OXFORD UNIVERSITY PRESS | Digital Scholarship in the Humanities

Uncovering Environmental Change in the English Lake District: Using Computational Techniques to Trace Shifting Practice in the Historical Documentation of Flora

Journal:	*Digital Scholarship in the Humanities*
Manuscript ID	DSH-2019-0078.R3
Manuscript Type:	Full Paper
Date Submitted by the Author:	11-Jun-2020

Complete List of Authors:	Smail, Robert; Lancaster University, Lancaster Environment Centre Donaldson, Chris; Lancaster University, History Govaerts, Rafaël; Royal Botanic Gardens Kew, Biodiversity Informatics and Spatial Analysis Rayson, Paul; Lancaster University, Computing Stevens, Carly; Lancaster University, Lancaster Environment Centre

Keywords:	Floristic change, Historical flora, Geoparsing, Digital humanities, Corpus linguistics, Geographic information systems (GIS), English Lake District, Natural Language Processing (NLP), Plants of the World Online (POWO)

SCHOLARONE™ Manuscripts

Digital Scholarship in the Humanities

Uncovering Environmental Change in the English Lake District: Using Computational Techniques to Trace the Presence and Documentation of Historical Flora

1 Introduction

The English Lake District has long been regarded as a place of outstanding cultural and environmental importance. Situated in North West England, the region has been described and celebrated by naturalists, poets and painters since the late 1600s and ranks among one of the most iconic upland landscapes in Europe. In 2017 the Lake District became a UNESCO World Heritage Site (WHS) under the 'cultural landscape' category. UNESCO introduced its 'cultural landscape' category in the 1990s in order to establish guidelines for acknowledging places whose 'outstanding universal value' derives from 'the combined work of nature and mankind' (Denyer, 2016; UNESCO, 2017). The official designation of the Lake District as a cultural landscape has raised questions about how the character of the region has been shaped by natural processes and human industry. These questions are of interest for historians and environmental scientists, but they are also important for heritage and conservation organisations in the region, including the National Trust, Natural England and the Lake District National Park Authority, who are under mounting pressure to preserve, to protect and (in some cases) to restore the Lake District's historical environmental character. Answering these questions is not easy, not least because we lack a sufficiently coherent body of empirical evidence about environmental conditions in the Lake District before the mid-twentieth century (Lake District National Park, n.d.; UNESCO, 2017).

For Peer Review

The root of the problem is not a lack of evidence, but instead the need for a methodology that can enable researchers to compile and analyse evidence from a body of otherwise disparate historical sources. In this article, we demonstrate the implementation of such a methodology. Using a combination of computational techniques, we show how it is possible to consolidate and interrogate a diverse assortment of texts that provide evidence about the environmental character of the Lake District from the seventeenth century to the twentieth century. These texts contain a wealth of information about the region's historical environment, including accounts of flora, fauna and weather conditions. For the purposes of this study, we have chosen to focus on information relating to the region's flora, and we have done so for two reasons: firstly, because the Lake District's flora has received particular and sustained interest historically, meaning that there is an extensive body of empirical evidence dating back to the seventeenth century on which one can draw; secondly, we have selected flora because it is a good indicator of broader environmental changes (Ellenberg 1974). By tracing changes in plant species distribution and composition, one can also trace the transformation of habitats and landscape characteristics more generally.

The method we present combines techniques from the digital and spatial humanities, including Natural Language Processing (NLP), Named Entity Recognition (NER), corpus linguistics (CL) and Geographic Information Systems (GIS). These techniques have been shown to be effective in guiding the investigation and interrogation of geospatial themes across historical textual corpora (Donaldson, Gregory, & Taylor, 2017; Gregory & Donaldson, 2016). A few pioneering research projects have, moreover, indicated the benefits of applying these sorts of techniques for environmental history (Cervera, Pino, Marull, Padró, & Tello, 2019; Hinrichs et al., 2015), and there have recently been comparable developments in the environmental sciences (Roll, Correia, & Berger-Tal, 2018).

As Kherai and Oosthoek note, however, this is still 'an emerging area' (Kheraj & Oosthoek, 2016, p. 245), and as yet few studies have attempted to use computational methods to analyse source materials drawn from across the sort of broad span of time that we are concerned with in this study. We are therefore keen to take this opportunity to advance environmental research, and in the following pages we do so by demonstrating three things: firstly, how information about the Lake District's flora can be identified across a range of digitised historical sources on a large scale; secondly, how this information can be extracted and transformed into a structured dataset; thirdly, how this dataset can be plotted and analysed geospatially using digital tools including GIS. The method we outline here can be adapted to support research into the environmental history of other locations.

The historical texts we shall consider constitute a corpus of 92 digitised works published between 1682 and 1904 (a list of these works appears in the appendix). Collectively, these texts total nearly 19 million words and include examples of genres ranging from regional guides and travelogues to scientific journal articles and reports. The historical breadth and variety of the corpus is deliberate, as working with both non-scientific and scientific writing enables one to conduct more historically nuanced analyses. Selecting texts published between the late seventeenth century and the early twentieth century was also deliberate, as it enables us to study the composition of the Lake District's flora over a much longer time frame than current datasets allow. We chose not to include works published after 1904 due to UK copyright laws, which constrain our ability to work freely with some more recent sources. All the texts in the corpus have been compiled from existing open-source material, which collectively presents a uniquely rich and diverse body of information about the Lake District's environmental history.

For Peer Review

Our analysis will focus on how our computational methodology can enable researchers to track changes in the language used to describe and document plant species across the corpus and therefore across time. Plant taxonomy and nomenclature underwent sustained change during the period represented in our corpus. To uncover broader themes collectively registered across the corpus, it is essential that our computational methodology can trace variations and changes in naming conventions. As part of our investigation, we have drawn on the *Plants of the World Online* (POWO) database (POWO, 2018). Based on the extensive plant collections of the Royal Botanical Gardens, Kew, POWO is one of the world's most extensive databases of historical binomial plant name synonyms. Our use of this database reveals the importance of adopting a temporally sensitive approach that allows for synonyms to be evaluated in tandem with modern standardised plant names.

The methodology we outline and the dataset we derived from the corpus contains information about 802 plant species, 510 (63.5%) of which are linked to locations within the boundary of the Lake District National Park. This boundary, although not established until 1951, broadly corresponds with both the principal area of interest documented by the texts in the corpus and the area in which the members of the Lake District World Heritage Site Partnership (LDWHSP) have a vested interest. ~~In tracing the evolving language used to describe plants we have been able to form a richer dataset that more accurately reflects the historical observations recorded across our corpus. By uncovering changes in plant naming conventions over time, we have been able to increase our understanding of where different plant species were observed historically and also how they were documented in the past.~~ By extracting information about plant species from the corpus and collating that information in a structured and accessible digital form, this work stands to make an important contribution to academic researchers investigating the historical distribution and composition of plant

For Peer Review

species in the Lake District. We have also increased and enriched the historical knowledge available to organisations, including members of the LDWHSP, who are directly involved in landscape management and policy decisions in the region.

