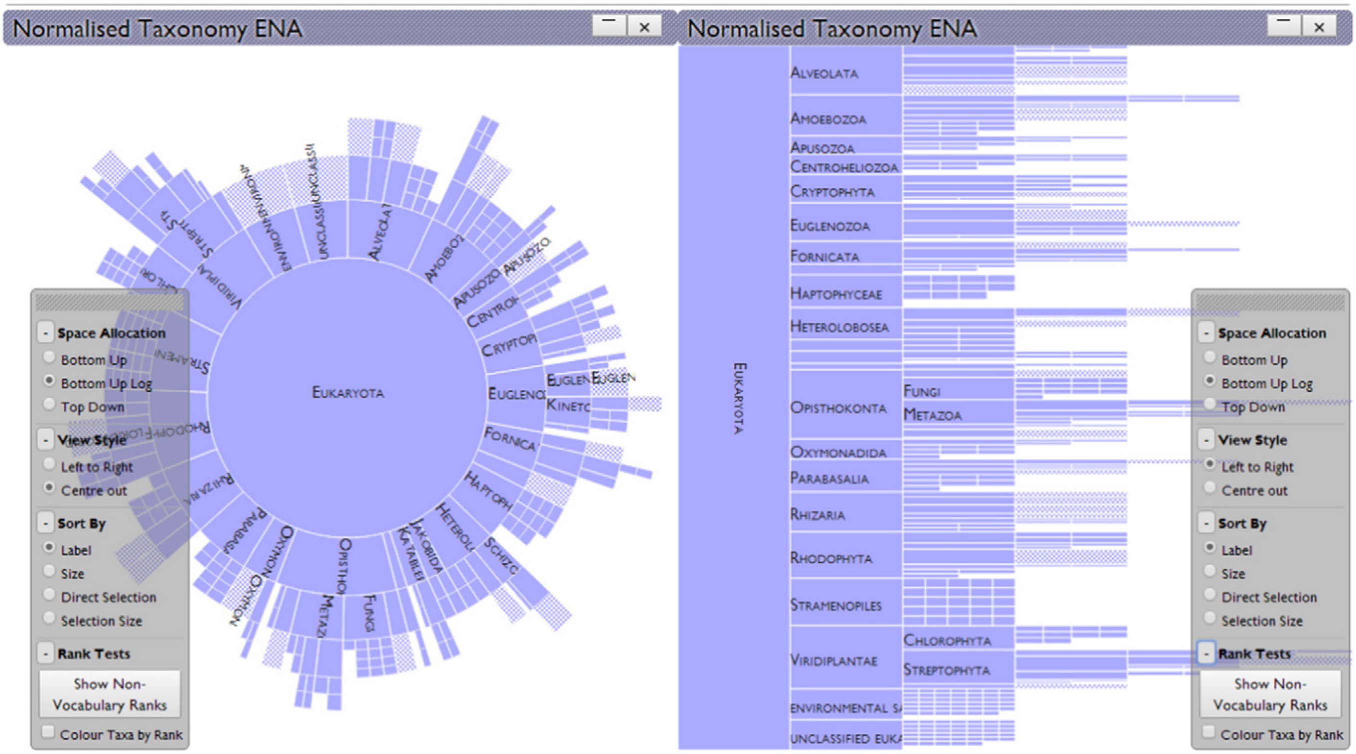
suitable candidates for human-readable labelling of specimens/taxa—scientificName, acceptedNameUsage etc. and these are again offered as a choice prior to the beginning of the parsing. Then, taking advantage of the fact that the more verbose, descriptive data fields are not used in the visualisations, we extended an existing open-source JavaScript zip file reader (JSZip) to retain only the decompressed data in each file required for visualisation purposes. This substantially reduced the memory footprint of an unzipped archive, some of which would have otherwise been too large to process within the browser. An example of the dialog presented within the browser after both the DwC-A has been loaded and the metadata unzipped and processed is shown in Fig. 2.

2.3. Shared view properties

Once parsing has finished, the browser initialises the chosen views and presents them to the user as in Fig. 3. Each view is presented in its own area of the browser, and can be repositioned using each visualisation's draggable top bar. Also in this top bar are controls to minimise/maximise and close that particular view—and closing the "Controls" sub-view closes all associated views for that dataset.

Each of the three main types of view—taxonomic, geographic and temporal—all employ visual clustering to show data that would otherwise overwhelm the visualisation, so for the most part are displaying aggregates rather than individual records. This is unsurprising in the first instance; a taxonomy after all is a mechanism for recursively grouping organisms into sets based on similarity. The geographic view uses a clustering plug-in that groups data points by geographical proximity based on the current map scale, and zooming in or out splits or joins the marker clusters into smaller or larger clusters. Lastly, the temporal visualisation employs a bar chart that uses a fixed range of bars to show in the visualisation—too many makes them difficult to see and interact with, too few makes the visualisation too granular to be of practical use. So depending on the timescale under consideration a bar can contain data spanning days or decades.

All the visualisations share a consistent set of interactions for basic tasks. A left-button mouse-click drills down into the data, splitting the



Normalised Taxonomy ENA

- Space Allocation
○ Bottom Up
● Bottom Up Log
○ Top Down
- View Style
○ Left to Right
● Centre out
- Sort By
● Label
○ Size
○ Direct Selection
○ Selection Size
- Rank Tests
Show Non-Vocabulary Ranks
□ Colour Taxa by Rank

Normalised Taxonomy ENA

- Space Allocation
○ Bottom Up
● Bottom Up Log
○ Top Down
- View Style
● Left to Right
○ Centre out
- Sort By
● Label
○ Size
○ Direct Selection
○ Selection Size
- Rank Tests
Show Non-Vocabulary Ranks
□ Colour Taxa by Rank

Fig. 5. Two views of the ENA reference taxonomy rooted at *Eukaryota*, centre-out (sunburst) and left-to-right (icicle). The control panel allows the view to be switched along with other settings for the visualisation.

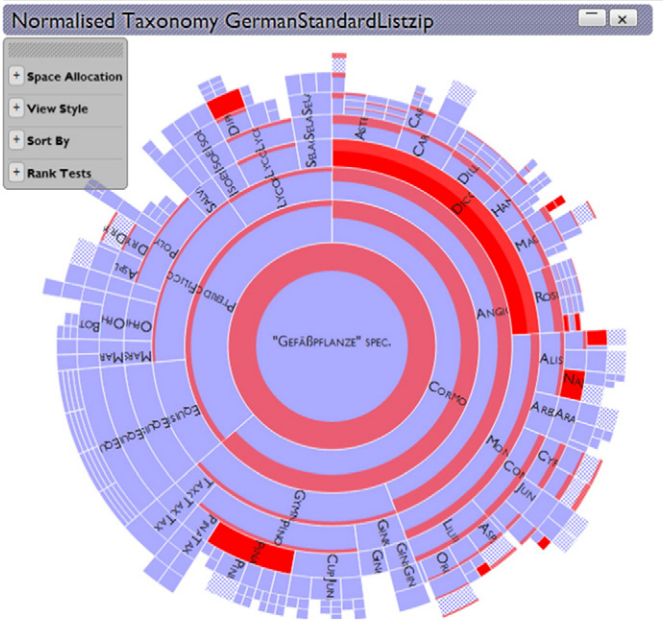


Fig. 6. The deeper red shows whether a taxon has been selected directly. The less saturated red indicates the proportion of subtaxa that have been selected.

chosen aggregate into smaller constituent parts—the taxonomic view will move deeper into the hierarchy, the geographic view will split a cluster into smaller sub-clusters, and the temporal view will expand the chosen bar into a set of bars at a higher level of temporal resolution. A right-button mouse-click will select all the records covered by the chosen aggregate, whether it's a taxon in the taxonomic view, a cluster in the geographic view, or a time period (represented by a bar) in the temporal view.

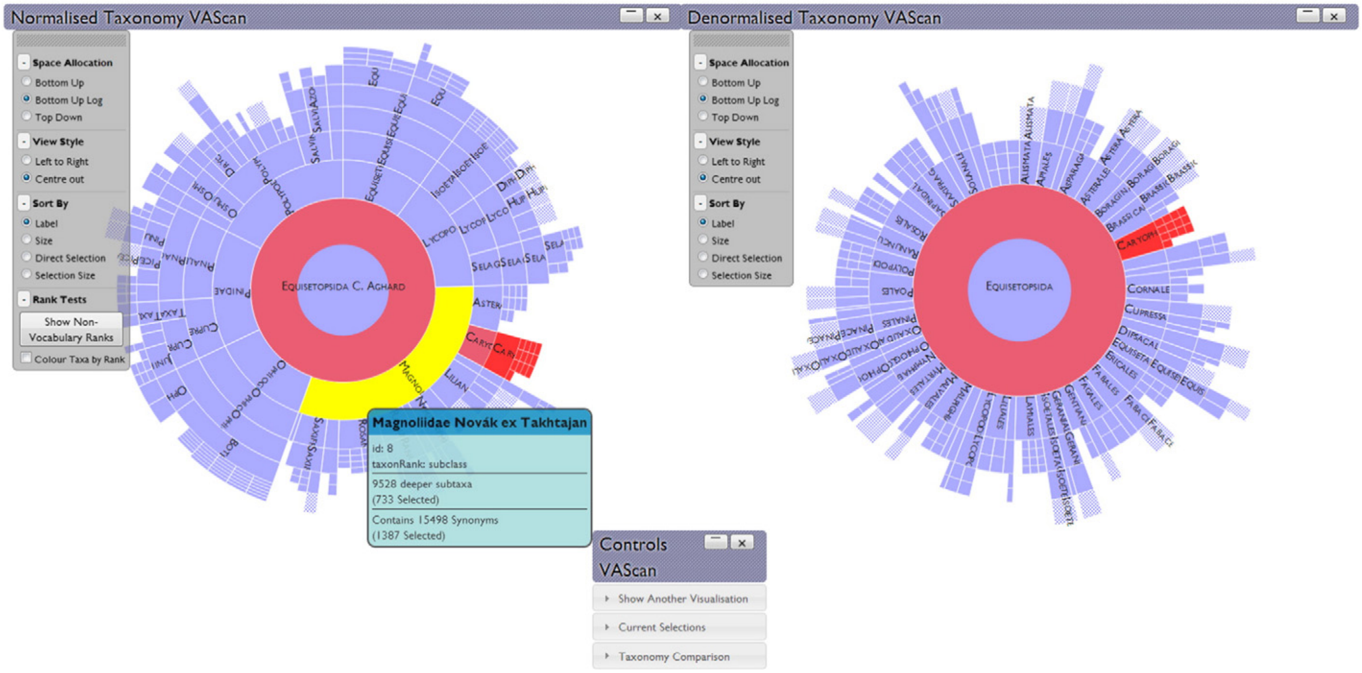
Further, there is a consistent principle of sizing the aggregate representations (taxa, map clusters, bars) in the visualisations by the natural logarithm of the number of records each aggregate represents. This is because trying to visualise data counts on a linear scale, whether taxonomic, geographic or temporal, almost immediately leads to situations where smaller aggregates shrink to the point of invisibility if the larger aggregates are shown at a reasonable scale, or if the scaling is such so the smaller aggregates are made visible, the larger aggregates become too large to display. The choice of a logarithmic scale allows large aggregates to be visibly more prominent than their smaller counterparts, but at the same time allows smaller aggregates a perceivable visual presence. This is of particular relevance in a visualisation designed to show data quality as extremely large aggregates or, at the other end of the size spectrum, individual outlying records are often a sign of compromised data quality.

