Plant Ecology and Evolution 148 (2): 256–266, 2015
http://dx.doi.org/10.5091/plecevo.2015.1048

REGULAR PAPER

Using legacy botanical literature as a source of phytogeographical data

Quentin J. Groom

Botanic Garden Meise, Nieuwelaan 38, BE-1860 Meise, Belgium
E-mail: quentin.groom@br.fgov.be

Aim – Paper-based publications were the main repository for phytogeographical information until the end of the 20th century. These texts are still an important reference source for phytogeography and potentially a valuable source of data for research on environmental change. The recent digitization of biodiversity publications, text-mining and mark-up protocols means that these data are now more accessible than ever before. Here I examine the value of legacy literature specifically for studies on phytogeography.
Methods – Three contrasting data mobilisation projects are used as case studies for the extraction of phytogeographic data. Two were digitisations and XML mark-up of floras, the *Flore d'Afrique Centrale* from the 20th century and the *Flora of Northumberland and Durham* from the 19th century. A third case study used *Chenopodium vulvaria* L. as a test case, where I attempted to recover as much phytogeographic data as possible for one species, both from literature and from herbarium specimens.
Results – A large amount of useful information was extractable from legacy literature. The main limitations are that most localities need georeferencing and that observations are only rarely associated with a precise date. In the case of *C. vulvaria* literature contributed about 20% of all available observations of the species. Literature becomes a progressively more important source of data the further back in time one looks. However, useful observations become much rarer earlier than about 1850.
Main conclusions – Sourcing phytogeographic data from legacy literature is valuable. It contains observations and links to other data that are unavailable from any other source. Nevertheless, its extraction takes a substantial investment in time. Before commencing on such a project it is important to prioritise work and understand the limitations of such data, particularly with regard to georeferencing.

Key words – XML mark-up, georeferencing, botanical literature, phytogeography, observations.

INTRODUCTION

Reliable biogeographic data are required for many purposes including conservation, distribution modelling and understanding biogeographic change. Much of current observational data goes directly into digital media and can be used immediately; however, it is widely acknowledged that there is a large legacy of historical information that is stored on the labels of specimens and in the text of biological literature (Meier & Dikow 2004, Hardisty & Roberts 2013, Nelson et al. 2013).

Many museums and herbaria are already involved in large-scale digitization, transcription and georeferencing of their herbarium collections. Such data have found use in many biogeographic studies (Holland 1975, Loiselle et al. 2008, Lavoie 2013). Another source of biogeographic data is the corpus of biodiversity literature. This literature dates back at least four hundred years and the information contained within texts has increased with time. The categories of biogeographic information contained in literature include localities, chorology, estimates of abundance, habitats, collector names and collection numbers. Biogeographic data in literature can provide a more refined source of data than specimens, but also connects specimens and localities to taxon concepts and to other literature. Indeed, specimens and biodiversity literature are intimately connected, each supporting and validating the other. It is estimated that there are 350 million herbarium specimens in 3,400 herbaria in the world and as of April 2014 the Biodiversity Heritage Library has 42 million pages of scanned biodiversity literature (New York Botanical Garden 2012, Kalfatovic 2014). Combined, this constitutes a vast interconnected source of biogeographic information, however these connections are only implicit and their true value will only be made explicit through digitization.

One method for extraction of biogeographic data requires digitisation; mark-up of the data elements of each observation; georeferencing and extraction in a useable format. This method of data extraction has the advantage that the data can be connected back to its original context within the

All rights reserved. © 2015 Botanic Garden Meise and Royal Botanical Society of Belgium – ISSN 2032-3921

document. Various projects have done this, but there is no standard method (Kirkup et al. 2005, Curry & Connor 2008, Hamann et al. 2014). Literature is generally scanned and converted to digital text by OCR or double keying. The digital documents are then marked up with XML tags to create a structured document. The means used to mark-up text varies, documents can either be marked-up manually or automatically, however, even automated processes require some manual proofreading due to variability in the layout of published texts and inevitable errors in the original and the digitised text (Sautter et al. 2007).

Manual methods, using text or XML editors, can also proceed quite rapidly, if linked to a schema or document type definition. Automated methods generally use an XML parser to scan the text and regular expressions to identify elements for tagging (Kirkup et al. 2005). Scripts or macros can be written to identify features within the document to aid the mark-up, these features might be keywords, such as subsection titles or they might be particular formatting given to a particular type of text, such as italics for Latin names. In some highly structured text, considerable use can be made of punctuation, which delineates items of interest. If documents are small and weakly structured the simplest option can be to edit them manually.

Ideally, a biodiversity observation should have four essential elements encapsulated in the phrase "what, when, who and where". Though there are several other pieces of information that improve the usability of observations, including estimates of abundance, habitat descriptions, phenology, maturity and size. Yet, such data are frequently collected in an ad hoc manner and are not always complete. Furthermore, data collection is extremely biased taxonomically, temporally and spatially. Indeed, one can only hope to interpret these data correctly if one better understand these biases. More complete data mobilization is not a solution to the problem of bias, but having the full complement of data does help reveal these biases and therefore helps interpret these data correctly. Indeed, scarcity of sources and biased data are the norm in historical research and while this represents a challenge, they are not an insurmountable barrier to interpretation.

Biogeographic data are not the only type of data contained within biodiversity literature. These works contain information on morphological traits, habitats, bibliographic references, nomenclature etc. Extracting all of these data simultaneously has advantages in terms of completeness and efficiency, but it also adds costs to extracting priority data. Whether it is worth the additional cost of highly atomised mark-up depends on the use case. However, repositories of marked up documents, such as Plazi (www.plazi.org), can be used to store documents at varying levels of atomization, so that different users can mark-up the text elements as they require.

This paper summarises experience I have gained from three projects in the mark-up of legacy literature. The first project, the *Flore d'Afrique Centrale*, is a monographic flora series for the vascular plants of the Democratic Republic of the Congo, Rwanda and Burundi (Various editors 1948–1971, 1973–2005). It was started in 1948 and is published by the Botanic Garden Meise (formerly the National Botanic Garden of Belgium). It is the only comprehensive source of botanical information for this region and the Garden wanted to make it more accessible internationally. The series consists of about 8000 pages of text containing taxonomic treatments with details of nomenclature, bibliographic references, descriptions, distributions, collection details of specimen, habitats, vernacular names etc. Next, the *Flora of Northumberland and Durham* by Nathaniel John Winch is an excellent example of an early phytogeographic flora (Winch 1838). It contains details of localities for vascular plants, Bryophytes, algae and fungi in the North-east of England. Additionally, the flora contains bibliographic references, indications of abundance, habitats and details of botanists who either first found or observed the species at these sites.