2 Background

2.1 The natural environment as recorded in historical accounts

For Peer Review

The flora of the Lake District has been recorded with increasing enthusiasm and dedication since the seventeenth century. Naturalists including John Ray (1627–1705), Archbishop William Nicholson (1655–1727) and Thomas Lawson (c.1630–1691) are credited with producing some of the earliest records of Lake District flora (Arber, 1943; E. Jean Whittaker, 1981; Hodgson & Goodchild, 1898, p. xxiii–xxiv). Over the course of the eighteenth and nineteenth centuries, interest in uncovering and recording the Lake District's natural environment grew as the region attained greater renown (Denyer, 2016; Lindop, 2005; Nicholson, 1955). In addition to naturalists, the Lake District also attracted the attention of tourists, travellers, writers and poets, many of whom published accounts that contain information about the region's flora. Consequently, there is an extensive and wide-ranging body of source material on which to draw. This material provides a wealth of empirical information about historical flora in the Lake District. The problem, however, is that this material appears in a variety of forms, which makes it difficult to consult and to collate. To understand how these sources can be combined we first need to understand the different sorts of information being described and the varying ways it is presented.

To illustrate this point, consider briefly the species *Drosera anglica* (*English sundew* or *great sundew*) and its description in two texts. *Drosera anglica* is a species commonly found in the wetter parts of the region, including bogs and lake shores. The species has been declining in England due to drainage, eutrophication and peat extraction (Stace, 2005, p. 217). Describing the species in *Flora of The Lake District* (1885), John Baker writes:

Drosera anglica, Huds. (Great Sundew). Native.

Scottish type. Range i.

C[umberland]. Ullock Moss near Portinscale.—(W. Dickinson.) Helvellyn,—(J. Flintoft.) Moss at Grange, abundant.

(J. C. Melvill.) Seathwaite in Borrowdale.—(Miss Edmunds.) Side of Crummock.—(W; B. Waterfall.)

W[estmorland]. Foulshaw Moss and Brigstear Moss near Kendal, First recorded by Wilson.

L[ancashire]. Stickle Pike, Donnerdale.—(W. F. Miller.)

(Baker, 1885, p. 44)

For Peer Review

Baker (1834–1920) was a Fellow of the Royal Society and principal assistant to the Keeper of the Kew Herbarium (he became Keeper in 1890; see Desmond, 1977, p. 36). His *Flora* was the culmination of years of patient research, and it was intended for the botanical community ~~(for the development of regional floras; see, David E Allen, 2003, p. 271–280)~~, providing a detailed (if rather dry) account of the different species found in the region. Baker's *Flora* is essentially a reference work, it contains concise descriptions that feature specialised abbreviations and technical terminology. In this format, *Drosera anglica* is documented at eight different locations around the Lake District. In addition, the species is also noted as being 'native' to the region and growing at *range i.*, which means it was commonly observed to grow from sea level up to an altitude of 900 feet.

Contrast this with the description of *Drosera anglica* in Frederick Malleson's (1819–1897) *Holiday Studies of Wordsworth by Rivers, Woods and Alps*. Malleson writes:

From the [Ulpha] bridge the main road leads along the foot of the Dunnerdale Fells to Broughton and to Millom. The scenery, though grand and noble, is bare and wild. In the boggy streams running off these fells is found the great sundew (*Drosera anglica*), a rare plant. (Malleson, 1890, p. 68)

This account is rather different than Baker's. A minister in the Church of England, Malleson served as vicar of Broughton-in-Furness between 1870 and 1897. The observations just quoted were made during a two-day tour through the Duddon Valley in 1882. This tour, he tells us, was inspired by the writings of the Romantic poet William Wordsworth. Specifically, Malleson is referring to Wordsworth's River Duddon sonnets, which were published in 1820. Malleson's aim is to impart a vivid account of the valley's scenery: the surroundings are described as 'surpassingly beautiful', as both 'grand' and 'noble' though 'bare and wild'. A competent botanist, Malleson offers an account of *Drosera anglica* that is knit into his poetically inspired account of this landscape; flora is intrinsically linked to the emotional impressions the landscape makes upon him and which, in turn, he attempts to impress upon the reader. Though the account is quite different to Baker's (both in its purpose and its style), it too imparts some useful information about *Drosera anglica*. Malleson informs us of a locality in which the plant can be found. He also records its

For Peer Review

conservation status, telling us it is a 'rare' species, and he gives us some details about its habitat, noting that it can be found in 'boggy streams running off these fells'.

Despite their quite different outward appearances and intended audiences, both commentators provide insightful accounts of Lake District flora. Moving past differences in language, descriptive styles and layout, commonalities can be discerned regarding the flora being described. In addition to recording a specific plant species, these accounts indicate where those species can be found, as well as providing details about the plant's traits and habitat. There is therefore a clear benefit in bringing these two sources together and combining them with other accounts of *Drosera anglica*. Doing so enables us to build a more comprehensive picture of the historical presence and distribution of plant species in the Lake District.

For Peer Review

2.2 Correlating empirical evidence across historical sources

As we have seen in the previous section, historical sources can provide a rich documentary record of Lake District flora and its distribution. Taken in isolation, individual accounts give us glimpses into the activities and interests of specific individuals, furnishing us with a snapshot of flora seen at specific places and points in time. ~~In order t~~To ~~improve further~~advance our understanding of the region's historical environmental character, and to begin to examine changes, it is necessary to combine and correlate evidence from multiple sources. Consequently, it is important that the sources collectively provide sufficient geographic and temporal coverage. By geographic coverage, we refer to how observations of a specific site can be iteratively combined with observations of other sites to enable one to *construct* a picture of the whole Lake District. Similarly, by temporal coverage, we refer to how accounts of specific sites made at different points in time can be compared to determine if and when local environmental conditions changed.

As previously noted, Lake District flora has received interest from naturalists, plant collectors, tourists, travellers, writers and poets alike. The individuals who constituted these groups often had different motivations for exploring and recording the region's flora. These motivations influenced the aspects of the natural environment they investigated and recorded, including which plants they documented and which they overlooked. Even observations collected by the most studious scientific researcher are still likely to contain some degree of bias, reflecting the time spent observing a particular site, the skill level of the observer and their specific interests, as well as the standards and conventions of scientific practice at that time (this is a persisting issue in botanical field research; see, Rich, 1997). Drawing from multiple 'witness groups', as they will be termed here, serves to elucidate the interests of each group and to provide a sense of perspective. It is also essential if we are to reveal a broader narrative of historical Lake District flora that is not limited to a single group.