Finally, Vesper operates as a co-ordinated multiple view system, so data points selected in one visualisation are also highlighted in the other visualisations. This enables questions such as "are the earliest specimens more likely not be geo-referenced?" or "are the specimens not allocated to a taxon lacking temporal and geographic data too?" to be asked. Selections may be inverted to negate situations, such as selecting the topmost map cluster and then inverting it to show the taxonomic and temporal distribution of non-georeferenced data. Selections can be saved as lists of taxon or occurrence IDs to an HTML file if the browser supports the latest HTML5 FileWriter API—at the time of writing this is limited to newer versions of Chrome and Opera (Deveria, 2014). Selections that are composed of suspicious or questionable quality data points can thus be exported for use in other tools.

2.4. Individual views

2.4.1. Geographic view

As stated previously (Steiniger and Hunter, 2013), there is a wide selection of available mapping software and services for generating plots of geographic data. Thus, rather than build our own, we chose from this existing range of alternatives and settled on the Leaflet.js library. This choice was due to its JavaScript codebase, active development community that has extended the base Leaflet functionality, and use and support of HTML5 functionality which means it is optimised for the same generation of browsers that are targeted by the D3 library.


Normalised Taxonomy VAScan

- Space Allocation
○ Bottom Up
● Bottom Up Log
○ Top Down

- View Style
○ Left to Right
● Centre out

- Sort By
● Label
○ Size
○ Direct Selection
○ Selection Size

- Rank Tests
Show Non-Vocabulary Ranks
□ Colour Taxa by Rank

Magnoliidae Novák ex Takhtajan
id: 8
taxonRank: subclass
9528 deeper subtaxa
(733 Selected)
Contains 15498 Synonyms
(1387 Selected)

Denormalised Taxonomy VAScan

- Space Allocation
○ Bottom Up
● Bottom Up Log
○ Top Down

- View Style
○ Left to Right
● Centre out

- Sort By
● Label
○ Size
○ Direct Selection
○ Selection Size

Controls VAScan
Show Another Visualisation
Current Selections
Taxonomy Comparison

Fig. 7. Some data sets contain information to construct both normalised and denormalised taxonomies, shown respectively to the left and right in this figure. The normalised taxonomies include synonymy and have a greater range of ranks. Here, in the VAScan archive, the normalised taxonomy has taxa at the subclass rank that the denormalised taxonomy does not.

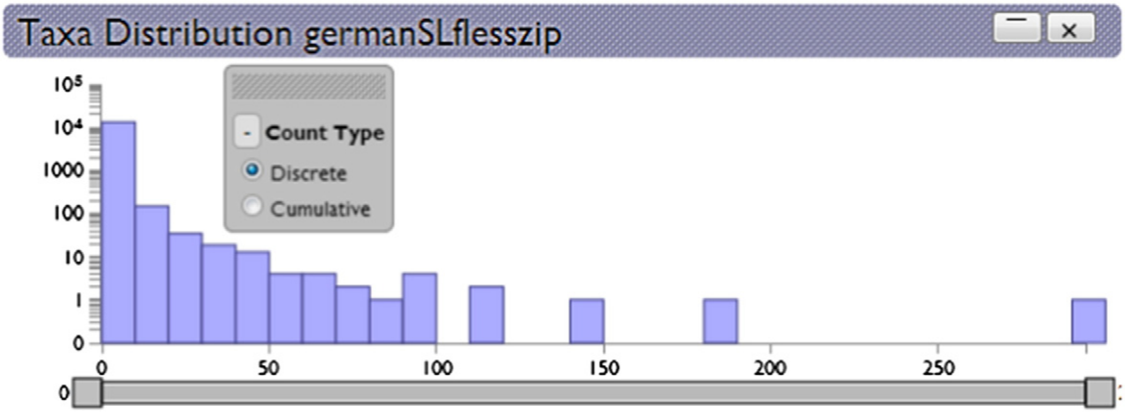


Fig. 8. Fan-out bar chart for a small reference taxonomy which follows the common pattern: most taxa have no (are leaves) or few sub-taxa; a few taxa have many sub-taxa.

When available, geographic (decimal longitude/latitude) data is plotted on the map via a clustering plugin, *leaflet-markercluster*, which produces zoomable clusters of data at interactive speeds, and has been extended to show selected record counts within each cluster. Another plugin, *leaflet-draw*, has been used to allow selection on the map by circle, rectangle or polygon if selecting by marker cluster is unsuitable in any situation. This functionality can also be used when a third plug-in, *leaflet-maskcanvas*, is activated to show the distribution as a more common dotplot across the map. Fig. 4 shows the difference between the clustering and dotplot plug-ins over the same data, with the left-hand side showing the clustered version. The mapping service used is OpenStreetMap as it offers free unrestricted use as long as they are credited within the application.

The clusters are sized logarithmically so that larger clusters appear bigger, but smaller clusters are not overwhelmed. Other visual map clustering techniques tend to use colour hue to indicate the size of a cluster, but perception research states whilst size is a good visual variable for conveying quantitative information, hue is less so (Garlandini and Fabrikant, 2009) e.g. is red 'bigger' than green? Also, we reserve colour to indicate selection totals within clusters once some records are selected.

2.4.2. Taxonomic views

2.4.2.1. Taxonomy visualisation. The taxonomy visualisation uses two standard space-filling techniques to visualise tree data, the icicle plot (Kruskal and Landwehr, 1983) and the sunburst (Andrews and Heidegger, 1998). These had to be adapted to cope with the large taxonomies that were sometimes encountered (the latest release of the ENA reference taxonomy has over 1,100,000 taxa). The naïve D3 layout algorithms allocated space for every single node in the taxonomy, even when most of them were calculated to take up a sub-pixel area. This slowed the rendering enormously, and the memory footprint of the extra fields and DOM objects produced for a million records often crashed the browser. Instead we adapted the layouts to stop calculating node positions when the size dipped below a threshold such as a given pixel height in the icicle plot, in effect giving them a horizon beyond which calculation and rendering was deemed pointless. Fig. 5 shows the sunburst version on the left and the icicle plot version on the right for the same taxonomy, the ENA reference taxonomy.

Visualising selections in the taxonomy visualisations is slightly more complicated than the approach for the geographic and temporal dimensions. Clusters in the map and timeline simply need to reflect the proportion of selected items in that aggregate; however the aggregates in the taxonomic views are considered first-order items in themselves i.e. a higher taxon is just as much a legitimate object as one of its sub-taxa, any may have a selected state even if none of its sub-taxa are. Thus each rendered taxon in the taxonomy view needs to show two selection states: 1) has it been selected itself and 2) how many of its sub-taxa are selected? Thus, we applied a visual scheme of colouring the entire node if directly selected, with a paler colour wash applied on top to indicate the proportion of selected sub-taxa, and an example of this can be seen in Fig. 6.

There are also two flavours of taxonomy dependent on whether it is generated from a set of parent-child ids, termed normalised taxonomies, or from collating fields of rank data in the file, these named denormalised taxonomies. As a rule of thumb, reference taxonomies tend to be stored as the former and species collections as the latter. This dichotomy also matches the DwC-A specification that says a core file can either be a taxonomy or an occurrence list (GBIF, 2011a), though

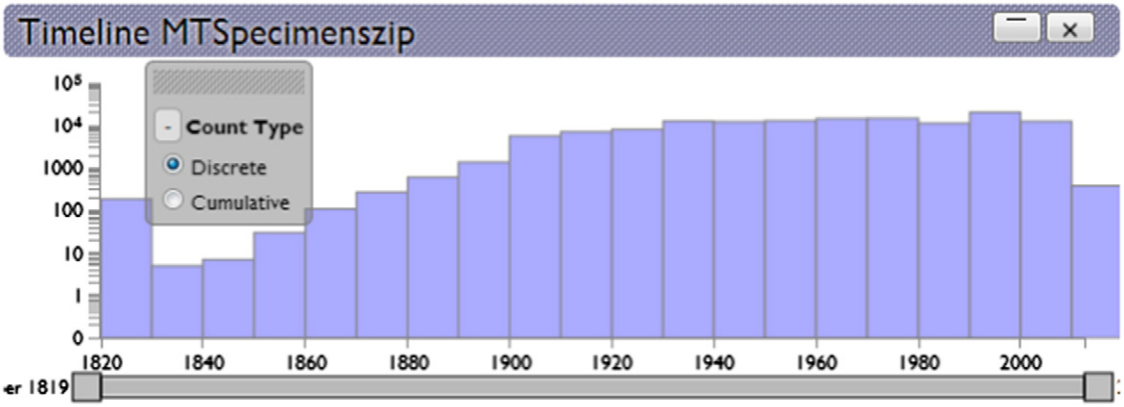


Fig. 9. Aggregated bar chart of specimen collection dates. This data set, MTSpecimens, covers almost two centuries.