Lastly, in a project to understand the phytogeography of *Chenopodium vulvaria* L., an attempt was made to discover as many observations of this species as possible from a wide variety of sources, including observation databases, specimens and published observations. These data have been collated and provide indications of the volumes of data available from different sources (Groom 2015).

The aim of this paper is to evaluate the value of legacy literature as a source of phytogeographic information. By drawing on my experience of digitising a variety of texts from a wide variety of literature, publication dates and different languages, I assess how well legacy literature provides the 'what', 'when', 'who' and 'where' of a useful observation. The focus here is on the literature pertaining to vascular plants, but I anticipate that the results will be applicable to other organisms. This paper will help others embarking on similar projects to prioritise their work and avoid common pitfalls.

MATERIAL AND METHODS

Flore d'Afrique Centrale

The text of the *Flore d'Afrique Centrale* (FAC) was parsed into a custom XML schema using a semi-automated process, then the finely atomised data were imported into a database from where it is now displayed online (http://www.br.fgov.be/RESEARCH/DATABASES/FOCA/index.php). Scanning and optical character recognition was outsources to SunTec Web Services Pvt. Ltd. The digitized text was delivered as Microsoft Word documents. The treatments were then split into species treatments and family and generic treatments. This simplified the design of scripts used for automated mark-up so that they could process one format of treatment layout at a time. A series of scripts were written in Perl (version 5.8.8) using the Expat XML parser. Each script progressively marked-up the text with finer granularity using regular expressions to identify key elements within the text. After each script the document was confirmed against an XML schema to ensure that the XML was well formed and valid. Errors were manually corrected at each step in the mark-up process. Errors in mark-up from the scripts were largely due to OCR errors resulting in the regular expressions not recognising their targets. The rigor imposed upon the process by confirming each step ensured quality and led to the correc-

tion of many errors which would have otherwise been passed on to the final product.

The biogeographic data in the flora consists of both chorological information and details of specimens. The chorology is a simple list of countries, whereas the specimen details consist of the collector name and collection number together with the name of the collecting locality and one of eleven phytogeographic zones of the region.

Three different approaches were taken to georeferencing the named localities. Firstly the data were submitted to the web application GEOLocate creating automated geolocalities for the names of collection sites (Rios & Bart 2010). Secondly, a digitised gazetteer of Central African collecting sites was applied to the data (Bamps 1982). Thirdly, the collector names and collection numbers on already geolocated herbarium specimens were matched against the collector names and collection numbers in the flora so that the georeferencing of the specimens could be reused for the flora. These georeferencing methods were used in succession with each step, superseding georeferences from the previous step. In this manner I ensured that coordinates from herbarium specimens were preferred over coordinates from the gazetteer and those were preferred over coordinates from GEOLocate.

Flora of Northumberland and Durham

The text of the *Flora of Northumberland and Durham* (FND) was taken from Winch (1838). A digital version of the text was downloaded from the Internet Archive as a DjVu file (https://archive.org/details/transactionsofna1838natu). This document was originally scanned for the Biodiversity Heritage Library by the Ernst Mayr Library of the MCZ, Harvard University (http://biodiversitylibrary.org/page/33669545). The original OCR had been conducted using ABBYY FineReader 8.0, but the accuracy of the OCR was too poor to be used without significant correction. Therefore, the text was uploaded into Wikisource (wikisource.org) where the text was corrected manually. (https://en.wikisource.org/wiki/Index:Transactions_of_the_Natural_History_Society_of_Northumberland,_Durham,_and_Newcastle-upon-Tyne_1838_Vol.2.djvu). Wikisource is an online collaborative library where text can be loaded, corrected and downloaded for reading. Once a corrected text was available it was downloaded and custom Perl scripts were used to convert the text to XML and then manual marking up was used to identify text elements such as location, descriptions and people's names. Once the mark-up was complete the localities were georeferenced manually using a variety of online and paper maps, gazetteers and floras. The text was then republished in the Advanced Books system of the publisher Pensoft (Winch 2014).

Chenopodium vulvaria distribution data

Observation and specimen details were collected in a Common Data Model (CDM) database which is the central component of the EDIT Platform for Cybertaxonomy (Ciardelli et al. 2009). Two methods were used to extract observations from literature, either XML mark-up or direct data entry. Digitised treatments were marked up with XML using the GoldenGate editor (Sautter et al. 2007); uploaded to the PLAZI taxonomic treatment repository (www.plazi.org) and imported to the CDM database. Alternatively the observation details were copied from the treatment and entered manually into the CDM database using the EDIT Taxonomic Editor (Ciardelli et al. 2009). Observations where gathered from the biodiversity literature by reading the Biodiversity Heritage Library corpus systematically after searching for *Chenopodium vulvaria* L. and its synonyms *C. foetidum* Lam., *C. olidum* Curt., *Atriplex vulvaria* Crantz and *Vulvaria vulgaris* Bubani. Other published observations were gathered from the Library of the Botanic Garden, Meise. A complete survey of non-digitised literature is impossible, but there was an effort to check multiple floras of every European country and any other country with a climate suitable for *C. vulvaria*. A reference list to the extracted observations is available in Groom (2015).

Digitised observation data were also gathered from databases, primarily from the Global Biodiversity Information Facility (GBIF) (http://gbif.org, accessed 8 Nov. 2013; see Groom 2015), but also from the Atlas of Living Australia (www.ala.org.au, accessed 25 Feb. 2013); the Botanical Society of Britain and Ireland (http://bsbidb.org.uk, accessed 23 Feb. 2013) and Herbaria@home (http://herbariaunited.org, accessed 23 Feb. 2013). Scientific articles and websites containing observations were also discovered using search engines (http://scholar.google.be/; www.google.be). Data from databases were imported directly into the CDM database.

Specimen data were gathered from either databases or from herbaria by transcription of label information. Specimens from the following herbaria are included in the study, names and abbreviations follow those in the Index Herbariorum (http://sweetgum.nybg.org/ih/). University of Wales (ABS); University of Birmingham (BIRM); Botanical Museum Berlin-Dahlem (B); Bulgarian Academy of Sciences (SOM); Charles University in Prague (PRC); Herbier J.H. Fabre (FABR); Institut Botànic de Barcelona (BC); Institute of Biodiversity and Ecosystem Research, Bulgarian Academy of Sciences (SOMF); Nationaal Herbarium Nederland (L); The Natural History Museum, London (BM); University of Manchester (MANCH); Botanic Garden, Meise (BR); Moscow State University (MW); Botanische Staatssammlung München (M); Museu Nacional de História Natural e da Ciência (LISU); Museum national d'Histoire naturelle (P); National Academy of Science, Kyrgyzstan (FRU); Natural History Museum of Denmark (C); New York State Museum (NYS); National Museum in Prague (PR); Reading University (RNG); Royal Botanical Gardens, Kew (K); Sapienza University of Rome (HFLA), Sofia University (SO); South London Botanical Institute (SLBI); Universidad Nacional del Sur Herbario (BBB); Universität Wien (WU); Universidad de Concepción, Chile (CONC); University of Alaska Herbarium (ALA); University of California (UC); University of British Columbia (UBC); Wageningen University (WAG) and others contributing data to GBIF. Numerous other herbaria and herbarium catalogues were searched without finding specimens and several herbaria were contacted and either contained no specimens or did not respond.