How might we expect observational records to vary between different witness groups? Genre aside, specialisation is one obvious cause of variation. Here, we do not use the term in its more rigid sense to mean specialisation between scientific disciplines, as it has been shown that the boundaries between disciplines in natural history were never to become rigidly demarcated (David Elliston Allen, 1994; Kuklick & Kohler, 1996; Nicholas Jardine, James A. Secord, 1996). Rather, specialisation refers here to the particular interests of each observer: that which led them to favour the recording of specific taxonomic groups, habitat types or geographical localities over others. One might turn to the species *Impatiens noli-tangere* (*Touch-me-not-balsam*) to elaborate this point*. Impatiens noli-tangere* is rare in the UK, but it is relatively common in the Lake District*.* Some observers might overlook this

For Peer Review

species in favour of rarer or more distinctive ones. For the lepidopterologist, however, *Impatiens noli-tangere* is a species of interest since it is the sole food source of the rare moth *Eustroma reticulate* (*Netted carpet moth*). Within the discipline of Lepidopterology, then, places where *Impatiens noli-tangere* grew attracted particular attention.

Other witness groups including tourists, writers, travellers and poets approached the region from yet different perspectives. Though the individuals affiliated with these groups often made detailed observations, on the whole they were less interested in whether a particular plant species belonged to a particular taxonomic group or habitat type than they were in describing some facet of the landscape that surrounded them. Writers intent on revealing the region's aesthetic qualities were, for example, more likely to focus their attention on a plant or combination of plants that helped to define the landscape they were trying to understand and describe. Consider Elizabeth Lynn Linton's *The Lake Country*. Linton tells us her aim in writing this work was to 'illustrate and describe the most beautiful places—both those popularly known, and those which only the residents ever find out.' (Linton, 1864, p. x). Describing the walk up from Ambleside to the celebrated waterfall of Stockghyll Force, it is the dramatic qualities of the locality that Linton attempts to impress upon her reader. She writes:

> following the wild path of rock and running water and twisted tree-roots—the rocks below getting larger and more broken, the rift between them deeper and shaper—the roar of the waters loader, and the rush more fierce and rapid … there you come upon the "loosening silver" of the fall (Linton, 1864, p. 18-19)

For Peer Review

At this point, Linton uses a reference to the species *Pyrola media* to appeal to the reader's imagination. She observes that, 'for those who have stout nerves … the *Pyrola media*, a rare growth of the Winter-green, [may be] found only among the rocks in the centre of the fall' (Linton, 1864, p. 18-19). Linton's decision to include these observations indicates how her interest in flora was intrinsically linked with her interest in the rugged landscape of the region. The scarcity of the plant imbues the locality with a heightened sense of uniqueness, while the difficulties involved in gaining a glimpse of it helps impress the dramatic landscape of the waterfall.

Combining these sorts of discreet observations from texts that span the entire timespan and geographic extent under investigation allows us to uncover a broader picture of the region's flora as a whole. This picture becomes even richer and more nuanced when observations from multiple witness groups, including guidebook writers, are included.

2.3 Using digitised material and source selection

The texts that form the corpus have been digitised using automated Optical Character Recognition (OCR) software. This software uses NLP methodologies to recognise letters and words from scanned images (Bennamoun & Mamic, 2012, p. 199-220). Automated OCR digitisation is conventionally less expensive and time-consuming than manual digitisation, however in digitising historical texts its accuracy has been shown to be variable, with a percentage of characters and words being misidentified (Blanke, Bryant, & Hedges, 2012; Tanner, Muñoz, & Ros, 2009). One way of dealing with OCR error is to use fuzzy matching. Fuzzy matching relies upon the use of edit-distance to measure the similarity between two words (Tanner et al., 2009). ~~This measure is frequently represented as a quantitative metric~~

~~such as 0–100, where 100 represents an exact match.~~ Specifying a similarity threshold allows for words that are not an exact match, due to OCR error, to be matched provided their similarity falls above a specified quantitate threshold. However, for fuzzy matching to be effective the threshold needs careful consideration. Setting a threshold that is too high is unlikely to improve match rates significantly, while setting a threshold that is too low can result in false positives. Achieving an effective balance is especially difficult when OCR accuracy varies across a corpus, and for this reason it was not used in this study (Amelia, 2017; Gregory et al., 2016; Tanner et al., 2009).

Our decision to use OCR digitised texts, despite potential variations in accuracy, is three-fold. Firstly, the manual digitisation of almost one hundred texts was beyond the resources of this project. Secondly, a large amount of historical textual source material has already been digitised at considerable expense to both the public and the private sector, and this makes it desirable to explore the potential of this material and to assess its limitations before investing in costly re-digitisation projects. Thirdly, we thought it important to determine whether our developing methodology was robust enough to cope with 'noisy' corpora.

In the selection of texts, our intention was to form a corpus that reflected the broad range of literature documenting Lake District flora. To achieve this several factors were taken into consideration. These factors included the date of publication, geographical focus and the text's genre. Coverage of a broad time frame is essential if changes in plant distribution are to be examined historically over time. The corpus therefore spans more than two hundred years (1682–1904). Texts that focused on the Lake District or North West England more generally were favoured, but texts covering a wider geographical area, such as John Hull's *The British Flora* (1799), were also admitted if they were deemed of special significance to the recording of flora at a specific point in time.

For Peer Review

Making selections based on genre proved especially challenging. As previously noted, capturing the observations of different witness groups is essential for forming a more comprehensive picture of plant distribution in the region. However, these observations were often found to extend across several textual genres. Whereas the observations of the national scientific elite are most commonly found in botanical floras, specialist botanical journals and the reports and transactions of scientific societies, the observations of amateur collectors and enthusiasts are more commonly found in published diaries and travelogues. Consequently, we determined that it was necessary to include regional and national botanical floras, botanical journals, scientific society reports and transactions as well as periodicals, collected letters, diaries, botanical handbooks, travelogues and Lake District tourist literature.

2.4 Historic name variations: plant species synonyms

In order to examine the distribution of flora recorded in a corpus of historical texts, it is necessary to identify every plant name recorded in that corpus. This is made more challenging as the scientific names of plants underwent substantial and sustained change during our period of investigation. Many plant species have been recorded under different names at different points in time and some plants have been recorded under several different names at the same point in time. Changing plant taxonomy was one reason for name variation. During the eighteenth and nineteenth centuries plant nomenclature was considered intrinsically linked to taxonomic classification, with a plant's name serving to both identify and distinguish it as a unique 'species', while at the same time linking that plant to other plants with similar traits (Pickstone, 2001, p. 71; Sanderson, 2017). There was,

For Peer Review

however, considerable debate over what traits should be considered the most important basis for grouping plants together, and by the mid-eighteenth century several rival taxonomic systems had been proposed, each setting out different naming principles as part of their classification system (David Elliston Allen, 1994; Gledhill, 2002; Scharf, 2009).