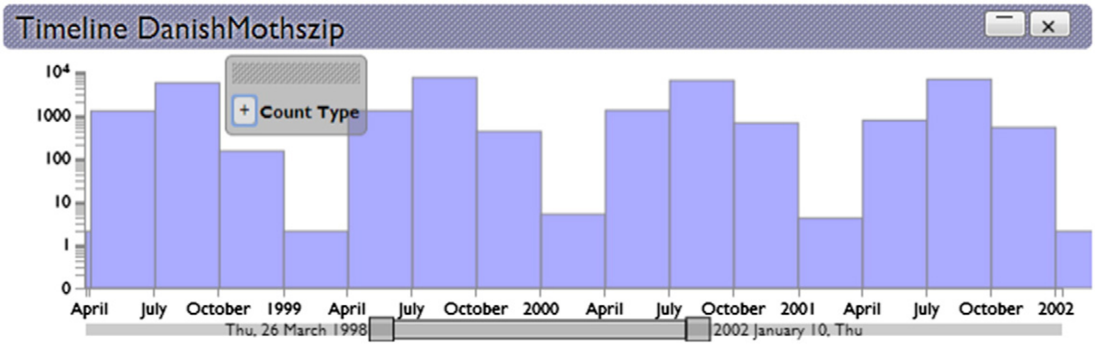


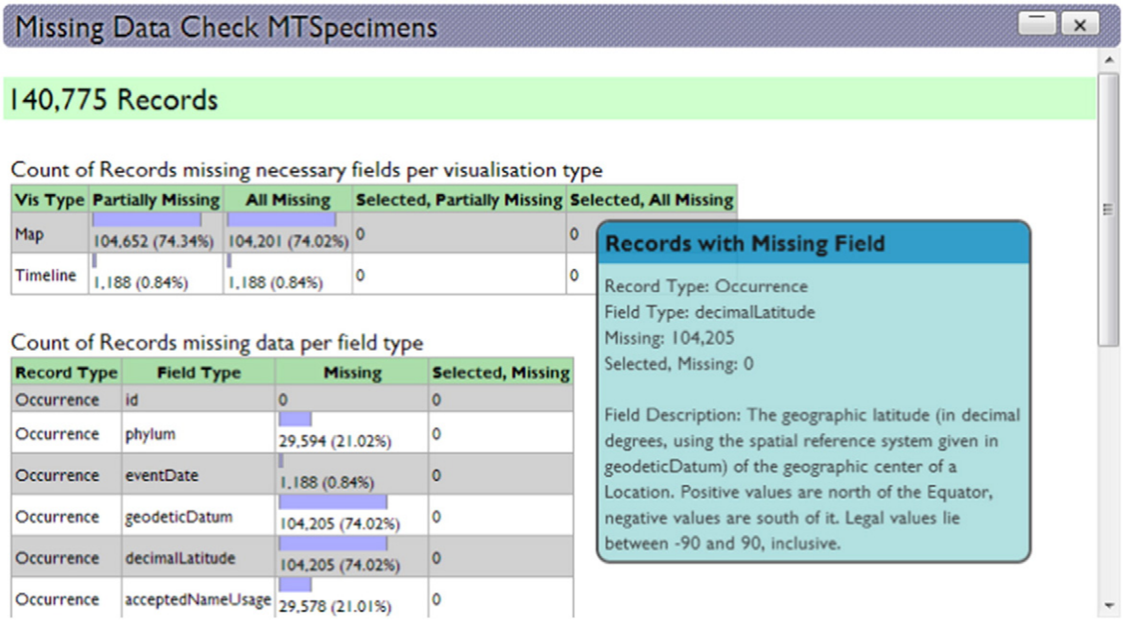
Fig. 10. The timeline can reveal a seasonal pattern of moth specimen collection in Denmark.

some archives allow the possibility of generating both types of taxonomy, as shown in Fig. 7.

Another decision made when constructing the visualisation was to not encode rank through fixed depths in the taxonomy, unlike previous taxonomy visualisations (Graham and Kennedy, 2007a; Spenke and Beilken, 2003). Instead, child taxa, whatever their rank, are shown immediately below their parent taxa. This is a deliberate decision as one sign of poor quality data is taxa with no rank assigned, or taxa that are impossibly related e.g. a genus that is a subtaxon of another genus. Data like these would break fixed rank visualisations. These are particular problems in the normalised taxonomies, where rank is a possibly uncontrolled field value, whereas the denormalised taxonomies are generated from data using a controlled set of known rank types such as Kingdom, Family, and Genus. There are two functions in the taxonomy view control panel to visualise the extent of rank standardisation in normalised taxonomies—firstly taxa with ranks that do not occur in the GBIF standard rank vocabulary (http://rs.gbif.org/vocabulary/gbif/rank.xml) can be selected, and secondly the taxa can be coloured by rank value. Both these functions can give insight into whether ranks are being used correctly or uniformly within a classification.

2.4.2.2. Taxonomic fan-out. Linnean taxonomies have been observed as following a *hollow curve* distribution (Chamberlin, 1924)—simply put most taxa tend to contain very few sub-taxa, often having only one child taxon, whilst a few taxa contain many sub-taxa, sometimes numbering well into the hundreds. Various reasons have been proposed for this distribution, grouped by two categories (Holman, 1985): either evolutionary—the effect of evolution on the organisms being classified—and psychological—the effect of the taxonomists doing the classifying. Though roughly a power law, the exact modelling is still an active research topic (Bokma et al., 2014) as they either under or over-estimate tail distributions.

As a rule of thumb, any taxonomy that strays far from the 'hollow curve' distribution described previously may have some underlying quality issues: either too many single-taxon units or unfeasibly large holding taxa that act as dumping grounds for various unclassified organisms. Whilst the taxonomic tree view can help reveal such patterns, often the extreme situations can overwhelm it: tree visualisations do not cope well with extremely large fan-outs; even research into alleviating this situation (Song et al., 2010) considers a large fan-out to be under a hundred sub-items. So, we introduce a taxonomic fan-out chart that aggregates the distribution of subtaxa per taxon as a bar chart. Fig. 8 shows an example of this pattern by visualising the fan out distribution for the plants in the German Standard List. The need for the logarithmic scale on the count axis becomes apparent as without it the larger (but less common) taxa wouldn't be visible.


Missing Data Check MTSpecimens

140,775 Records

Count of Records missing necessary fields per visualisation type

Vis Type	Partially Missing	All Missing	Selected, Partially Missing	Selected, All Missing
Map	104,652 (74.34%)	104,201 (74.02%)	0	0
Timeline	1,188 (0.84%)	1,188 (0.84%)	0	0

Count of Records missing data per field type

Record Type	Field Type	Missing	Selected, Missing
Occurrence	id	0	0
Occurrence	phylum	29,594 (21.02%)	0
Occurrence	eventDate	1,188 (0.84%)	0
Occurrence	geodeticDatum	104,205 (74.02%)	0
Occurrence	decimalLatitude	104,205 (74.02%)	0
Occurrence	acceptedNameUsage	29,578 (21.01%)	0

Records with Missing Field

Record Type: Occurrence
Field Type: decimalLatitude
Missing: 104,205
Selected, Missing: 0

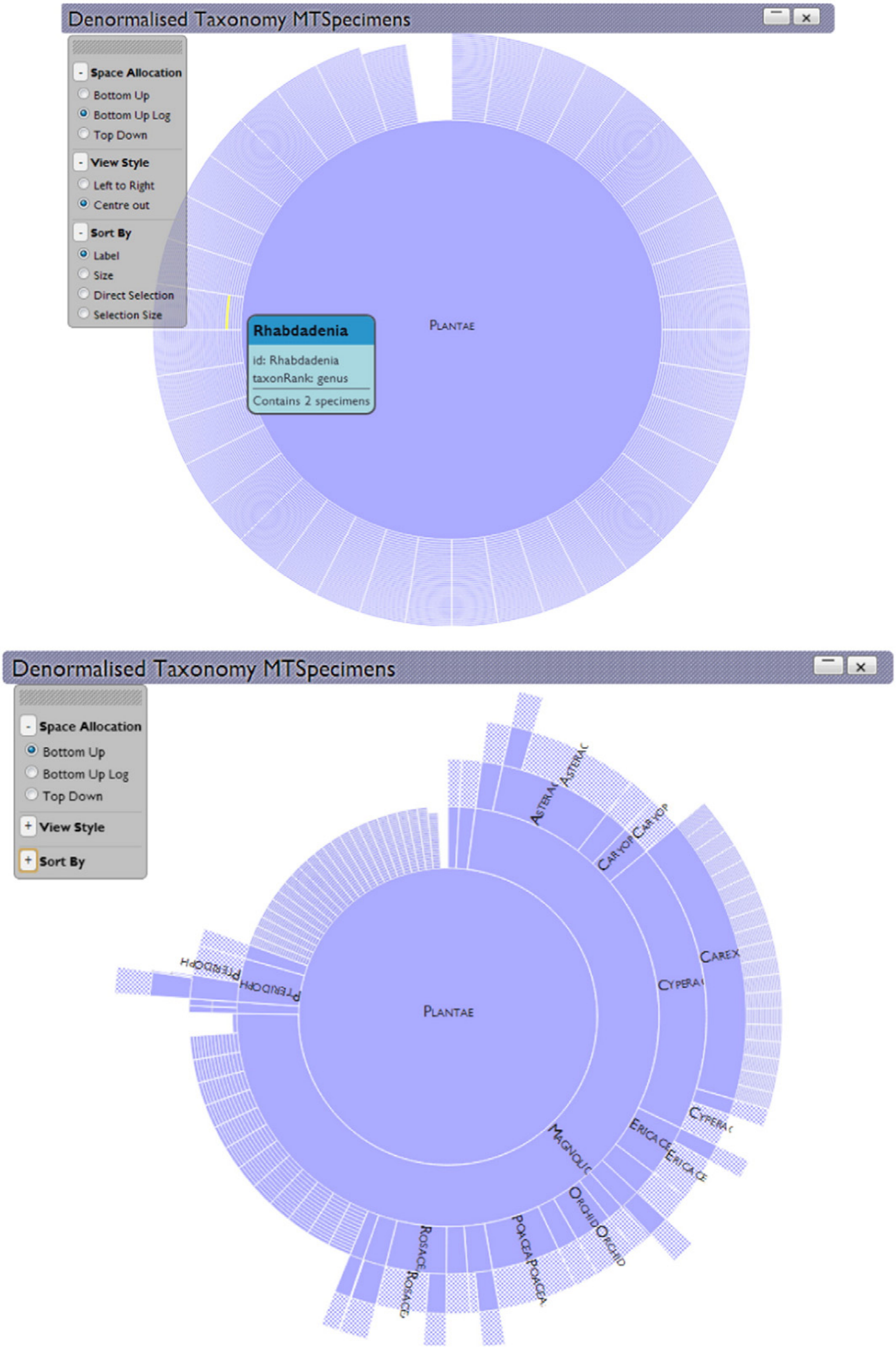
Field Description: The geographic latitude (in decimal degrees, using the spatial reference system given in geodeticDatum) of the geographic center of a Location. Positive values are north of the Equator, negative values are south of it. Legal values lie between -90 and 90, inclusive.