Georeferencing was carried out manually, except for the rare occasions when coordinates were available with the specimen or observation (Chapman & Wieczorek 2006).

A potential source of data that was not included were dot maps in distribution atlases, such as the Flora Europaea (Jalas & Suominen 1980). These data have been derived from specimens or other observations. While these are a potential source of distribution data they lack additional metadata, such as dates, location names and collector names and they have no corroborative information to aid verification.

RESULTS

Flore d'Afrique Centrale

The FAC contained 72,769 references to herbarium specimens, with 22,208 collection localities, of which only 15.7% could be georeferenced through direct reference to gazetteers and previously georeferenced specimens. Two of the main obstacles to more complete georeferencing were the large number of homonymous place names and spelling variants.

Figure 1 shows the results of the partially automated georeferencing of observations from one of the phytogeographic regions in the FAC. If there were no errors in the georeferencing nor in the original publication all the observations should be within the regional boundary, however, there are many mislocated sites, either because the authors were incorrect in their assignment of specimens to a phytogeographic region or, more likely, that the georeferencing was wrong.

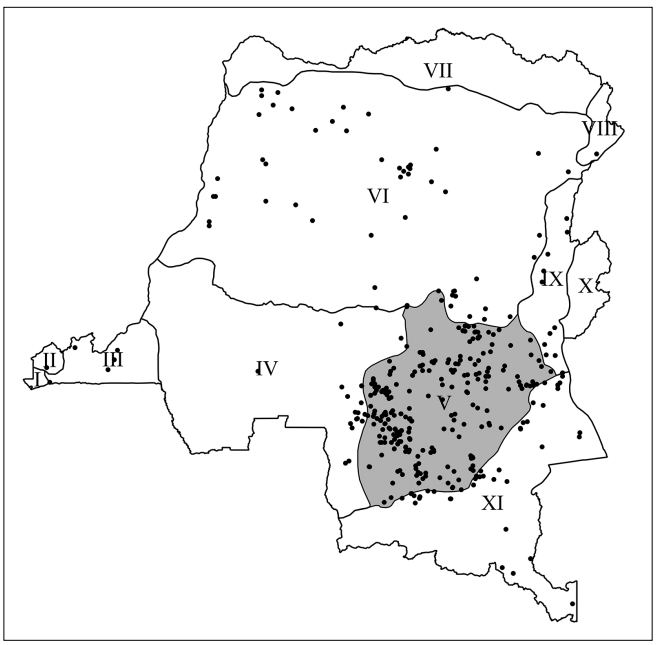

Figure 1 – A map of the automated georeferencing of observations from Bas-Katanga taken from the *Flore d'Afrique Centrale*. The region Bas-Katanga is in grey. Points outside the borders of the phytogeographic region are errors from mistakes either in the georeferencing or in the original publication. The phytogeographic regions are numbered: I, Côtier; II, Mayumbe; III, Bas-Congo; IV, Kasai; V, Bas-Katanga; VI, Forestier Central; VII, Ubangi-Uele; VIII, Lac Albert; IX, Lacs Édouard et Kivu; X, Rwanda-Burundi; XI, Haut-Katanga.

Regarding additional data, no observation dates are found in the FAC over seventy years of publication, but nearly all accounts have detailed habitat description and a summary of the global distribution. Furthermore, there are detailed descriptions of each taxon, notes on ethnobotany and vernacular names.

Flora of Northumberland and Durham

The FND has entries for 2,599 species and about 5,800 observations. For vascular plants, where there are observations, there is an average of 4.3 observations per species, but these are not normally distributed, 27% have only one observation.

Even though the FND was published over 170 years ago the taxa are recognisable. Only certain critical taxa were difficult to assign to modern concepts. Indeed, for the 1,050 treatments of vascular plants, only ten could not be assigned to accepted modern taxa. Exceptions include "*Rosa gracilis*" and "*Salix hirta*". Even though taxonomic authorities are not cited, as they are in modern texts, the names are well supported by references to other authorities, though not in many cases to the person who first proposed the name. These references helped to clarify the name where it was unclear.

The FND cites people and sometimes literature together with the observation. It is not clear if these citations refer to the person who first recorded the species at the location or if this was the only record of the species at the location. Certainly, it is evident from the text that the author visited at least some of the sites in question. For example, under the entry for *Woodsia ilvensis* (L.) R.Br., two sites are mentioned with the observers James Backhouse and S. Halestone. The author writes "These localities cannot be far asunder" suggesting he has either seen or has current knowledge of the sites. Indeed, this species did become extinct at these sites.

The georeferenced observations from the FND are mapped in fig. 2. They are believed to be more accurate than the georeferencing for the FAC, because their locations were georeferenced manually and because better maps and gazetteers are available for England. Their distribution is non-random being more frequent in towns, along river corridors and at sites of rare species, such as in Teesdale at the south-west corner of the map.

In the FND a number of texts are cited for each taxon using an abbreviated shorthand. At the time, the cited texts would have been well-known to the readers, but these now require some investigation to decipher. For example, "*Sm. Eng. Fl. iv. 239*" refers to The English Flora by James Edward Smith, 1828, volume 4 page 239. References such as these were determined with confidence by referring to the original text where the page numbers could be verified. A large proportion of these texts are available on the Biodiversity Heritage Library and other digital libraries.

Additional information in the flora is limited. Most taxa have a brief habitat description, but only twenty-one dates are present in the FND out of over 5,700 observations.

The *Chenopodium vulvaria* corpus

Given the considerable biases involved in the collection and digitization of biodiversity data there are no truly unbiased datasets with which to compare. Nevertheless, I have digitized such a large number of the available specimens and literature on *C. vulvaria* that I believe this gives a useful indication of trends. There were good reasons to choose *C. vulvaria* as a test subject. It is a largely European species so it will be well represented in historical texts; *C. vulvaria* is generally rare so the volume of data is reasonable to collect; it is an ugly plant, so it has not been over-collected for aesthetic reasons; its foul smell makes it quite unique and, at least in Europe, impossible to mistake for any other plant and finally its habitat is always associated with humans, so its locations are comparatively easy to georeference.