Even within a single taxonomy, it was not uncommon for plant names to change over time. The identification of new species collected from around the globe frequently required plants to be re-grouped and re-named to incorporate new knowledge. ~~In these cases, plant names were sometimes also changed to reflect the re-grouping~~ (David Elliston Allen, 1994; Scharf, 2009). A requirement of scientific naming systems is that each plant has a single 'accepted' name. However, it was frequently the case that plants collected and named in different localities were later found to be in fact the same species. In these instances, one name was selected as the accepted name and the other names were dropped. In both situations the adoption of a single name was often a gradual process, and it was not uncommon for naturalists in one social group or geographical region to continue using plant names that had been rejected many years before by other groups or in other regions.

~~Tracking historical changes in naming conventions across our corpus posed challenges for even state-of-the-art NLP methodologies. Conventionally, w~~When working with modern corpora, one can employ NER techniques that identify and extract different types of named entities by automatically comparing the contents of a corpus to a designated search list. However, these lists normally use modern naming conventions and this makes them less adept at detecting historic name variations (Butler, Donaldson, Taylor, & Gregory, 2017). In order to maintain the accuracy of NER techniques when working with historical texts, it is necessary to develop naming inventories that map modern names to any known historic variations. With this in mind, we compiled a list of historical plant synonym names

For Peer Review

using the POWO database, which lists both currently accepted plant names alongside documented alternative synonym names (POWO, 2018).

~~In total POWO contains over 982,000 plant names. In spite its size though, the database is not an exhaustive list. Certain plant groups are still being evaluated and therefore are not yet available online (POWO, 2018). This limitation is not an issue for this study, however, as all the plants selected to form our search list were present in the POWO database. Furthermore, T~~the database primarily contains synonyms formed within the binomial paradigm established by the Swedish naturalist Carl Linnaeus (1707–1778) during the mid-eighteenth century. Before Linnaeus, naturalists had tended to use 'diagnostic phrase names', which also served as a brief description of the plant and its traits (these names could be very long, sometimes up to half a page in length; see, Gledhill, 2002; Koerner, 1996, p. 149; Ogilvie, 2008; Reddy, 2007). As a result, most synonyms listed within POWO date from the mid-eighteenth century onwards. Consequently, it is likely that POWO will be less effective at identifying plant synonym names in texts published before the 1750s. In order to determine the extent to which this was the case, we analysed the plant name match rates per text to reveal if any decrease could be discerned across the corpus.

3 Methodology

3.1 Forming the historical plant list

In order to compile a historically sensitive plant search list, we first needed to draw together a list of currently accepted plants relevant to the Lake District. We completed this task by conducting a polygon query of the National Biodiversity Network (NBN Atlas) to find all plant species listed as present in the Lake District (Atlas, 2019). The NBN Atlas is an open-source

For Peer Review

online resource that provides access to extensive species records. It is linked to several extensive databases, including the Botanical Society of Britain & Ireland (BSBI) and the Biological Records Centre (BRC) (Atlas, 2019). We decided to focus on high-level species, as these species are often more conspicuous in the landscape and are therefore more likely to have been observed and recorded historically. As a result, all mosses, algae and liverwort were filtered out of the downloaded dataset, leaving vascular plants and ferns. We then compared this list with POWO, compiled the historical synonyms and ~~and~~ 'mapped' them to their modern names. The final list used to search the corpus contains 952 currently accepted plant species to which 9340 synonyms were linked (~~the plant search list, alongside the corpus and Python scripts have all be made available on GitHub and University online repository - URL~~ reference anonymised for review). Within the compiled list a number of historical synonyms were linked to two or more currently accepted names. As this could cause potential ambiguity in the analysis of the results, duplicate names were filtered out if matched in the corpus.

3.2 Forming and formatting the corpus

The corpus was formed entirely of existing digitised texts downloaded from open-source online repositories including Archive.org, Google Books and BioDiversity.org (~~URL~~ reference anonymised for review). Once all the texts had been downloaded, the first stage was to standardise and format the corpus. Non UTF-8 characters were removed, and words split over two lines were re-joined. Both processes were preformed using an automated approach with regular expressions in Python 2.7 (~~URL~~ reference anonymised for review).

For Peer Review

3.3 Extracting species and geographical locations from the corpus

To improve efficiency, we located plant species and place names within the corpus in two phases. Firstly, we used a keyword search method to identify plant species named in the corpus, including synonym variations and abbreviated forms. For certain search terms, particularly those where a word can have multiple meanings, keyword searching can lead to misleading results. However, the complex structure of plant names (which are formed of two Latin words) and their appearance in texts principally written in English, helped to reduce ambiguity between plant names and other word tokens. By manually checking a random sample of 500 match instances, we found that only 1.8% of the sample was misidentified. To support subsequent analyses of the data, all plant name match instances were then mapped to their current accepted name and a span of text (or co-text) adjacent to each match was extracted. The process was performed using an automated approach, with ~~the~~ scripts written and run in Python 2.7. The extracted text was then geoparsed using the Edinburgh Geoparser (Grover et al., 2010; Tobin, Grover, Byrne, Reid, & Walsh, 2010). This process enabled the automated identification and georeferencing of place name entities across text, which is a prerequisite for performing geospatial analysis using GIS software. Geoparsing the extracted co-text, and not the whole corpus, improves efficiency as only the relevant sections are examined (Rupp et al., 2015). This is especially valuable when working with the more general natural history texts, which are likely to contain numerous place names related to other research areas in the environmental sciences.

Place names can have a variety of potential meanings, and they are thus more likely to be affected by issues of ambiguity than plant names. As Rupp et. al. have noted, the place name Lancaster can be a town, but it can also be the name of a person (Stuart

For Peer Review

Lancaster) or even an honorary title (the Duke of Lancaster). Furthermore, a single place name can refer to multiple localities: there are settlements named Lancaster in England, the USA and South Africa (~~Donaldson et al., 2017, p. 47;~~ Rupp et al., 2015). To mitigate the potential errors introduced by these ambiguities, the Edinburgh Geoparser includes two interlinked components: a 'geo-tagger' and a 'geo-resolver'. The geo-tagger runs through a sequence of NLP analysis steps (including tokenization, sentence splitting, POS tagging, chunking and a rule-based named entity recogniser) to identify place names within the text and to disambiguate them from other word token types (Grover et al., 2010; Tobin et al., 2010). The geo-resolver then attempts to assign a pair of geographical coordinates to each identified place name, using a ranking algorithm that considers population size and the geospatial relation of the place to others in the document. This algorithm gives preference to places that cluster with other locations in the same document (Grover et al., 2010; Tobin et al., 2010). Assessing the impact of these heuristics on the output can be difficult. However, manual checking of the output indicates that the ranking of locations in our corpus was more frequently influenced by the geographical clustering of place names than by population size, which reflects the regional focus of many of the texts in the corpus. In addition to these metrics, it is possible to add a weighting to a particular geographical area when the geographical locality under investigation is already known (Grover et al., 2010; Tobin et al., 2010). For this study, the geoparser was used in conjunction with the Ordnance Survey 1:50,000 scale gazetteer, as this gazetteer provided the best coverage for the Lake District, and a weighting of two was added for locations which fell inside the North West of England ("OS Open Data," n.d.).