Fig. 11. Missing data summary for the MTSpecimens data set. Of 140,775 records, over 100,000 are missing geo-referenced data fields such as *decimalLatitude*. A much smaller proportion, just over 1000, are missing time data.

2.4.3. Temporal view

Temporal data associated with specimen records are shown using the same bar chart technique as the taxonomic fan-out visualisation. A bar chart is preferred to a line graph simply because bars are more visually prominent than a set of points joined by a line and because they are easier targets to move a pointer over for tooltip querying and other interactions. Further, bar charts display no interpolated values, unlike line graphs where there is a tendency to read meaning into the in-between values of the line between points, even where there isn't any.

Fig. 9 shows the aggregated timeline for the MT Specimens collection, revealing that the collection started slowly in the 1820s and only passed 1000 new specimens per decade by the turn of the 20th century.



- Space Allocation
○ Bottom Up
● Bottom Up Log
○ Top Down
- View Style
○ Left to Right
● Centre out
- Sort By
● Label
○ Size
○ Direct Selection
○ Selection Size

Rhabdadenia
id: Rhabdadenia
taxonRank: genus
Contains 2 specimens

- Space Allocation
● Bottom Up
○ Bottom Up Log
○ Top Down

Fig. 12. In a degenerate taxonomy, thousands of items are directly under the Plantae kingdom. This leads to obvious difficulties if all these immediate subtaxa are visualised together, as seen in the top half of the figure. But the non-logarithmic size distribution then used in the screenshot in the figure's bottom half shows that there is a recognisable taxonomy of some sorts in the data set; however, it achieves this by ignoring the majority of the extraneous data attached to the root.

However since then it has regularly amassed 10,000 new specimens or more per decade (the bar at the far right is smaller because we are only part way through the 2010–20 decade.)

Zooming in/drilling down is achieved either by clicking on an individual bar, which then expands into a new set of bars to fill the display, or by using the range slider underneath the bar chart. This range slider is also used to zoom back out and can be dragged to change the time period under consideration. Zooming in to a level where monthly aggregations are visible can often reveal seasonal patterns in data collection, as in Fig. 10, which shows Danish moth specimen collection (unsurprisingly) drops off in the winter.

2.5. Miscellaneous controls

Vesper offers a pair of simple search widgets for querying the data. The first, “Search”, does a partial string match of entered text against the chosen name field e.g. scientificName, and then selects every matching record. The second, “Record ID”, allows querying of a specific record by the coreID field and returns associated data such as parentage/synonymy in the taxonomy. A final option, shown in Fig. 11, gives a panel that details the number of missing/unparseable values in the fields returned by the initial parsing and how these affect the chosen visualisations which are often composed of data from several fields. These options are accessible from the control panel view for each visualisation, which can be seen in the bottom left of Fig. 3.

2.6. Customisation

Vesper uses an open-source JavaScript internationalisation framework (i18next) and thus can be localised to display instructions/details in languages other than English. An English language example of the templates can be found in the src/locales folder of the Vesper source code, though this is aimed more at developers than typical data users. Also, some properties of the visualisations such as colouring and border styling are adjustable by editing the CSS classes in the Vesper source, an advantage of using the CSS/SVG/HTML triumvirate championed by the D3 library.

3. Discussion

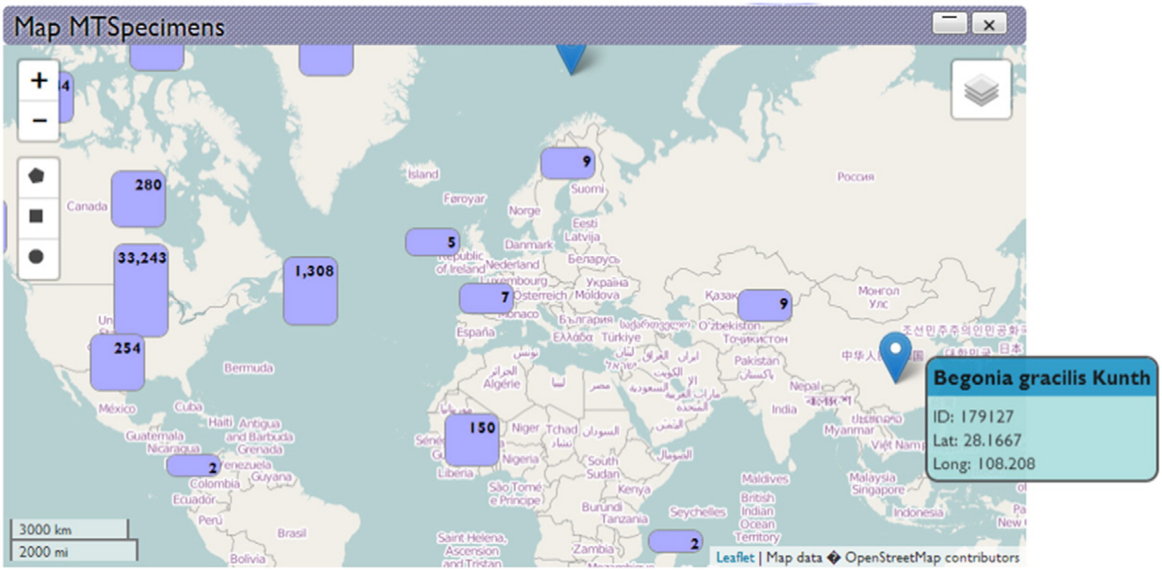
The advantages of Vesper for analysing DwC-A datasets are demonstrated by describing a number of scenarios with real-world data sets. The first scenario describes exploring a specimen collection with spatial and time-based data, and what issues are revealed by Vesper's visualisations and multiple view capabilities. The second scenario describes exploring a large reference taxonomy and investigating some of the more extreme outliers in the taxonomy fan-out chart. This evidences that it can show data quality issues within a large taxonomy. The final scenario shows how multiple archives can be loaded and the taxonomies compared to give a quick idea of coverage between a specimen collection and a reference taxonomy.

3.1. Example scenario 1—MTSpecimens

The MTSpecimens dataset is a 140,000 + specimen-based collection, mostly collected from North America, especially Canada. The DwC-A version explored here is from the 19th April 2012, although there is now a revised version available (autumn 2013) with more data. Selecting the data set from the examples and then selecting the three available options (denormalised taxonomy, map and timeline) parse the DwC-A file for the necessary fields per record. The taxonomy is produced by aggregating the denormalised taxonomic data in the archive, and, after parsing has finished, produces what appears to be a singularly unrewarding visualisation, as shown in the top half of Fig. 12.

A quick investigation with the mouse tooltip reveals thousands of taxa gathered directly under Plantae, most of them at the genus level with no intervening family or order ranks. As discussed previously, tree visualisations have not been designed to show set sizes that would overwhelm most list visualisations, and whilst the logarithmic scaling means we can see every single item under Plantae none of them are shown at a usable level of detail. This can be alleviated by switching to a non-logarithmic view which reveals a taxonomy previously hidden by the thousands of spuriously attached taxa at the root—see the screenshot in the bottom half of Fig. 12. This non-logarithmic view doesn't give enough space for the smaller taxa to be plotted; in fact looking at this view gives the impression they are not there and thus not a problem, but by starting with the logarithmic-based layout we immediately have evidence of a quality issue in the data.

This data set contains a large amount of geo-referenced data which can be visualised using the map view, and Fig. 13 shows how the clustering appears at the top level for the MTSpecimens data set. What is apparent is that the bulk of the data is in North America, with smaller clusters elsewhere. One individual data point is sitting in China, and using a tooltip to query the marker reveals it to be a specimen of *Begonia*



Begonia gracilis Kunth
ID: 179127
Lat: 28.1667
Long: 108.208

Fig. 13. Roughly 36,000 records have geographic coordinates in the MTSpecimens data set. The outlier in China is apparently caused by a missing sign in the longitude data for that record.

gracilis—a plant widespread in the mountains of Mexico. This appears to be an instance of a longitude inversion, which is commonly revealed by geovisualisation techniques. What is less apparent, but is revealed by the missing value summary described earlier, is that 36,000 geo-referenced records represent only a quarter of the 140,000 records in the archive. In some cases, a visualisation may manage with partial data: taxonomies often legitimately omit ranks even within the same dataset, whereas any geo-location data that is missing longitude or latitude information is essentially broken as the best it can then generate is a position along the equator or Greenwich meridian.

Further, we can explore the cross-sections of the different views of the data. Exploring the timeline visualisation reveals the beginning of this collection to have started in the 1820–30 decade with 186 records. Selecting this first bar highlights it in red and zooms the map view in to the highest magnification that contains selected, geo-referenced records, as seen in Fig. 14. The map now shows 88 geo-referenced records out of 186 (a higher geo-referenced proportion than the collection overall), though when the larger sub-cluster is drilled into it is soon revealed that they, as part of a group of 2815 records, are all recorded as being collected at the same exact point. This geographic data is obviously an approximation of some sort.