The corpus of *C. vulvaria* contains 2,497 georeferenced observations or specimens, though at least 100 of these are duplicates. It is not always possible to determine whether a specimen or observation is a duplicate or just a similar observation in time and space. Duplicates were more or less obvious where vouchers have been dispersed to multiple herbaria, however, published observations were not always easy to identify as duplicates of specimens, unless the collector, collection number and date are given in the publication. Within all the literature discovered on *C. vulvaria*, only 4% has a precise date and 11% has the observation year. The remainder are dated by the year of publication.

A comparison of the change in the number of observations with time is shown in fig. 3. The number of observations shows a considerable increase around 1850, but largely levels off during the 20th century. Overall, 18% of all the observations are from the literature. Using these data it was possible to breakdown the sources of these different observations and track how this has changed with time (fig. 4). Literature is proportionately more important for older dates. For the period prior to 1850 it is the most important category.

There is a large variation between countries in the proportion of observations that come from literature. Eight countries only had observations from literature, though these only account for 33 observations in total, examples are Azerbaijan, Croatia and Lebanon. A further eleven have more than half their records from literature; including Iran, Italy, Poland and Romania. In contrast, 25 countries have less than 10% of their observations from literature, including Bulgaria, Czech Republic, France, Greece, Spain and Sweden.

The nomenclature of botanical literature is not consistent. Observations of many species are published under a number of synonymous Latin names. *C. vulvaria* has five synonyms,

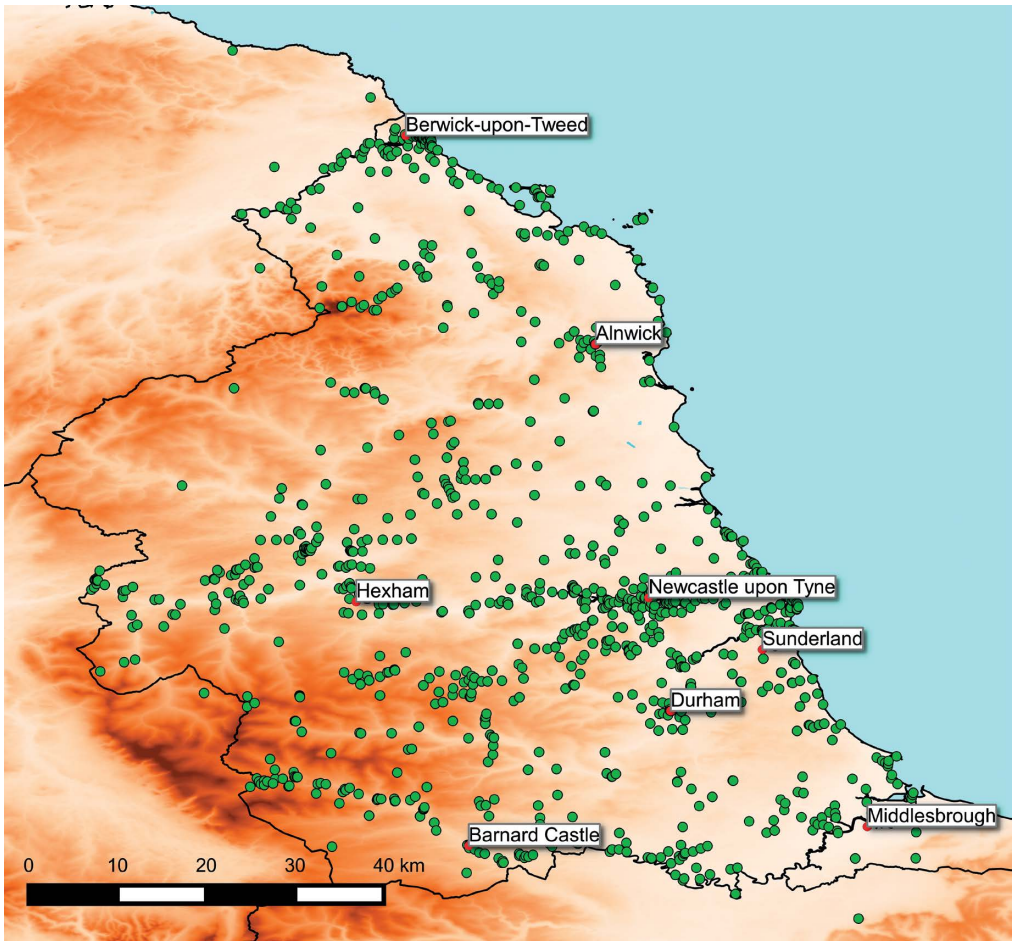

Figure 2 – A map of the georeferenced observations from the *Flora of Northumberland and Durham* (Winch 2014), showing their non-random distribution. Observations outside the boundaries of the region are not errors, but additional sites mentioned in the text.

but only the names *C. olidum* Curt. and *C. vulvaria* L. are commonly used. The only significant complication was the homonym *Chenopodium olidum* S.Watson, which is an illegitimate name for *Chenopodium watsonii* A.Nelson. This required particularly scrutiny of North American observations, where this species is native.

In the georeferenced *C. vulvaria* corpus, 87% had an error radius below 10 km and the mode was about 5 km, however, many of the observations derived from GBIF providers lack error radii. Georeferencing of literature was considerably easier than georeferencing specimens, because text was legible and because it was usually in the context of a publication with a particular geographic focus. Also, additional clues to the locations came in the form of maps, and habitat descriptions. The *C. vulvaria* corpus has habitat descriptions in 12% of the species accounts.

DISCUSSION

Pre-digital sources of observations are only from specimens, published literature, maps, unpublished notebooks, pictures and record cards. On occasion these sources duplicate information, while others are unique. Nevertheless, even when the same observations are digitised from different sources the presence of two independently digitised records is valuable for the discovery of errors. Where the same data are available from several sources it is usually sensible to prioritise the digitization of one reliable source, the one that will be both cost effective and complete. So far most progress has been made with specimens, but this is probably due to their uniqueness and fragility not necessarily due to their value as a source of data. In fact, when the herbarium of Nathaniel Winch is digitised and transcribed, it will be useful to access

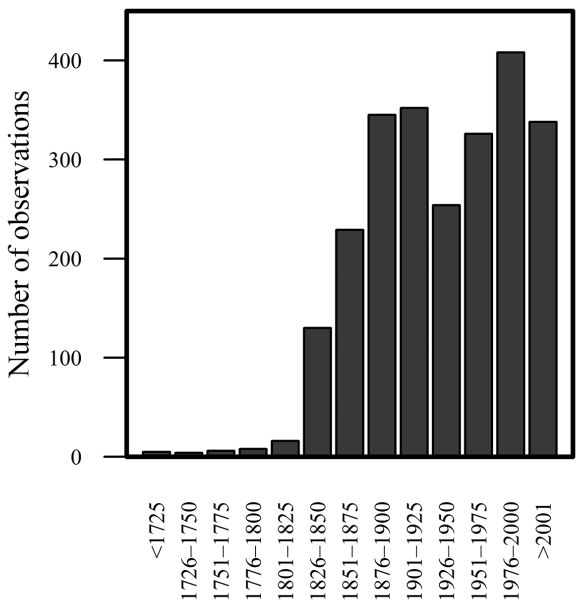

Figure 3 – The total number of observations of *Chenopodium vulvaria* gathered from all sources and grouped in 25-year periods.