3.4 Determining plant and location collocates

For Peer Review

Collocate analysis was used to trace the recorded localities of plants across the corpus. An established analytical method within the fields of lexicography and corpus linguistics, collocate analysis can be used to identify automatically every time a pair of co-occurring search terms appear in close proximity to one another within a text. As a collocate is here determined by the proximity between terms within a text, the span of the collocate 'window' is of some significance: if the span is too narrow, place names associated with the search term may be missed; if the span is too wide, place names are more likely to be erroneously linked to the search term. Commonly, a collocate window is of a fixed size. In this study, however, our initial experiments using a fixed size collocate window led to place names being erroneously linked to plant names. Manual checking of these errors suggested that they were principally a result of the very 'compact' format of the entries in many botanical works, which frequently list information about flora and their geographical locations in a compressed space (an example of which is shown in Fig. 1). In these instances, we found that a fixed window would extend across multiple plant species match instances, leading to false positives between plant and place. To overcome this problem, we adopted a 'dynamic' collocate window. This was done by identifying all plant species matches and their abbreviated forms across the whole corpus and setting a collocate boundary that extended either up to 300 characters to the right of each match or up to the point where the next plant species match occurred.

The direction of the span of the collocate window was also taken into consideration. A span can be made to the left or right of each plant name in order to include the words that co-occur immediately before and after each match. Close evaluation of the initial output of the collocate analysis revealed that including a span to the left of the matched plant names

For Peer Review

increased the number of false positives. Inspection of these false positives suggests that they were a consequence of the compact format of many of the texts in the corpus and of the tendency of the texts in the corpus to record the location where a species was observed after naming the species itself. Accordingly, we decided to extend the collocate window only to the right of each match instance, which improved the overall accuracy of the results. This decision means that our results do not include instances where a location is mentioned before a plant name. Given the early stage of this research we felt that prioritising the accuracy of the results over the recall was justified, as it gives more weight to the potential value of the generated dataset. We are confident that it should be possible to augment our analyses once a methodology for accurately identifying collocate pairings with place names mentioned before match instances has been developed.

For Peer Review

4 Results and Findings

4.1 The impact of historical synonyms on match instances

Before we proceed to examine the geographies of historical Lake District flora that the collocate analysis revealed, it is helpful first to consider the plant species matches and how searching for historical synonyms influenced the results. Searching the corpus for modern accepted plant names resulted in a total of 16216 match instances. The number of match instances increased to 22659 when the search was expanded to include both accepted plant names and their historical synonyms. We can account for this marked increase of 28% in two ways. Firstly, the number of times a species was matched in the corpus can be seen to have increased. When we searched the corpus for only modern accepted plants names, 673 species were matched at an average of 24.1 match instances per species. When synonyms were included in the search, the matches for these 673 species increased to 28.8 match instances per species. Secondly, the number of different species being matched also increased.

As just noted, 673 of the 952 plant species listed in the search list were matched in the corpus when we searched for modern accepted names. The total increased to 802 species matches when synonyms were added to the search list. This increase suggests that many modern accepted species names were not in use during the timeframe under investigation. Although further fine-grained analysis would be required to reveal if there are any patterns regarding plant name changes, these results do demonstrate the improvements in match instances that can be gained by using temporally sensitive search lists when examining flora across historical source material. Through the results we can see an increase not only in the number of times a plant species was matched across the corpus, but also in the number of different unique species matched. More complete identification of plant species across the whole corpus is critical if subsequent collocate analysis and geoparsing are to provide an accurate picture of the geographies of historical Lake District flora.

4.2 The impact of plant recording practices on match rate

The identification of historical synonyms can also assist in the investigation of observer practices. ~~Plant synonyms account for 6494 of the total match instances.~~ Fig. 2, created using the pyplotlib library in Python 2.7, shows the ratio between the number of times a plant species was matched in the corpus under a historical synonym name and the number of times it was matched under its modern accepted name. As is evident in Fig. 2, just over half of the plants matched across the corpus were recorded under a single name: 367 were

For Peer Review

recorded under only their modern accepted name, whereas 129 were recorded under a single historical synonym name. This finding indicates that the use of many plant names was relatively stable across the time period under investigation. This result surprises, especially given the heterogeneity of the corpus and the span of time it represents.

For plants matched under two or more names, the transition from one name to another can be used to explore evolving naming conventions and, more broadly, changes in recorder habits. On the one hand, a smooth shift from one name variation to another might tend to suggest a smooth transition in the modification of plant-naming habits, with a new name being suggested and adopted by the botanical community over a relatively short span of time. On the other hand, the use of several variant plant names over a sustained period suggests a more staggered transition between names, either because some observers were not aware a new name had been proposed or because they resisted adopting the new name. Both trends can be detected in the 306 species recorded under two or more names.

Consider the species *Platanthera bifolia* (*lesser butterfly-orchid*) and *Ranunculus aquatilis* (*Water Crowfoot*). Both are documented in modern floras as having taxonomies that remained unresolved until the second half of the twentieth century (Preston, Pearman, Dines, & others, 2002; Stace, 2010). Our findings extend this understanding, revealing distinct differences in the trajectory of naming conventions for these two species over the past two hundred years. *Platanthera bifolia* was matched 67 times across 34 texts, spanning the entire timeframe of the corpus. The earliest text in which the species was recorded was Volume 2 of Ray's *Historia Plantarum* (1688); the latest text was Crump and Crossland's *Flora of the Parish of Halifax* (1904). In total, the species was recorded under five different names: *Orchis bifolia*, *Orchis alba*, *Habenaria bifolia*, *Platanthera bifolia* and *Gymnadenia bifolia.* Despite being recorded under a variety of different names, there is a discernible

For Peer Review

pattern in the use of these names over time. From the 1680s until around the mid-eighteenth century the plant was matched under the names *Orchis bifolia* and *Orchis alba*. After this period, *Orchis alba* no longer appears in the matches and the species is instead most commonly recorded under the name *Orchis bifolia,* as well as *Platanthera bifolia*, *Gymnadenia bifolia* and *Habenaria bifolia.* This last name was first matched in Volume 4 of Smith's *English Flora* (1828) and appears with increasing regularity until the mid-nineteenth century, after which point it becomes the only name under which the species is recorded.

With *Ranunculus aquatilis* the situation is rather different. This species was matched 71 times across 31 texts in the corpus. As with *Platanthera bifolia*, the match instances for *Ranunculus aquatilis* occured across nearly the entire timeframe of the corpus, with the earliest text being Volume 1 of Ray's *Historia Plantarum* (1686) and the latest text being an edition of *The Naturalist* from 1893. Over the course of the 207 years that separate these texts, *Ranunculus aquatilis* was recorded under six different synonyms: *Ranunculus aquatilis, Ranunculus aquaticus, Batrachium heterophyllum, Ranunculus heterophyllus, Ranunculus hydrocharis* and *Ranunculus diversifolius*. However, the transition between these synonyms is not as smooth as with *Platanthera bifolia.* Instead, the synonyms under which *Ranunculus aquatilis* is recorded appear to have been used interchangeably. The name *Ranunculus aquatilis* was first matched in 1686 and last matched in 1890. Similarly, the first match for the synonym *Ranunculus aquaticus* was in 1686 and the latest was in 1870. The other synonyms ~~(*Batrachium heterophyllum, Ranunculus heterophyllus, Ranunculus hydrocharis* and *Ranunculus diversifolius*)~~ were all matched over the course of the nineteenth century.