Selecting this suspicious cluster in the map then highlights when this entire cluster of specimens were recorded, and reveals that they are distributed across most of the dataset's time span but with a peak in the 1930s, as shown in Fig. 15. In conclusion, we can see that of the 36,000 geo-referenced specimens in this data set, nearly 3000 of those are recorded as being collected at the same exact point, indicating perhaps a historical data quality issue.

3.2. Example scenario 2—The ENA Reference Taxonomy

The ENA (European Nucleotide Archive) Reference Taxonomy is a classification built by EMBL-EBI to support querying and browsing of their large nucleotide collection. The latest release in DwC-A format (9th April, 2014) has over 1.1 million records, and is thus a taxonomy composed of a similar number of taxa.

Loading this data set gives us a single option of viewing the normalised taxonomy. Once this is selected, the DwC-A is parsed (which takes a few seconds, but a progress bar indicates the current parsing status). The visualisations launched are the taxonomy visualisation itself and the taxonomy fan-out chart, and it is the latter which is of interest here. Looking at the fan-out chart quickly enables the identification of

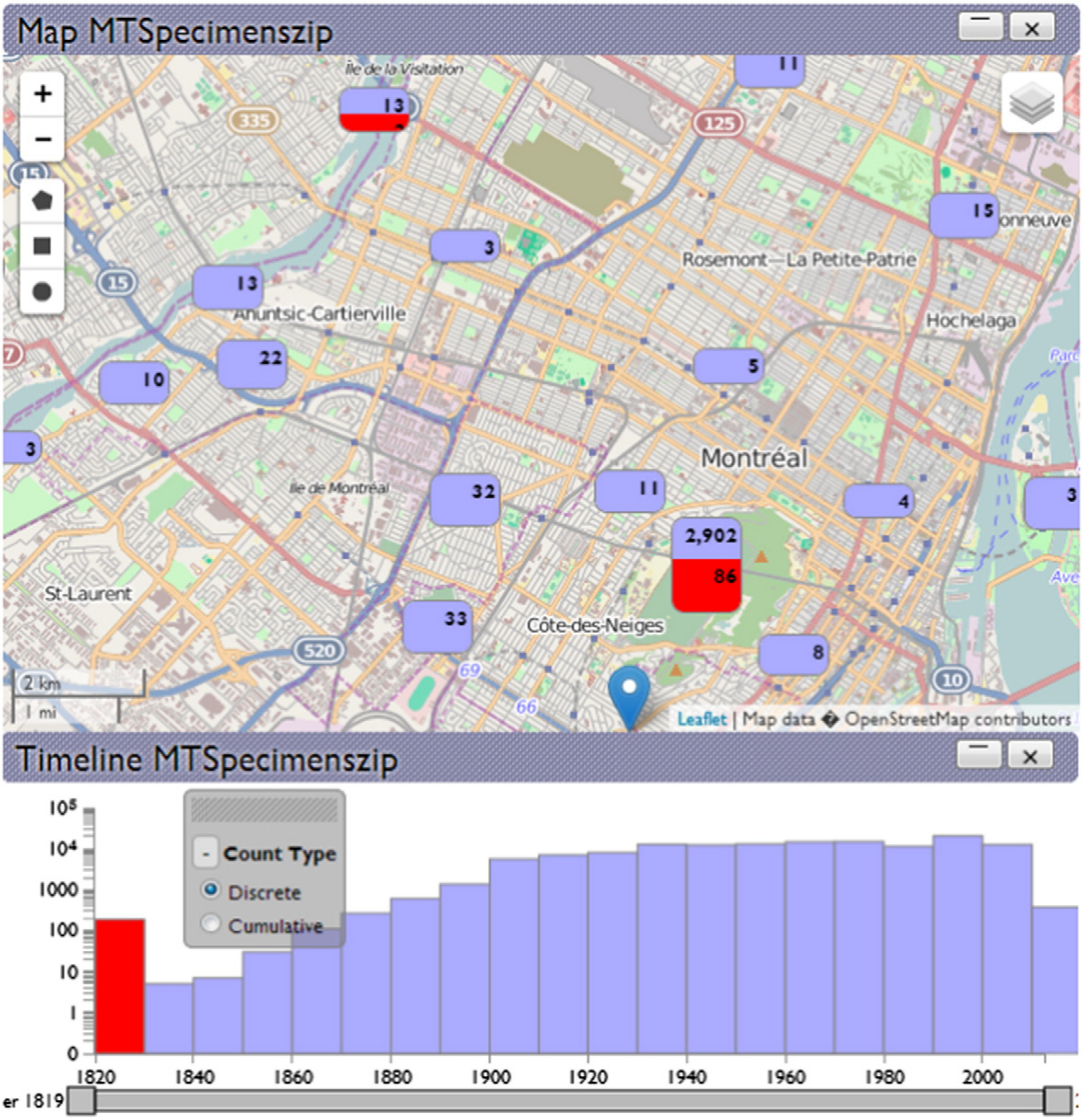
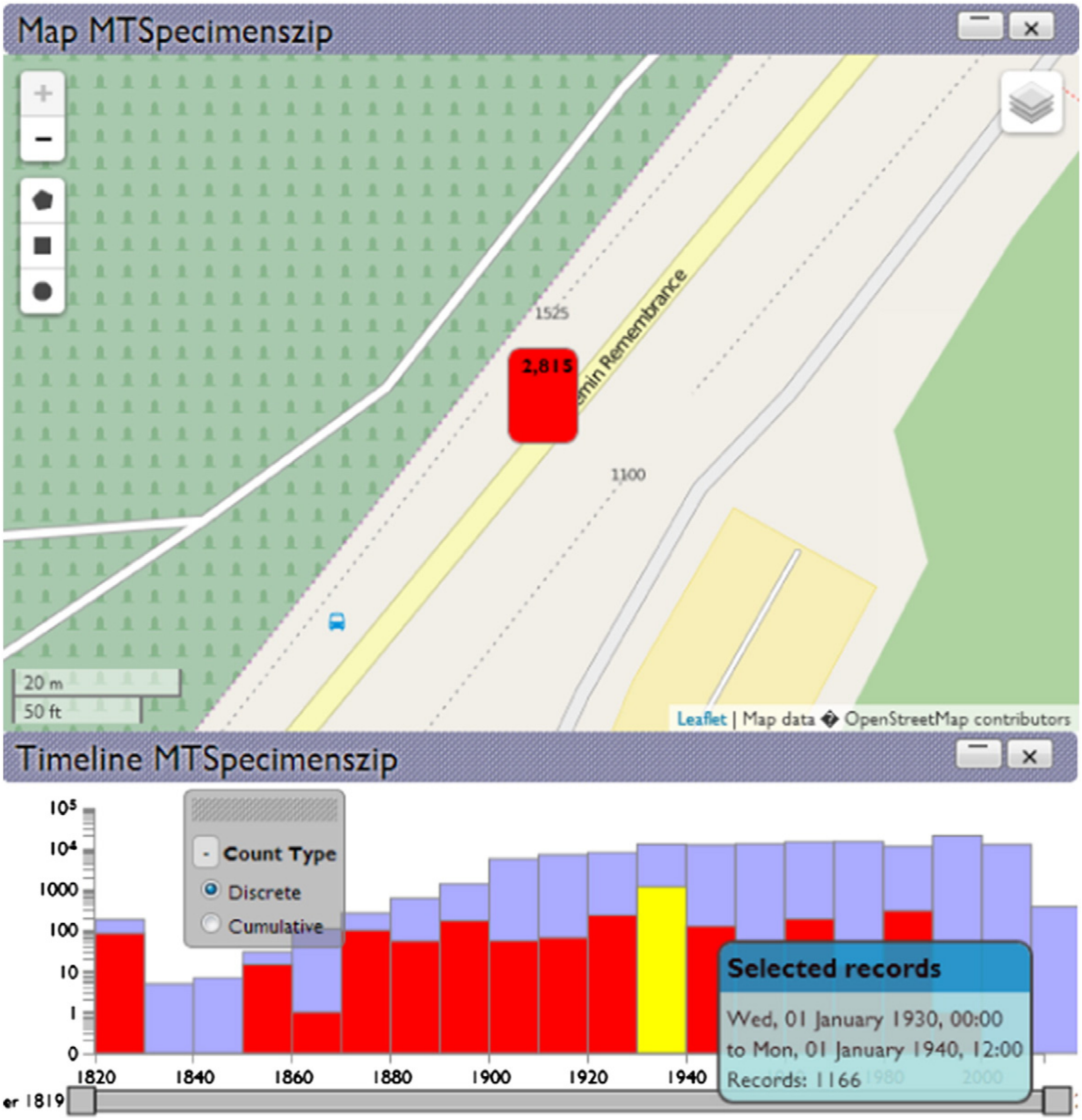


Fig. 14. Plotting the distribution of the earliest collection specimens with known longitude and latitude data reveals them to be in central Montreal.