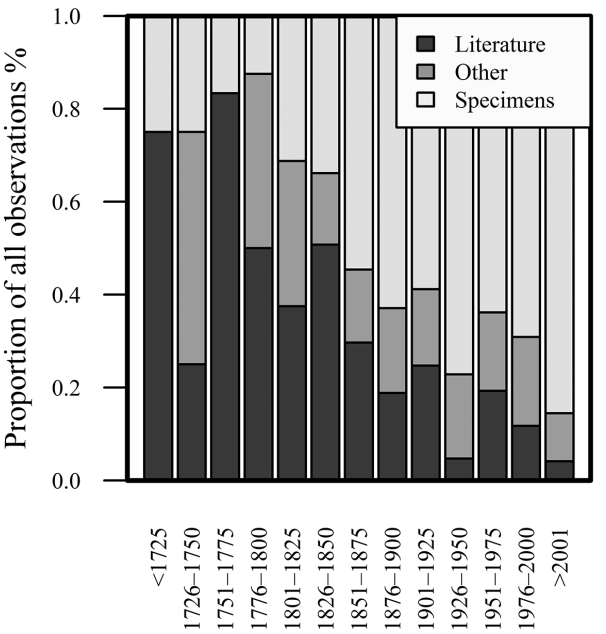

Figure 4 – The proportion of observations of *Chenopodium vulvaria* retrieved from specimens and literature, grouped in 25-year periods. Dark grey bars are the proportion of observations extracted from literature; light grey bars represent the proportion of observations extracted from herbarium specimens and mid-grey bars represent other observations from databases that could not be ascribed to either herbarium specimens or literature, in recent years some of these will have been direct observations entered directly into databases. In older cases these will have been either specimens or published observations that have been digitised without reference to their source.

a digital version of his Flora where his taxonomic names, observer names and locality names have been deciphered from printed text, rather than from handwriting.

The total number of observations is both dependent on the activities of botanists and on the changes in abundance of *C. vulvaria*. The increase in observation around 1850 seems likely to be the result of an increase in botanical activity; this is earlier than the modern increase seen in the analysis of dated observations elsewhere (Rich 2006, Otegui et al. 2013). It can be explained by the inclusion of proportionately more literature observations in the *C. vulvaria* data. In northern Europe the increase of botanical activity in the 20th century is offset by the decline in abundance of *C. vulvaria*, resulting in little overall change in the number of total observations in the 20th century (fig. 3).

For the *C. vulvaria* corpus the total contribution of literature observations may seem small compared to specimens. However, published observations filled important gaps. In this example, literature was important for early dates and for certain countries with inaccessible herbaria. When investigating the distribution of a widespread species, personal visits to all the appropriate herbaria are impossible and inefficient, even contacting individual herbaria for details of their collections is slow and often fruitless work. So until digitiza-

Table 1 – The differences between specimens and literature as sources of phytogeographic data.
Specimens and literature are complementary sources of phytogeographic data. They often relate to each other, but they are not the same. Each source has particularly characteristics that make it more suitable for different applications.

	Literature	Specimens
Taxonomy	Standardised and cited	Variable and uncited
Dates	Usually undated, except for the publication date	Usually dated
Collector/Observer	Well-recorded and standardised	Often not recorded or with illegible signatures
Location	Often vague, but more standardised than on specimens and related to higher geographic units	Often vague, but generally more specific than in literature
Metadata	Often well connected to other information including habitat morphology and other literature	Rarely with additional notes, particularly on habitat
Legibility	Good in all languages, except for black letter text	Often difficult, with frequent abbreviations and spelling variants

tion of herbaria in common place literature will remain particularly important for some parts of the world. Some fragile and seldom collected species may also be better represented in literature than in specimens. Palms, for example, are difficult to collect and store.

Collectors tend to over-represent rare species in their collections (Garcillán & Ezcurra 2011, ter Steege et al. 2011). Common species are either ignored or only collected when they or their location differs in some respect from the norm. *C. vulvaria* is considered rare in northern Europe, where it may be over-represented in collections in comparison with southern Europe where it may be underrepresented. In literature a similar bias occurs. In the FND no observation details are given for common species, on the contrary their abundance is often described in words such as 'common' or 'abundant'.

Unlike specimens, biodiversity literature reveals, sometimes subtly, abundance, habitat and location, an example is the phrase "On the sea shores of Northumberland and Durham abundant, not very common on the coast near Berwick" (Winch 1838). These expressions of abundance are usually related to areas within the landscape, which in the case of the FND are English counties. While, such qualitative descriptions seem unhelpful, if large numbers of such descriptions were available they could be analysed by coding the expressions of abundance numerically and by using ordinal statistics.

Specimens are more fragile than books and although they are sometimes duplicated, many are unique; there are many examples of lost herbaria and countless specimens destroyed by such things as floods and insects. Books, in contrast, are printed in many copies, reprinted, revised and translated. Therefore, books last longer than specimens. Furthermore, although herbarium specimens have been collected for more than 200 years, early specimens are often poorly annotated, often with little more information than the species name. Carl Linnaeus, for example, gave practically no information on his specimens yet, his *Species Plantarum* does often indicate where a species grows.

Below I examine how well legacy literature provides the four essential elements of an observation i.e. 'what', 'when', 'who' and 'where'. These conclusions have been summarised in table 1 where the differences between literature and specimens are compared as sources of data.

What was observed? Taxonomy

In each example of digitised literature, taxonomy and nomenclature can differ from modern accepted use. Nevertheless, for the vast majority of species the synonymy is simple and unambiguous. There are exceptions, but as a rule of thumb, difficult genera in modern literature are difficult in historical literature. There is a large quantity of supplementary information in the FAC to facilitate understanding of the name concept, including nomenclatural references, synonymy and details of the type specimen. The taxonomic information in the FND was less, but was sufficient to establish the identity of all but a few taxa. Generally speaking, synonymy is not a major limitation to the use of the data in these texts. The only exception is where a species has been subsequently split into multiple species and it is not obvious how the details in the taxonomic treatment should be subdivided, though in some cases the division has already been recognised in the Flora at a subspecific level.