These findings provide greater insight into the different ways naming conventions evolved for these two species. Specifically, the findings indicate that new names proposed

For Peer Review

for *Platanthera bifolia* were accepted and adopted swiftly, whereas newly proposed names for *Ranunculus aquatilis* were taken up much more slowly, if at all. Identifying these trends provides opportunities for further analysis into the factors that influenced plant-naming conventions in the region. More immediately, these trends remind us that the texts in our corpus reflect historical scientific conventions as well as biases, and that these conventions and biases should be taken into account in the methods implemented in analysing the corpus.

4.3 Potential limitations of the computational methodology

Despite the apparent improvement in match instances when using historically sensitive search lists, one must still be cautious when analysing the results. Above all, one must be mindful that some historical name variations may still be missed, either as a result of spelling errors introduced during the OCR process or because the historical search list is still incomplete. Assessing where computational methods have failed to identify plant names is very challenging and requires further checks to be performed. One approach is to plot the results across the corpus and assess changes to the number of match instances.

Fig. 3, created using the pyplotlib library in Python 2.7, plots the match instances of accepted names (X) and accepted and synonym names combined (O). As already established, match instances are noticeably increased when historical synonyms are included in the search. However, Fig. 3 reveals a degree of variability in match rate per text across the corpus, with a greater number of matches being discernible between 1820 and 1890. Before 1800 the number of match instances per text decreases sharply. Given that these texts were selected on account of their relevance to the recording of Lake District flora,

For Peer Review

it seems unlikely that they do not contain any plant names. A more plausible explanation is that the plant names used in earlier texts in the corpus are not being detected. Using Fig. 3 as a guide, a detailed reading of the texts published around 1800 was performed to understand why there was such a sharp drop-off in results. This evaluation revealed that texts published before 1800 increasingly used non-Linnean taxonomic systems, such as those of John Gerard (1545–1612), Caspar Bauhin (1560, 1624), Joseph Pitton de Tournefort (1656–1708) and above all the English naturalist John Ray (1627–1705) (Charles, 1947; Ogilvie, 2008; Scharf, 2009). As a result, the names used to record plants in many of the texts published before 1800 predate the names compiled from POWO and were therefore being missed. Such omissions are noteworthy as they give us a better understanding of the adoption of Linnaean plants names in Britain.

A further disparity in the results was uncovered when the distribution of match instances, presented in Fig. 2, were considered in relation to witness groups. Here, match rates across travel accounts were found to be frequently lower than scientific floras and journals, as well as more general histories of the region. Closer inspection revealed that tourists and travellers to the region frequently documented the plants they observed using common names rather than scientific binomial names. Again, these names were not compiled from POWO and had therefore been missed. These findings are helpful in revealing not only how observing conventions varied over time and between social groups, but also where and why our methodology has failed to identify plant names. Such issues alert us to the allowances that need to be made in future analyses of the results. These findings also provide insight into how our methodology can be refined and improved going forwards. They highlight that the addition of pre-Linnean plant names and common plant names into our plant species search list would likely capture further information.

For Peer Review

4.4 Mapping extracted information

We shall now examine how searching for historical synonyms across the corpus affects the geographical distribution of the geoparsed results. The collocate results for modern accepted names returned a total of 1982 location matches, 515 of which were unique. This figure increased to 2569 total location matches and 576 unique locations when the modern accepted plant name collocates were combined with the synonym plant name collocates. Closer investigation of the geoparsed results revealed a discernible increase in both the composition of plant species that were geolocated and the locations to which they were linked. For the modern accepted plant name collocates, 400 individual plant species were linked to 515 different Lake District locations. This number increased to 510 plant species and 576 locations when the plant synonym collocates are added.

Assigning geographical coordinates to each location enables the results to be visualised. Fig. 4, created using ArcGIS 10, plots the results of the geoparsed modern plant name collocates against the combined modern and synonym collocates. This visualisation further exposes the geospatial differences that result from implementing the two search methodologies. Specifically, Fig. 4 helps reveal how both the *extent* (or, in other words, the geospatial distribution) and the *depth* (in other words, the number of plants linked to each location) improved when historical synonyms were included in the search list. Each dot on the map marks a location with which a plant species was collocated, with the size of the dot representing the number of times a plant was matched to that location. When one focuses on the extent of the geospatial dispersion of the results across the whole of the Lake District, the geoparsed modern name collocates and the geoparsed modern and synonym collocates

For Peer Review

appears to be similar. However, subtle shifts can be discerned when one focuses in on specific localities, such as the northwest of the Lake District. Here, place names including Cardurnock, Blackdyke, Wampool, Thurstonfield, Howrigg and Harker can all be seen to match with synonym names only. These place names stand out from other place name collocate match instances as they all received comparatively less attention across the corpus. Consequently, the impact of searching for historical synonyms alongside modern plant names is accentuated for these place names when the findings are mapped geospatially, as other place names with a higher number of collocate-match instances are more likely to collocated with modern plant names.

Linked with the extent of place names distribution, the depth of plant names collocated to each location also improved when historical synonyms were included in the search list. As can be seen in Fig. 4, the number of times plant names were linked to a particular site visibly increases in several localities across the region. This increase is especially evident around Buttermere and Derwent Water and around Coniston, Langdale and Hawkshead. This provides a more complete picture regarding the species composition in each locality; it also provides further insights into observer habits and the range of different people making these observations.