Map MTSpecimenszip

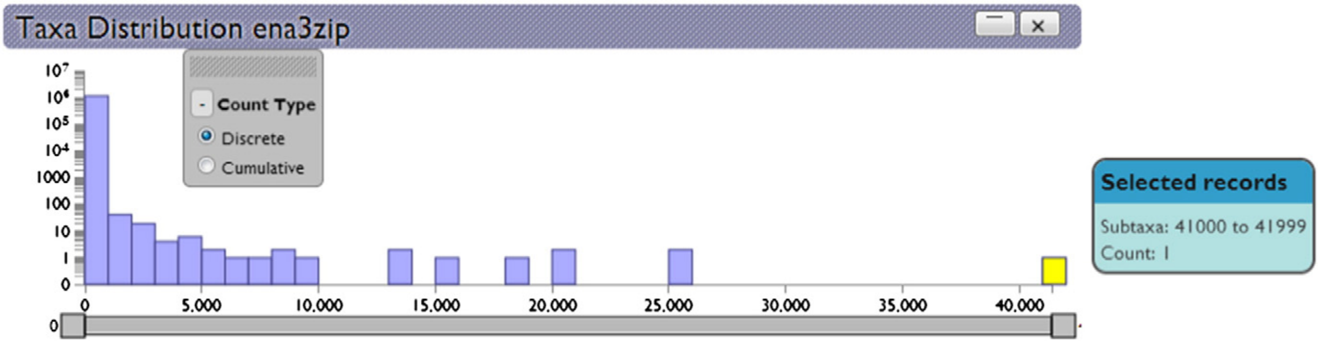
Timeline MTSpecimenszip

Selected records
Wed, 01 January 1930, 00:00
to Mon, 01 January 1940, 12:00
Records: 1166

Fig. 15. Temporal distribution of specimens recorded as being collected at one specific spatial point.

possible rogue taxa within the classification, as demonstrated in Fig. 16 where several taxa are reporting over 10,000 subtaxa each in the ENA reference taxonomy.

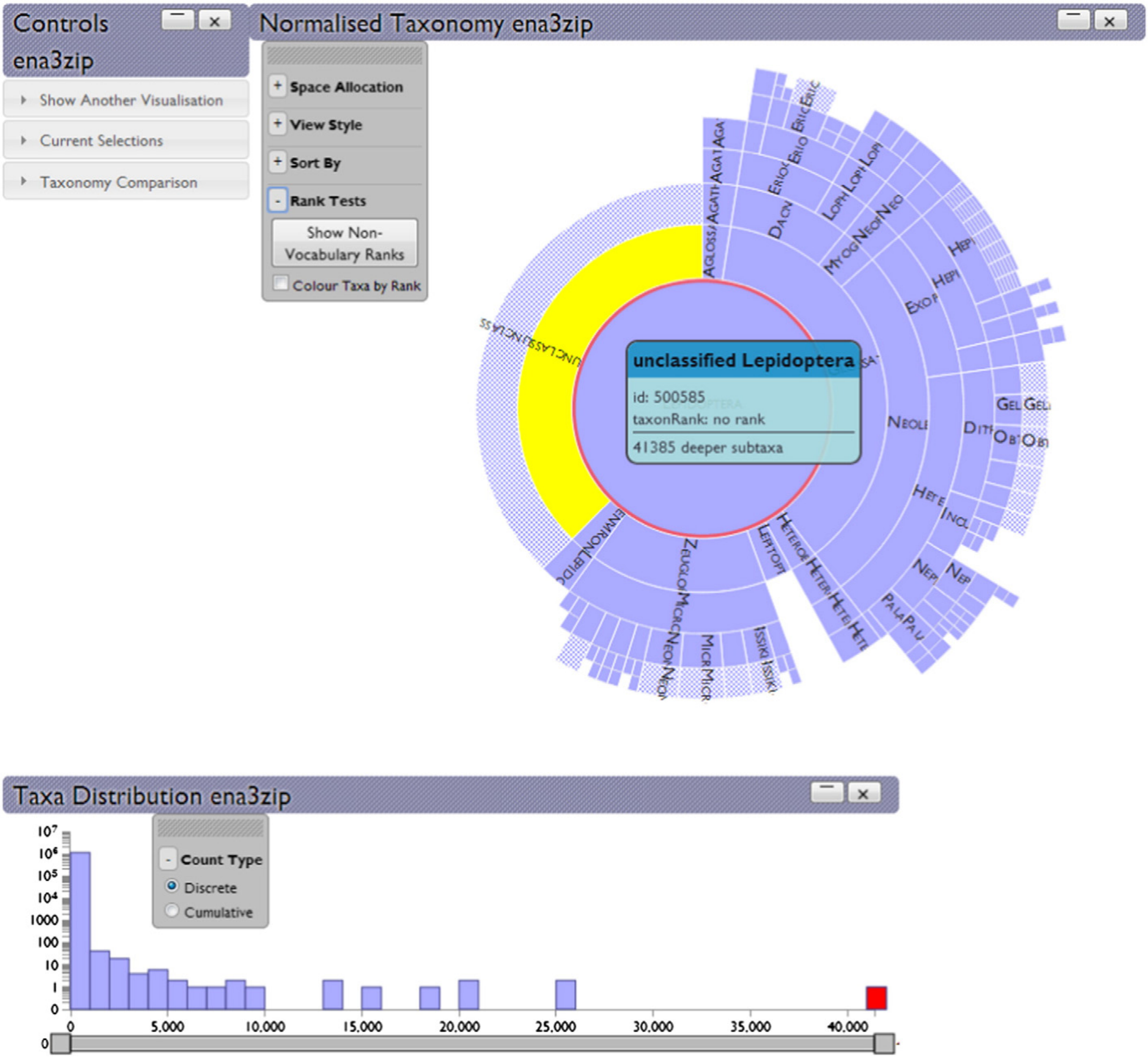
Selection of the most extreme example in this bar chart reorients the current root node in the taxonomic tree view to show the offending taxa, as shown in Fig. 17. Here, it turns out to be *unclassified Lepidoptera* with no assigned taxonomic rank, a clear indication the taxonomy still has some work pending, as just this one taxa represents almost 4% of the records within the million-plus sized taxonomy. Similar investigation of the remaining extreme taxa finds, amongst other things: various



Taxa Distribution ena3zip

Selected records
Subtaxa: 41000 to 41999
Count: 1

Fig. 16. Identifying extreme sized taxa within an archive is best approached through the taxa distribution view.



unclassified Lepidoptera
id: 500585
taxonRank: no rank
41385 deeper subtaxa

Count Type: ● Discrete ○ Cumulative

Fig. 17. Selecting the bar in the chart reveals the outlier to be "unclassified Lepidoptera"—a dumping ground within the taxonomy for as-yet unclassified specimens.

unranked Influenza taxa, unclassified Bacteria and 15,000 species of *Pseudomonas* collected together under the genus of that name.

As a follow-up, we launch the Search widget (and switch off the keystroke level search which would be unresponsive on this size of taxonomy), and search for any other taxa containing the string "unclassified". The visualisation returns as shown in Fig. 18, showing as well as the Lepidoptera there are many other taxa containing that particular string.

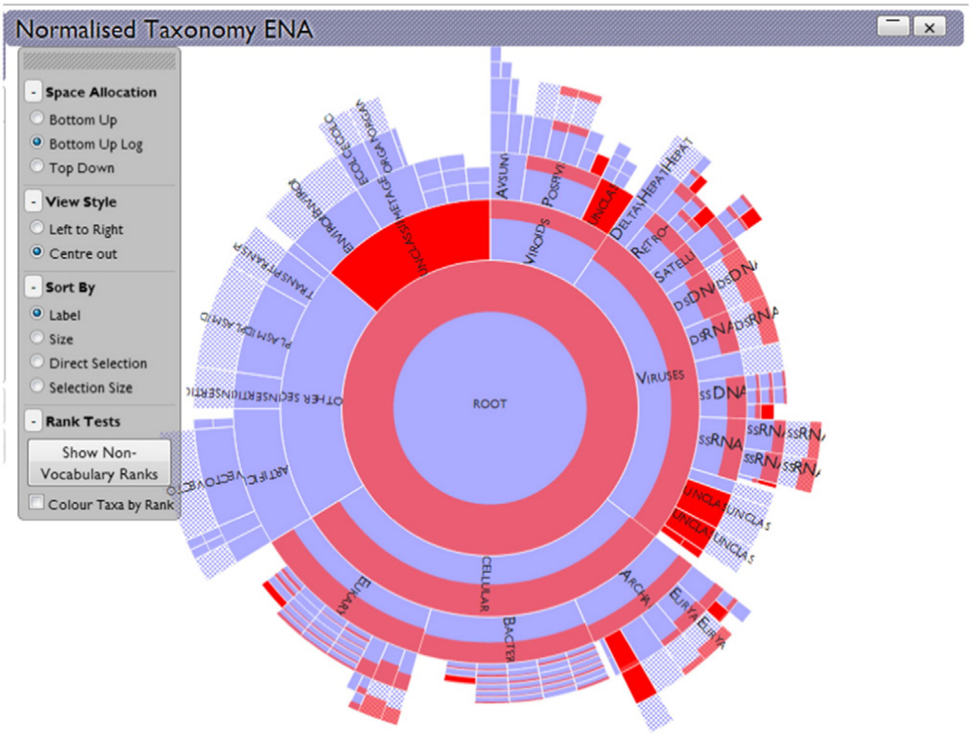
Similarly, selecting the "Show Non-Vocabulary Ranks" button in the taxonomy controls can show how many taxa have non-standard or non-existent rank information, and where they are concentrated in the taxonomy. In the ENA taxonomy's case this reveals over 100,000 taxa with no or non-standard rank labels; however, the vast majority of these are concentrated under "Viruses"; of the near 900,000 taxa under cellular organisms, only 12,000 (roughly 1%) are lacking rank information, as shown in Fig. 19.

In summary, this scenario shows that Vesper can quickly reveal data quality issues in large taxonomies of over a million taxa.

3.3. Example scenario 3—comparison to reference taxonomy

One of the features of Vesper is that different DwC-A files can be opened and visualised simultaneously, permitting the development of functionality that would compare the taxonomies of two different archives. In this scenario, a species collection of ladybirds with 235,000 records is compared against a small reference taxonomy for ladybirds of 600 taxa (part of the Catalogue of Life mega-taxonomy). Both archives are loaded in, taking particular care to use the same field as a label for each taxonomy (e.g. scientificName), as this is what the comparison keys on. In each archive's control panel view there is a section for taxonomy comparison. Selecting the "Compare names" option for one archive and then selecting the same option for another open archive will compare the two name sets, with the overlap shown using the same technique as other selections, as seen in Fig. 20.