When did the observation occur? Date of publication

Observations in biodiversity literature almost always lack a precise date. This omission of useful information is not as critical as it might appear. The publication date fixes an end date to the presence of the organism at the location. Indeed, one only ever knows the last sighted date of an organism.

If a work, such as the FND, was the work of a single individual then one can assume that the observations date from the active period of the botanist. For example, Nathanial Winch was born in 1769 and based upon the few dates in his Flora he was actively observing plants from 1797. His Flora was read before the Natural History Society of Northumberland, Durham, and Newcastle-upon-Tyne on 20 June 1831. At this reading he stated "...it [the Flora] is the result of more than 30 years' attention to our native Botany,...", which is consistent with the start date of 1797. Although the reading before the society was in 1831 the text was not published until 1838 and it is not clear if any changes occurred in the text after 1831. The last observation date mentioned in the text was in 1828 and the last citation was from 1830. Therefore,

it seems reasonable to fix the dates of observations between 1797 and 1831.

Similar deductions can be made for most observers, particularly where good biographical information is available. At least in the case of plants, these data can still be used for biogeographic modelling, but it does rule out any study of short term periodic phenomena, such as phenology.

Who made the observation? The observer

Attribution of an observation to a person is not only done to give credit to the observer. It provides corroborating evidence to the veracity of the observation and can help verify the location and dates attributed to an observation. For example, the clergyman and natural historian William Turner (circa 1508 – 13 Jul. 1568) described *C. vulvaria* in his book *The names of herbes* (Britten 1881). He mentioned two sites for this plant, one in Callice [Calais], France and another in Bon [Bonn], Germany. These observations seem odd localities for an English clergyman of this period, unless you know from his biography that he travelled widely in Europe and lived for a time in Cologne Germany, close to Bonn (Raven 1947). Likewise, Winch in the FND mentions the abundance of *Funaria hygrometrica* Sibth. on Vesuvius, which would seem a peculiar observation unless you know that he travelled in the Mediterranean (Whewell 1839).

Overall, attribution of observations to observers was one of the most reliably recorded data elements. This underscores the importance of maintaining and disseminating bibliographic information on botanists for the verification of observation details. An important requirement for future biodiversity informatics is a repository containing biographical details of collectors and observers, containing information on when they were active, what they studied and where they worked. Such a resource would be invaluable for the validation and georeferencing of historical biodiversity observations. It could also be an important resource for investigations into the history of science (Groom et al. 2014a).

Where was the observation made? Chorology and georeferencing

Georeferencing of observations from literature is similar to georeferencing specimens, indeed the same methods can be used (Chapman & Wieczorek 2006). However, the clear type and literary context can make georeferencing easier, especially in comparison with handwritten labels. Within a publication, site names are often consistent and being able to cross reference localities makes situating them easier.

There have been attempts, at least in part, to automate georeferencing (Beaman & Conn 2003, Guralnick et al. 2006, Rios & Bart 2010). However, I did not find automated georeferencing particularly useful. This was, in part, due to the lack of good quality gazetteers for Africa, but perhaps most importantly to the fact that such systems create a relatively high proportion of false positive errors due to mislocation. While this is a problem, the additional regional information in the FAC did allow grossly mislocated observations to be identified more easily (fig. 1).

Manual georeferencing is a time consuming process that requires experience, local knowledge and a good understanding of the pitfalls, such as homonyms, changing places names and moving boundaries. In these examples, manual rather than automated georeferencing was more reliable and complete. Yet, it is time consuming to manually georeference observations, even with detailed digital maps and gazetteers. Good quality georeferencing of old specimens and old literature requires fully synonymised gazetteers preferably with an indication of the dates when place names were used. It also requires a degree of interpretation, taking into consideration many factors including the publication date and the conventions of the author. Resolutions comparable to modern data are not possible for the majority of published observations and analysis of these observations has to reflect this in the interpretation.

The distribution of observations reflects both a bias in surveying area (fig. 2) and towards rare species. The FND contains no precise locations for common species. The rare species within towns are generally alien species. Newcastle-upon-Tyne and Sunderland were important ports in the 19th century and many aliens were introduced in the ballast of ships (Winch 1838). The frequency of observations along rivers may, in part, be because the main roads followed the river valleys. These biases can be seen by comparing these data with modern floras such as Swan (1993) and Groom et al. (2014b), with the georeferenced data (Groom 2014). Such biases probably exist in all floras and in collections of specimens (Rich 2006, Loiselle et al. 2008).

Floras are, in general, much better at describing the presence of a species in a geographic area than describing point locations, yet, such chorological information is not generally used in models of plant distributions, even though it could be considered more reliable and less subject to observer bias (Barbosa et al. 2012). Such data have many uses in ecological analysis, conservation and policy making.

Problems of very old texts

Old taxonomic literature, particularly those before the 20th century follow different conventions to modern literature. Fortunately, the typeface used for English, French and other western European languages has changed little over the past 200 years. However, black letter script, still widely used in the 19th century, particularly in Germany, presents problems for commonly used OCR engines. Furthermore, ligatures were often used within taxonomic Latin names, some of which are difficult to distinguish both for a human and automated reader. For example, the ligatures œ and æ can be indistinguishable and OCR has problems with combinations of characters such as ij and fi. These are problems particularly related to taxonomic names in early literature, but there are also modern characters that are problematic, such as ‘e’ and ‘c’, capital ‘i’ and lower case ‘l’ and Latin ‘ii’ and German ‘ü’. As OCR errors can still form correctly spelled words, there is often a need for manual proof reading.

The use of taxonomic authorities is relatively new within taxonomic literature. They only became widespread after the publication of rules for botanical and zoological nomenclature in the 19th century (Nicolson 1991). Earlier texts vary;

Table 2 – Criteria in the decision process for marking up texts.
Before embarking on a project decisions need to be made on which texts to work on and which methods to use. This table summarizes some of the criteria to evaluate in making these decisions. Ticks signify that that the literature conforms positively to the criteria and crosses negatively.

	Flore d'Afrique Centrale	*Flora of Northumberland and Durham*	The corpus of *Chenopodium vulvaria* Literature
High print quality	✓	×	×
Highly structured text	✓	✓	×
Uniform style	✓	✓	×
Up-to-date information	✓	×	×
Containing unique information	×	✓	✓
Large quantity of information	✓	✓	×
Contains coordinates	×	×	×
Links different data types	✓	✓	✓
Citations to sources	✓	✓	×
Usages	Many, including phytogeography, plant identification and extraction of traits, linking to specimens.	Phytogeography, conservation, environmental history	Phytogeography, environmental history, habitat analysis
Methods used	Automated transcription and automated mark-up	Manual transcription and automated mark-up	Manual transcription and manual mark-up

some cite the original author, but sometimes other literature containing the name is cited.