Evaluation of the impact of searching for historical synonyms on the match instances of individual plant species revealed that those plant species matched most frequently under synonym names had the most pronounced shifts in the geographical locations to which they were linked. For example, the species *Blechnum spicant* (*hard fern*) was matched 101 times across the corpus and 90 of these instances were under synonym names. During geoparsing, the species was collocated with 19 locations around the Lake District including Keswick, Buttermere, Grisedale Pike, Wasdale, Scale Force, Kirkstone Pass, Mosser,

For Peer Review

Ullock, Lamplugh, St. Bees Head, Glaramara, Styhead, Lingmell, Ponsonby, Bootle, Santon Bridge, Birks Wood, Ease Gill and Millom. Of all these locations, only the collocations with Bootle were matches with the modern accepted name of the species. The case of *Littorella uniflora* (*Shoreweed*) is even more extreme. This species was matched in the corpus 62 times, but only under the synonym *Littorella lacustris*. As a result, its collocation with Edenhall, Lodore, Blea Tarn, Styhead Tarn and Barrow-in-Furness would have been missed if we had only searched for its modern name. As these examples demonstrate, using modern plant species lists alone is likely to distort the results, giving an increased weighting to plant species whose modern accepted names remained more consistent throughout the period under investigation. Going forward, it will be necessary to assess the results of such analyses in greater detail in order to determine where any plants species have been incorrectly collocated to locations. This could be accomplished by reading a sample of the collocate windows to establish instances where plant names and place names have been misidentified. A further step would be to apply statistical measurements, such as Kulldorff's spatial scan statistic, to distinguish the degree to which the collocate pairings have been influenced by the underlying geography of the corpus (Kulldorff, 1997; Rupp et al., 2015). Using this sort of approach might reveal which collocates form clusters even when the underlying geographies of the text are taken into account (Rupp et al., 2014). This would provide confidence in the results and pave the way for further analysis of the data though the clustering of indicator species for different habitat types to assess changes in the observed environment over time. These clusters could in turn be used to help determine whether the discerned changes have been caused by human or natural events.

5 Conclusions

For Peer Review

This article has introduced a methodology that uses computational techniques to extract the geospatial information of Lake District flora from a corpus of disparate historical texts. Our findings reveal the potential of using historical sources to provide detailed empirical evidence regarding the historical distribution of different plant species across the Lake District. In total, 802 species were traced across the corpus and 510 of these species could be linked to locations around the Lake District. Re-organising the extracted information into a structured and searchable geo-temporal dataset enables further analysis; it also allows for the formation of advanced queries that use resources such as GIS to visualise and examine changes to plant species composition and distribution over time.

Our findings demonstrate the importance of developing temporally sensitive search lists in order to trace plant names accurately across a corpus that comprises texts from an array of genres and historical periods. Searching for historical synonym plant names alongside modern accepted names improved the number of unique plant species found across the corpus as well as the total number of finds. Focusing on the geoparsed results, the inclusion of plant name synonyms increased both the number of species that could be geolocated and the number of different locations to which they were geolocated. Consequently, our temporally sensitive methodology can be seen to have resulted in a more accurate reflection of the consulted corpus and to have reduced distortions that would have arisen were we to have only searched for modern accepted names.

But the inclusion of historical plant synonyms does more than simply improve the accuracy of the results. It also brings broader changes in plant classification systems and naming conventions into focus. The results indicate that many plant names are more stable than might initially be expected: 496 of the 802 plants found in the corpus matched under a

http://mc.manuscriptcentral.com/dsh

single name. It is important to bear in mind, however, that only 367 of the species that were matched under a single name were recorded under their modern accepted name. If the timeframe under investigation was extended farther into the twentieth century, it is likely that greater shifts in plant names would be observed. Closer investigation of the results revealed earlier texts were more likely to use pre-Linnean plant names, whereas travellers and tourists were more likely to use common plant names. Including these names in our search plant list would not only extend the dataset, but also reveal further shifts in naming conventions.

For the 306 plants species matched under two or more names, the transition from one name to the next across the corpus provides insight into the introduction and adoption of new names by those making observations on Lake District flora. In this case, some plant names were replaced by another relatively swiftly, which indicates that the new name was accepted and adopted with little resistance. Other names, however, appear to have been adopted less uniformly, with multiple plant names continuing to be used for a sustained period. These findings open up new lines of research into changing naming practices of Lake District flora and why some plant names were adopted more swiftly than others.

Even though searching for historical synonyms alongside modern accepted names improved results, other factors should not be ignored as a decline in match instances is still discernible in the results. In particular, the lower match rate before the year 1800 indicates that the Linnaean binomial naming system was only adopted in Britain at the turn of the nineteenth century. One option to overcome this limitation would be to extend the plant-naming lists by adding pre-Linnaean plant names. Furthermore, errors introduced by OCR also resulted in some plant and place names collocations being missed. A possible way to

For Peer Review

address such errors in further investigations of the corpus would be to draw upon emerging Machine Learning approaches to help identify plant and place names across the corpus.

Notwithstanding these limitations, the findings presented in this paper provides researchers with a firmer empirical grounding for making accurate assessments of historical flora. The structured digital dataset we have created from an otherwise disparate assemblage of texts is a composite knowledge base which can be used to assess broader floristic changes and to guide and complement more detailed analysis of specific sources. In its digital form the dataset can be easily shared, manipulated, queried and visualised geospatially; it can therefore be explored and exploited by a range of audiences to suit different interests and needs. These audiences include: academic researchers such as environmental historians and scientists, who have previously struggled to interrogate a wide range of historical accounts; heritage and conservation organisations including members of the LDWHSP, who are responsible for protecting and preserving the region; and members of the general public, who want to know more about the history of the Lake District's environment. It is hoped, moreover, that our computational methodology might be extended beyond the Lake District, taking in an even broader range of digitised historical texts to support research into the environmental history of other localities and natural features. Collectively, such research could yield a much more detailed and comprehensive understanding of the earth's environmental past.

For Peer Review

For Peer Review

Figure Captions

Fig. 1. Excerpt from Winch's *Flora of the Lake District* (Winch, 1825, p. 27)

Fig. 2. Ratio (%) of plant species being matched in the corpus under synonym names

Fig. 3. Match rates across the corpus

Fig. 4 Geographical distribution of plant species matched from across the corpus

For Peer Review

For Peer Review

References

Allen, David E. (2003). Four centuries of local flora-writing: some milestones. *WATSONIA-KINGS LYNN-BOTANICAL SOCIETY OF THE BRITISH ISLES-*, *24*(3), 271–280.

Allen, David Elliston. (1994). *The naturalist in Britain: a social history*. Princeton University Press.

Amelia, T. J. (2017). *Corpus Linguistics for History: The Methodology of investigating place-name discourses in digitised nineteenth-century newspapers.* Lancaster University. Retrieved from https://doi.org/10.17635/lancaster/thesis/143

Arber, A. (1943). A seventeenth-century naturalist: John Ray. *Isis*, *34*(4), 319–324.

Atlas, N. B. N. (2019). No Title. Retrieved from https://registry.nbnatlas.org/

Baker, J. G. (1885). *A flora of the English Lake District*. George Bell.

Bennamoun, M., & Mamic, G. J. (2012). *Object recognition: fundamentals and case studies.* Springer Science & Business Media.

Blanke, T., Bryant, M., & Hedges, M. (2012). Open source optical character recognition for historical research. *Journal of Documentation*, *68*(5), 659–683. https://doi.org/10.1108/00220411211256021

Butler, J. O., Donaldson, C. E., Taylor, J. E., & Gregory, I. N. (2017). Alts, Abbreviations, and AKAs: historical onomastic variation and automated named entity recognition. *Journal of Map & Geography Libraries*, *13*(1), 58–81.