The comparison is not nearly as accurate as concept-based approaches as name matching algorithms have many acknowledged

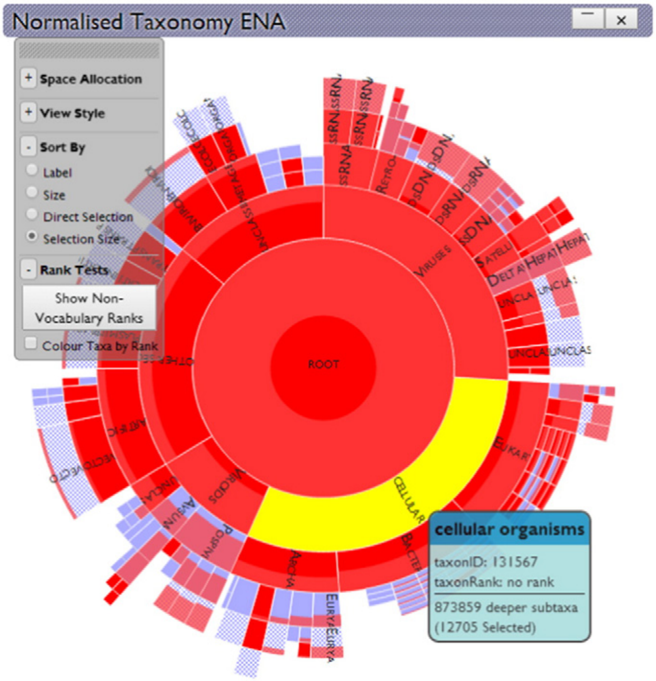


Normalised Taxonomy ENA

- Space Allocation
○ Bottom Up
● Bottom Up Log
○ Top Down
- View Style
○ Left to Right
● Centre out
- Sort By
● Label
○ Size
○ Direct Selection
○ Selection Size
- Rank Tests
Show Non-Vocabulary Ranks
□ Colour Taxa by Rank

Fig. 18. Searching for the string “unclassified” in taxa names reveals many such instances.

drawbacks (Kennedy et al., 2005): beyond simple spelling differences there are issues with renaming and repositioning of taxa over time that changes names between different classifications in a much more fundamental manner. However, the function here gives an intuitive feel of the overlap between two taxonomies, and does compare stems of names rather than every character (as in the scientificName field some archives add the authors after the name, some don't.) This is thus obviously an over-simplification of the situation but it is not intended to compete with taxonomic name reconciliation services, and is also necessarily limited name-wise by the computational power and time that would be needed to compare sets of names and concept-wise by the lack of concept information in most DwC-A files. In this sense, the aim is rather to give a quick view of the possible overlap between classifications.



Normalised Taxonomy ENA

+ Space Allocation
+ View Style
- Sort By
○ Label
○ Size
○ Direct Selection
● Selection Size
- Rank Tests
Show Non-Vocabulary Ranks
□ Colour Taxa by Rank

cellular organisms
taxonID: 131567
taxonRank: no rank
873859 deeper subtaxa
(12705 Selected)

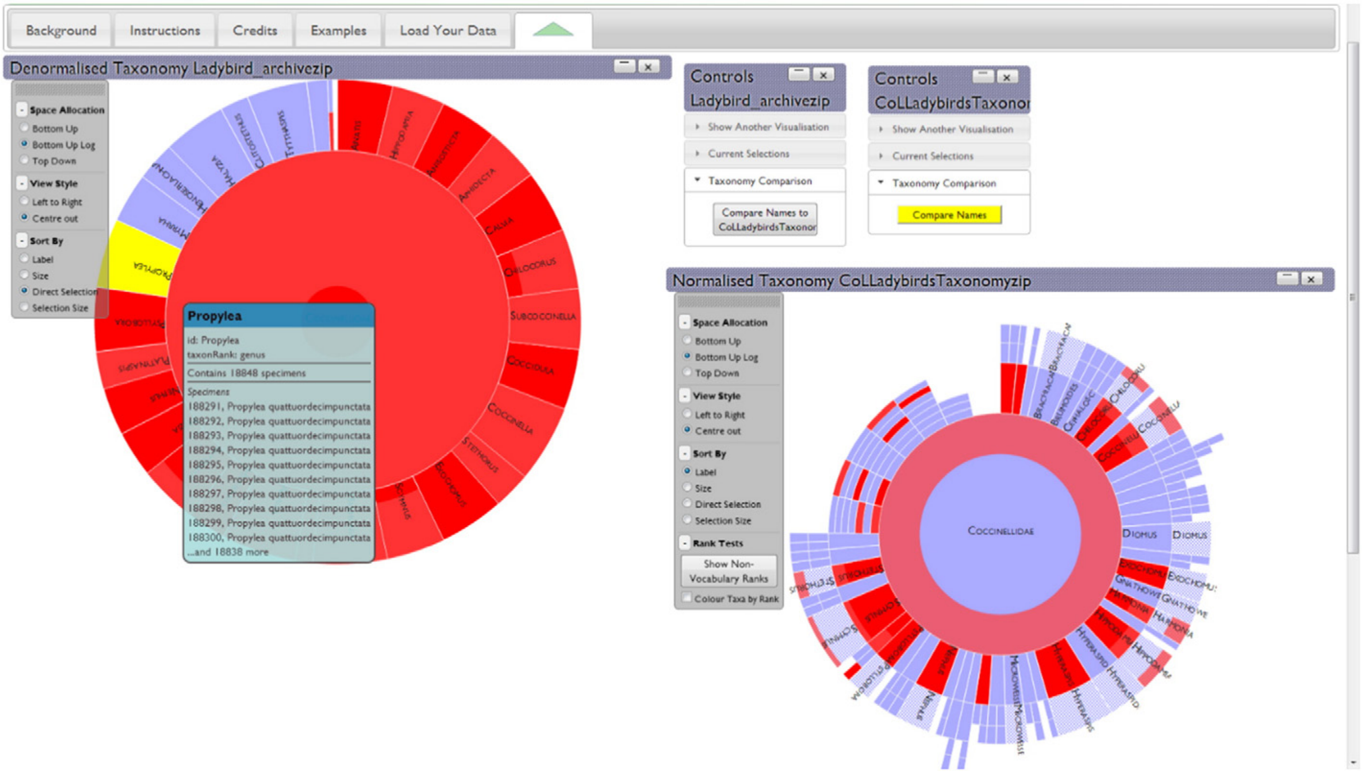
Fig. 19. Non-standard ranks within the ENA reference taxonomy are highlighted in red. Ordering the sunburst clockwise by selection size reveals most occur under “Viroids” and “cellular organisms” are relatively well standardised.

4. Conclusion

The visualisations in Vesper help achieve some of the recommendations put forward in (Faith et al., 2013) as responses to user concerns about existing biodiversity data. These specific recommendations are re-capped below:

- *Recommendations relating to data gaps, data volume, and data quality*
  6. *Initiate the following steps to enhance the trust-worthiness of GBIF mobilised data:*
     g. *Improve pathways for data publishers to provide warnings about biases or errors in the data at an early stage of discovery and publishing process.*
     j. *Expedite efforts in improving taxonomic and geo-spatial quality of GBIF mobilised data. This task includes attention to geo-referencing.*
     k. *Improve fitness-for-use of data at the data producer and/or primary publisher stage.*

By allowing users to perceive data issues such as taxonomic or geographic data “dumping grounds” Vesper is a tool that acts towards the fulfilment of point (j) above. By operating on data that is in DwC-A format Vesper can provide feedback on the quality of data before it is published, fulfilling point (g). Vesper acts as a detector for finding questionable data points, but does not provide mechanisms to then correct erroneous data. However, the ability to export selected taxon or



Propylea
id: Propylea
taxonRank: genus
Contains 18848 specimens
Specimens
188291, Propylea quattuordecimpunctata
188292, Propylea quattuordecimpunctata
188293, Propylea quattuordecimpunctata
188294, Propylea quattuordecimpunctata
188295, Propylea quattuordecimpunctata
188296, Propylea quattuordecimpunctata
188297, Propylea quattuordecimpunctata
188298, Propylea quattuordecimpunctata
188299, Propylea quattuordecimpunctata
188300, Propylea quattuordecimpunctata
...and 18838 more

Fig. 20. Comparing the reference taxonomy (bottom right) against the species-based denormalised taxonomy (top-left). Matching sections are shown in red; thus, the species taxonomy uses names that in the majority have correlates in the reference taxonomy. The smaller coverage of red in the reference taxonomy shows the species collected do not cover the full range of possible ladybird species. Propylea in the species taxonomy has no matches; it is spelled Propylaea in the reference taxonomy.

occurrence IDs in certain browsers means that it can begin a process that leads to other tools fulfilling point (k) above. Even further, as well as a final quality control step before publication, Vesper can act as a quality control during the process of dataset assembly. Generating DwC-A's is supported by the GBIF Integrated Publishing Toolkit and these can be generated at any time in the construction of a data set, so different DwC-A's of the same data set at different points in its construction could be used to show and guide progress towards data quality targets.

By operating entirely within an in-memory client-side browser environment there are obvious limits to VESPER's scalability. It will not take in the entire GBIF dataset as a DwC-A dataset, nor operate over archives containing many millions of records. However, it can take in respectably-sized archives, aided by both the memory efficiencies of partial zip decompression routines and the horizon-limited layout algorithms that stop the generation of unused layout data and DOM objects. The end result is that on a 32-bit 2GB Windows PC, Vesper can take in reference taxonomies of up to a million records, and specimen datasets with full taxonomic, geographic and temporal data fields of over 150,000 records.

Vesper thus supplies an extra option in the process of improving data quality within biodiversity data archives and at the crucial stage of operating before such data is propagated to publishers and then onto unsuspecting users. At the same time, users themselves can use the tool to inspect data archives and reassure themselves that the data contained is fit for purpose (or reveal it isn't, as the case may be).

Possible future work for Vesper may include improving the zip decompressing (the original library used has evolved recently, and the newer version is not currently used in Vesper) and in increasing the range of visualisation types. A grid map to count geospatial distributions is perfectly feasible, and a more visual representation of missing data beyond a table would be an interesting addition.