Subspecific treatments in 19th century Floras will be unfamiliar to modern readers. In the FND, and many other Floras, subspecific divisions are referred to by Greek letters, usually α, β, γ and δ. Furthermore, only in recent years has the rank of subspecies been used in the taxonomic hierarchy, in the FND and FAC almost all subspecific taxa are referred to as varieties.

Before the rediscovery of Mendel's work at the turn of the 20th century hybridization was not understood. In some cases hybrids were treated as subspecific taxa, whereas in other cases the generation of fertile hybrids were seen as evidence that those apparently different taxa were in actual fact one (Herbert 1822). While adding to the difficulty of working with legacy literature these issues, when understood, are not significant obstacles to comprehension or marking up.

CONCLUSION

Extracting and using data contained within literature is not without its difficulties. Furthermore, some of these data may be too vague, incomplete or derived to be useful. The costs involved in extracting these data are currently high and prioritization for data mobilization is essential. Costs could be reduced with the economy of scale, but tools and workflows need to be developed to do this. A prerequisite to data extraction are assessment criteria for prioritisation, which will identify the texts either containing high quality data; high value data or easily obtained data. High quality data would include precise details of the observation particularly with population estimates. High value data would be those data that are immediately required for use and unobtainable elsewhere. Easily obtainable data are those that are well structured, those that contain geographic coordinates and those that are cleanly printed in type that can be accurately converted to digital text through optical character recognition. Table 2 outlines some of the criteria that can be used to make these decisions and how these relate to the different examples of mark-up I have used in this paper.

A report of the pro-iBiosphere project had two recommendations related to the efficient use of mark-up resources (Mietchen et al. 2014). Firstly, to concentrate on recent revisionary works, not the whole of taxonomy and secondly to differentiate between the use of semantic mark-up and annotation, using the most appropriate method for the source document and the availability of resources. These recommendations are appropriate for providing maximum benefit to the whole taxonomic community, though there will always be a demand for a project based focus and it is a challenge to the biodiversity informatics community to ensure that all semantically enhanced documents remain available and useful in the future.

Mobilising published phytogeographic data presents challenges, but the rewards are numerous. The biosphere is in constant flux and we need information at different times and places to study temporal trends. For some parts of the world and time periods historical literature is the only source of information. Indeed, much of the data collected in the past are irreplaceable and as more digitised literature becomes available it will be increasingly seen as an important source for data.

ACKNOWLEDGEMENTS

Thanks to Ana Isabel D. Correia of the Faculdade de Ciências da Universidade de Lisboa; Prof. Dr. Roberto Rodríguez Ríos from the Universidad de Concepción, Chile; Vladimir

Vladimirov of the Bulgarian Academy of Sciences; Neus Ibáñez and Noemi Montes of Barcelona Herbarium; Michal Štefánek of Charles University in Prague; Alan Paton of the Royal Botanic Garden, Kew; Leni Duistermaat and Gerard Thijsse of Naturalis Biodiversity Center, Leiden; Robert Vogt of the Botanischer Garten und Botanisches Museum, Berlin; the specimen digitization team and Salvator Ntore at the Botanic Garden, Meise; Sabine Metzger and all those people in museums and herbaria around the world who have made their collections accessible. This work has been supported in part by the EU BON and pro-iBiosphere projects, which have been funded by the European Commission under the 7th Framework Programme (grant agreement numbers 308454 and 312848 respectively).

REFERENCES

Bamps P. (1982) Flore d'Afrique Centrale (Zaïre, Rwanda, Burundi): Répertoire des lieux de récolte. Meise, Jardin botanique national de Belgique.

Barbosa A.M., Estrada A., Márquez A.L., Purvis A., Orme C.D.L. (2012) Atlas versus range maps: robustness of chorological relationships to distribution data types in European mammals. Journal of Biogeography 39: 1391–1400. http://dx.doi.org/10.1111/j.1365-2699.2012.02762.x

Beaman R.S., Conn B.J. (2003) Geoparsing and georeferencing of Malesian collection locality data. Telopea 10: 43–52.

Ciardelli P., Kelbert P., Kohlbecker A., Hoffmann N., Güntsch A., Berendsohn W.G. (2009) The EDIT platform for cybertaxonomy and the taxonomic workflow: selected components. In: Fischer S., Maehle E., Reischuk R. (eds) Informatik 2009 – Im Focus das Leben, Beiträge der 39. Jahrestagung der Gesellschaft für Informatik: 625–638. Bonn, Gesellschaft für Informatik.

Chapman A.D., Wieczorek J. (2006) Guide to Best Practices for Georeferencing [online]. Available from http://herpnet.org/herpnet/documents/biogeomancerguide.pdf [accessed 12 Jul. 2014].

Curry G.B., Connor R.C.H. (2008) Automated extraction of data from text using an XML parser: an earth science example using fossil descriptions. Geosphere 4: 159–169. http://dx.doi.org/10.1130/GES00140.1

Garcillán P.P., Ezcurra E. (2011) Sampling procedures and species estimation: testing the effectiveness of herbarium data against vegetation sampling in an oceanic island. Journal of Vegetation Science 22: 273–280. http://dx.doi.org/10.1111/j.1654-1103.2010.01247.x

Groom Q.J., O'Reilly C., Humphrey T. (2014a) Herbarium specimens reveal the exchange network of British and Irish botanists, 1856–1932. New Journal of Botany 4: 95–103. http://dx.doi.org/10.1179/2042349714Y.0000000041

Groom Q.J., Young G., Richards A.J. (2014b) The Rare and Scarce Plants of South Northumberland 2013. Figshare. http://dx.doi.org/10.6084/m9.figshare.1030416

Groom Q.J. (2014) All observations extracted from the Flora of Northumberland and Durham 1831 [online]. Available from http://www.gbif.org/dataset/bfaa049a-90cd-412d-9660-5380591ba4a5 [accessed 11 Nov. 2014].

Groom Q.J. (2015) Piecing together the biogeographic history of Chenopodium vulvaria L. using botanical literature and collections. PeerJ 3: e723. http://dx.doi.org/10.7717/peerj.723

Guralnick R.P., Wieczorek J., Beaman R., Hijmans R.J., the BioGeomancer Working Group (2006) BioGeomancer: automated georeferencing to map the world's biodiversity data. PLoS Biology 4: e381. http://dx.doi.org/10.1371/journal.pbio.0040381

Hardisty A., Roberts D., The Biodiversity Informatics Community (2013) A decadal view of biodiversity informatics: challenges and priorities. BMC Ecology 13: 16. http://dx.doi.org/10.1186/1472-6785-13-16

Hamann T., Müller A., Roos M.C., Sosef M., Smets E. (2014) Detailed mark-up of semi-monographic legacy taxonomic works using FlorML. Taxon 63: 377–393.

Herbert W. (1822) On the production of hybrid vegetables; with the results of many experiments made in the investigation of the subject. Transactions of the Horticultural Society of London 4: 15–47. Available from http://biodiversitylibrary.org/page/44337878 [accessed 1 Dec. 2014].

Holland P.G. (1975) The use of herbarium records in biogeography. The professional geographer 27: 475–479. http://dx.doi.org/10.1111/j.0033-0124.1975.00475.x

Jalas J., Suominen J. (1980) Atlas Florae Europaeae, Vol. 5. Helsinki, The Committee for Mapping the Flora of Europe & Societas Biologica Fennica Vanamo.

Kirkup D., Malcolm P., Christian G., Paton A. (2005) Towards a digital African flora. Taxon 54: 457–466. http://dx.doi.org/10.2307/25065373

Kalfatovic M.R. (2014) Building for demand: The growth of the biodiversity heritage library. 2014 EOD Conference. Innsbruck, Austria. 11 April 2014 [online]. Available from https://www.slideshare.net/Kalfatovic/building-for-demand-the-growth-of-the-biodiversity-heritage-library [accessed 9 May 2014].

Lavoie C. (2013) Biological collections in an ever changing world: herbaria as tools for biogeographical and environmental studies. Perspectives in Plant Ecology, Evolution and Systematics 15: 68-76. http://dx.doi.org/10.1016/j.ppees.2012.10.002

Léonard J. (1994) Statistiques des Spermatophytes de la Flore d'Afrique centrale de 1940 à 1990. Bulletin du Jardin botanique national de Belgique 63: 181–194. http://dx.doi.org/10.2307/3668475

Loiselle B.A., Jørgensen P.M., Consiglio T., Jiménez, I., Blake J.G., Lohmann L.G., Montiel O.-M. (2008) Predicting species distributions from herbarium collections: does climate bias in collection sampling influence model outcomes? Journal of Biogeography 35: 105–116. http://dx.doi.org/10.1111/j.1365-2699.2007.01779.x

Meier R., Dikow T. (2004) Significance of specimen databases from taxonomic revisions for estimating and mapping the global species diversity of invertebrates and repatriating reliable specimen data. Conservation Biology 18: 478–488. http://dx.doi.org/10.1111/j.1523-1739.2004.00233.x

Mietchen D., Eckert S., Sierra S., Edmunds S., King D., Hagedorn G., Güntsch A., Müller A., Groom Q., Patterson D., Penev L., Maloney C., Catapano T. (2014) Report on progress during the coordination process of partners and non consortium partners. pro-iBiosphere project, deliverable D3.3.2. Call FP7-Infrastructures - 2012-1 [online]. Available from http://wiki.pro-ibiosphere.eu/w/media/2/2d/Pro-iBiosphere_WP3_MfN_D3.3.2_VFF30042014.pdf [accessed 1 Dec. 2014].

Nelson W.A, Dalen J., Neill K.F. (2013) Insights from natural history collections: analysing the New Zealand macroalgal flora using herbarium data. PhytoKeys 30: 1–21. http://dx.doi.org/10.3897/phytokeys.30.5889

New York Botanical Garden (2012) Index Herbariorum: A global directory of public herbaria and associated staff [online]. Available from http://sweetgum.nybg.org/ih/ [accessed 9 May 2014].

Nicolson D.H. (1991) A history of botanical nomenclature. Annals of the Missouri Botanical Garden 78: 33–56. http://dx.doi.org/10.2307/2399589

Otegui J., Ariño A.H., Chavan V., Gaiji S. (2013) On the dates of the GBIF mobilised primary biodiversity data records. Biodiversity Informatics 8: 173–184. http://dx.doi.org/10.17161/bi.v8i2.4125

Rich T.C.G. (2006) Floristic changes in vascular plants in the British Isles: geographical and temporal variation in botanical activity 1836–1988. Botanical Journal of the Linnean Society 152: 303–330. http://dx.doi.org/10.1111/j.1095-8339.2006.00575.x

Rios N.E., Bart H.L. (2010) GEOLocate (Version 3.22) [Computer software]. Belle Chasse, LA, Tulane University Museum of Natural History.

Raven C.E. (1947) English Naturalists from Neckam to Ray; a study of the making of the modern world. New York, Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511711152

Sautter G., Böhm K., Agosti D. (2007) Semi-automated XML markup of biosystematic legacy literature with the GoldenGATE Editor [online]. In: Altman R.B, Dunker A.K., Altman R., Hunter L. (eds) Biocomputing 2007 - Proceedings of the Pacific Symposium: 391–402. Stanford, World Scientific Press. Available from http://psb.stanford.edu/psb-online/proceedings/psb07/sautter.pdf [accessed 1 Dec. 2014].

Swan G.A. (1993) Flora of Northumberland. Newcastle upon Tyne, Natural History Society of Northumbria.

ter Steege H., Haripersaud P.P., Bánki O.S., Schieving F. (2011) A model of botanical collectors' behavior in the field: never the same species twice. American Journal of Botany 98: 31–37. http://dx.doi.org/10.3732/ajb.1000215

Britten J. (ed.) (1881) The names of herbes by William Turner, A.D. 1548. London, Y.N. Trübner.

Various editors (1948–1971) Flore du Congo, du Rwanda et du Burundi. Bruxelles, INEAC & Jardin botanique de l'Etat à Bruxelles.

Various editors (1973–2005) Flore d'Afrique centrale. Meise, Jardin botanique national de Belgique.

Whewell W. (1839) Address to the Geological Society, delivered at the anniversary, on the 15th of February, 1839. Proceedings of the Geological Society of London 3: 61–98. Available from http://biodiversitylibrary.org/page/30894466 [accessed 1 Dec. 2014].

Winch N.J. (1838) Flora of Northumberland and Durham. Newcastle, T. and J. Hodgson. Transactions of the Natural History Society of Northumberland, Durham, and Newcastle upon Tyne 2: 1–149. Available from http://biodiversitylibrary.org/page/33669541 [accessed 1 Dec. 2014].

Winch N.J. (2014) Flora of Northumberland and Durham. Reedition by Groom Q. (ed.) Advanced Books: e4002. Sofia, Pensoft. http://dx.doi.org/10.3897/ab.e4002

Manuscript received 11 Aug. 2014; accepted in revised version 1 Dec. 2014.

Communicating Editor: Elmar Robbrecht.