Acknowledgements

We are grateful to the UK BBSRC (Biotechnology and Biological Sciences Research Council) for funding this work (grant award no.: BB/K004115/1), and also to GBIF, the EMBL European Bioinformatics Institute (EBI) and Candensys for sharing their DwC-A datasets, though by now they have ironed out many of the issues that made them interesting for Vesper to explore.

The Vesper source code is available for download at http://github.com/martingraham/vesper under an ApacheV2 licence with a demonstration version sitting at http://www.vesper.org.uk/vesperDemo/vesper/demoNew.html. All the libraries used by Vesper are also MIT or ApacheV2 licenced software.

References

Aigner, W., Miksch, S., Müller, W., Schumann, H., Tominski, C., 2007. Visualizing time-oriented data—a systematic view. Comput. Graph. 31, 401–409. http://dx.doi.org/10.1016/j.cag.2007.01.030.

Andrews, K., Heidegger, H., 1998. Information Slices: Visualising and Exploring Large Hierarchies using Cascading, Semi-Circular Discs, IEEE Symposium on Information Visualization, Late Breaking Hot Topics. IEEE Computer Society Press, Research Triangle, North Carolina, USA pp. 9–12.

Andrienko, G., Andrienko, N., Demsar, U., Dransch, D., Dykes, J., Fabrikant, S.I., Jern, M., Kraak, M.-J., Schumann, H., Tominski, C., 2010. Space, time, and visual analytics. Int. J. Geogr. Inf. Sci. 24, 1577–1600. http://dx.doi.org/10.1080/13658816.2010.508043.

Ariño, A.H., Chavan, V., Faith, D.P., 2013. Assessment of user needs of primary biodiversity data: analysis, concerns, and challenges. Biodivers. Inform. 8, 59–93.

Auer, T., MacEachren, A.M., McCabe, C., Pezanowski, S., Stryker, M., 2011. HerbariaViz: A web-based client–server interface for mapping and exploring flora observation data. Ecol. Inform. 6, 93–110. http://dx.doi.org/10.1016/j.ecoinf.2010.09.001.

Block, F., Horn, M.S., Phillips, B.C., Diamond, J., Evans, E.M., Shen, C., 2012. The DeepTree Exhibit: visualizing the tree of life to facilitate informal learning. IEEE Trans. Vis. Comput. Graph. 18, 2789–2798. http://dx.doi.org/10.1109/tvcg.2012.272.

Bokma, F., Baek, S.K., Minnhagen, P., 2014. 50 Years of Inordinate Fondness. Syst. Biol. 63, 251–256. http://dx.doi.org/10.1093/sysbio/syt067.

Bostock, M., Ogievetsky, V., Heer, J., 2011. D^3: data-driven documents. IEEE Trans. Vis. Comput. Graph. 17, 2301–2309. http://dx.doi.org/10.1109/TVCG.2011.185.

Byron, L., Wattenberg, M., 2008. Stacked graphs—geometry & aesthetics. IEEE Trans. Vis. Comput. Graph. 14, 1245–1252. http://dx.doi.org/10.1109/TVCG.2008.166.

Chamberlin, J.C., 1924. Concerning the hollow curve of distribution. Am. Nat. 58, 350–374. http://dx.doi.org/10.2307/2456484.

Deveria, A., 2014. Can I use. http://caniuse.com (accessed 1st April 2014).

Faith, D., Collen, B., Ariño, A., Koleff, P., Guinotte, J., Kerr, J., Chavan, V., 2013. Bridging the biodiversity data gaps: recommendations to meet users' data needs. Biodivers. Inform. 8, 41–58.

Fegraus, E.H., Andelman, S., Jones, M.B., Schildhauer, M., 2005. Maximizing the value of ecological data with structured metadata: an introduction to ecological metadata language (EML) and principles for metadata creation. Bull. Ecol. Soc. Am. 86, 158–168. http://dx.doi.org/10.1890/0012-9623(2005)86[158:MTVOED]2.0.CO;2.

Ferreira, N., Lins, L., Fink, D., Kelling, S., Wood, C., Freire, J., Silva, C., 2011. BirdVis: visualizing and understanding bird populations. IEEE Trans. Vis. Comput. Graph. 17, 2374–2383. http://dx.doi.org/10.1109/TVCG.2011.176.

Fyfe, D.A., Holdsworth, D.W., Weaver, C., 2009. Historical GIS and visualization insights from three hotel guest registers in central Pennsylvania, 1888—1897. Soc. Sci. Comput. Rev. 27, 348–362. http://dx.doi.org/10.1177/0894439308329762.

Garlandini, S., Fabrikant, S.I., 2009. Evaluating the effectiveness and efficiency of visual variables for geographic information visualization. In: Hornsby, K., Claramunt, C., Denis, M., Ligozat, G. (Eds.), Spatial Information Theory. Springer Berlin, Heidelberg, pp. 195–211 http://dx.doi.org/10.1007/978-3-642-03832-7_12.

GBIF, 2011a. Darwin Core Archive Assistant User Guide. Global Biodiversity Information Facility, Copenhagen, Denmark, p. 33 (http://links.gbif.org/gbif_dwc-a_asst_en_v1.1).

GBIF, 2011b. Darwin Core Archives—How-to Guide. Global Biodiversity Information Facility, Copenhagen, Denmark, p. 21 (http://www.gbif.org/resources/2551).

GBIF, 2013. Darwin core archive validator. http://tools.gbif.org/dwca-validator/ (accessed 3rd April 2014).

Graham, M., Kennedy, J., 2007a. Exploring multiple trees through DAG representations. IEEE Trans. Vis. Comput. Graph. 13, 1294–1301. http://dx.doi.org/10.1109/TVCG.2007.70556.

Graham, M., Kennedy, J., 2007b. Visual exploration of alternative taxonomies through concepts. Ecol. Inform. 2, 248–261. http://dx.doi.org/10.1016/j.ecoinf.2007.07.004.

Haklay, M., Weber, P., 2008. OpenStreetMap: user-generated street maps. IEEE Pervasive Comput. 7, 12–18. http://dx.doi.org/10.1109/mprv.2008.80.

Holman, E.W., 1985. Evolutionary and psychological effects in pre-evolutionary classifications. J. Classif. 2, 29–39. http://dx.doi.org/10.1007/BF01908062.

Hong, J.Y., D'Andries, J., Richman, M., Westfall, M., 2003. Zoomology: comparing two large hierarchical trees. In: Plaisant, C., Fekete, J.-D. (Eds.), IEEE Symposium on Information Visualization Poster Compendium. IEEE Computer Society Press, Seattle, Washington, USA, pp. 120–121.

JSPerf, 2014. JavaScript performance playground. http://jsperf.com (accessed 1 April 2014).

Kennedy, J.B., Kukla, R., Paterson, T., 2005. Scientific names are ambiguous as identifiers for biological taxa: their context and definition are required for accurate data integration. In: Ludäscher, B., Raschid, L. (Eds.), 2nd International Workshop on Data Integration in the Life Sciences. Springer, San Diego, California, USA, pp. 80–95. http://dx.doi.org/10.1007/11530084_8.

Kruskal, J.B., Landwehr, J.M., 1983. Icicle plots: better displays for hierarchical clustering. Am. Stat. 37, 162–168. http://dx.doi.org/10.1080/00031305.1983.10482733.

Leaflet.js, 2014. Leaflet.js—an open-source JavaScript Library for mobile-friendly interactive maps. http://leafletjs.com (accessed 1st April 2014).

Munzner, T., Guimbretière, F., Tasiran, S., Zhang, L., Zhou, Y., 2003. TreeJuxtaposer: scalable tree comparison using focus + context with guaranteed visibility. ACM Trans. Graph. 22, 453–462. http://dx.doi.org/10.1145/882262.882291.

Otegui, J., Ariño, A.H., 2012. BIDDSAT: visualizing the content of biodiversity data publishers in the Global Biodiversity Information Facility network. Bioinformatics 28, 2207–2208. http://dx.doi.org/10.1093/bioinformatics/bts359.

Otegui, J., Ariño, A.H., Encinas, M.A., Pando, F., 2013. Assessing the Primary Data Hosted by the Spanish Node of the Global Biodiversity Information Facility (GBIF). PLoS One 8, e55144. http://dx.doi.org/10.1371/journal.pone.0055144.

Roberts, J.C., 2007. State of the Art: Coordinated & Multiple Views in Exploratory Visualization. In: Andrienko, G., Roberts, J.C., Weaver, C. (Eds.), Coordinated and Multiple Views in Exploratory Visualisation. IEEE Computer Society PressZurich, Switzerland http://dx.doi.org/10.1109/CMV.2007.20.

Schulz, H.-J., 2011. Treevis.net: A Tree Visualization Reference. IEEE Comput. Graph. 31, 11–15. http://dx.doi.org/10.1109/MCG.2011.103.

Schulz, H.-J., Hadlak, S., Schumann, H., 2011. The Design Space of Implicit Hierarchy Visualization: A Survey. IEEE Trans. Vis. Comput. Graph. 17, 393–411. http://dx.doi.org/10.1109/TVCG.2010.79.

Song, H., Kim, B., Lee, B., Seo, J., 2010. A Comparative Evaluation on Tree Visualization Methods for Hierarchical Structures with Large Fan-outs, ACM CHI. ACM Press, Atlanta, Georgia, USA, pp. 223–232. http://dx.doi.org/10.1145/1753326.1753359.

Spenke, M., Beilken, C., 2003. Visualisation of trees as highly compressed tables with InfoZoom. In: Plaisant, C., Fekete, J.-D. (Eds.), IEEE Symposium on Information Visualization Poster Compendium. IEEE Computer Society Press, Seattle, Washington, USA, pp. 122–123.

Steiniger, S., Hunter, A.J.S., 2013. The 2012 free and open source GIS software map – A guide to facilitate research, development, and adoption. Comput. Environ. Urban 39, 136–150. http://dx.doi.org/10.1016/j.compenvurbsys.2012.10.003.

Wertheimer, M., 1938. Laws of organization in perceptual forms. In: Ellis, W. (Ed.), A Source Book of Gestalt Pyschology. Routledge & Kegan Paul, London, pp. 71–88.