Cervera, T., Pino, J., Marull, J., Padró, R., & Tello, E. (2019). Understanding the long-term dynamics of forest transition: From deforestation to afforestation in a Mediterranean landscape (Catalonia, 1868--2005). *Land Use Policy*, *80*, 318–331.

Charles, E. (1947). Raven, English Naturalists from Neckam to Ray. Cambridge: Cambridge University Press.

Denyer, S. (2016). The Lake District Landscape: Cultural or Natural? In *The Making of a Cultural Landscape* (pp. 19–46). Routledge.

Desmond, R. (1977). *Dictionary of British and Irish Botantists and Horticulturalists Including Plant Collectors, Flower Painters and Garden Designers*. CRC Press.

Donaldson, C., Gregory, I. N., & Taylor, J. E. (2017). Locating the beautiful, picturesque, sublime and majestic: spatially analysing the application of aesthetic terminology in descriptions of the English Lake District. *Journal of Historical Geography*, *56*, 43–60. https://doi.org/10.1016/j.jhg.2017.01.006

Ellenberg, H. (1974) Zeigerwerte der Gefäßpflanzen Mitteleuropas. *Scripta Geobotanica*, 9. Göttingen.

Gledhill, D. (2002). *The names of plants*. Cambridge University Press.

Gregory, I., Atkinson, P., Hardie, A., Joulain-Jay, A., Kershaw, D., Porter, C., … Rupp, C. (2016). From Digital Resources to Historical Scholarship with the British Library 19th Century Newspaper Collection. *Journal of Siberian Federal University. Humanities & Social Sciences*, *9*(04), 994–1006. https://doi.org/10.17516/1997-1370-2016-9-4-994-

1006

Gregory, I., & Donaldson, C. (2016). Geographical text analysis: Digital cartographies of Lake District literature. In *Literary mapping in the digital age* (pp. 85–105). Routledge.

Grover, C., Tobin, R., Byrne, K., Woollard, M., Reid, J., Dunn, S., & Ball, J. (2010). Use of the {Edinburgh Geoparser} for Georeferencing Digitised Historical Collections. *Philosophical Transactions of the Royal Society A*, *368*(1925), 3875–3889.

Hinrichs, U., Alex, B., Clifford, J., Watson, A., Quigley, A., Klein, E., & Coates, C. M. (2015). Trading consequences: A case study of combining text mining and visualization to facilitate document exploration. *Digital Scholarship in the Humanities*, *30*(suppl_1), i50--i75.

Hodgson, W., & Goodchild, J. G. (1898). *Flora of Cumberland*. W. Meals and Company.

Kheraj, S., & Oosthoek, K. J. (2016). 18 Online digital communication, networking, and environmental history. *Methodological Challenges in Nature-Culture and Environmental History Research*, 233.

Koerner, L. (1996). Carl Linnaeus in his time and place. *Cultures of Natural History*, 145–162.

Kuklick, H., & Kohler, R. E. (1996). Science in the field, Osiris, 2nd Series, Vol. 11.

Kulldorff, M. (1997). A spatial scan statistic. *Communications in Statistics - Theory and Methods*, *26*(6), 1481–1496. https://doi.org/10.1080/03610929708831995

Lake District National Park. (n.d.). Lake District Nomination and Partnership Plan. Retrieved from http://www.lakedistrict.gov.uk/caringfor/projects/whs/lake-district-nomination

Lindop, G. (2005). *A literary guide to the Lake District*. Sigma Press.

Linton, E. L. (1864). *The lake country*. Smith, Elder and Company.

Malleson, F. A. (1890). *Holiday Studies of Wordsworth by Rivers, Woods, and Alps: The*

For Peer Review

Wharfe, the Duddon, and the Stelvio Pass. Cassell and Company.

Nicholas Jardine, James A. Secord, E. C. S. (Ed.). (1996). *Cultures of natural history*. Cambridge University Press.

Nicholson, N. (1955). *The Lakers: the adventures of the first tourists*. Hale.

Ogilvie, B. W. (2008). *The science of describing: Natural history in Renaissance Europe*. University of Chicago Press.

OS Open Data. (n.d.). Retrieved from https://www.ordnancesurvey.co.uk/opendatadownload/

Pickstone, J. V. (2001). *Ways of knowing: A new history of science, technology, and medicine*. University of Manchester Press.

POWO. (2018). Plants of the World Online. Retrieved from http://www.plantsoftheworldonline.org/

Preston, C. D., Pearman, D., Dines, T. D., & others. (2002). *New atlas of the British & Irish flora*. Oxford University Press.

Reddy, S. M. (2007). *University Botany-iii:(Plant Taxonomy, Plant Embryology, Plant Physiology)* (Vol. 3). New Age International.

Rich, T. C. G. (1997). Is ad hoc good enough. *Transactions of the Suffolk Naturalists' Society*, *33*, 14–21.

Rupp, C. J., Rayson, P., Gregory, I., Hardie, A., Joulain, A., & Hartmann, D. (2014). Dealing with heterogeneous big data when geoparsing historical corpora. In *2014 IEEE International Conference on Big Data (Big Data)* (pp. 80–83). IEEE. https://doi.org/10.1109/BigData.2014.7004457

Rupp, C. J., Rayson, P., Gregory, I., Hardie, A., Joulain, A., & Hartmann, D. (2015). Dealing with heterogeneous big data when geoparsing historical corpora. *Proceedings - 2014*

For Peer Review

IEEE International Conference on Big Data, IEEE Big Data 2014, (September 2013), 80–83. https://doi.org/10.1109/BigData.2014.7004457

Sanderson, J. (2017). *Plants, people and practices: the nature and history of the UPOV convention*. Cambridge University Press.

Scharf, S. T. (2009). Identification keys, the “Natural Method,” and the development of plant identification manuals. *Journal of the History of Biology*, *42*(1), 73–117.

Stace, C. (2010). *New flora of the British Isles*. Cambridge University Press.

Tanner, S., Muñoz, T., & Ros, P. H. (2009). Measuring mass text digitization quality and usefulness. *D-Lib Magazine*, *15*(7/8), 1082–9873.

Tobin, R., Grover, C., Byrne, K., Reid, J., & Walsh, J. (2010). Evaluation of georeferencing. In *Proceedings of the 6th Workshop on Geographic Information Retrieval - GIR '10* (p. 1). New York, New York, USA: ACM Press. https://doi.org/10.1145/1722080.1722089

UNESCO. (2017). The Operational Guidelines for the Implementation of the World Heritage Convention. Retrieved August 20, 2009, from https://whc.unesco.org/en/guidelines

Whittaker, E. J. (Ed.). (1981). *A Seventeenth Century Flora of Cumbria: William Nicolson's Catalogue of Plants 1690*. Publications of the Surtees Society,.

For Peer Review

Acknowledgements

We are grateful to Mike Bennett and Dr. Beatrice Alex at Edinburgh for their support with the geoparser and the Unlock service. We are also grateful to John Iacona at The Royal Botanical Gardens Kew, for help and guidance in accessing POWO data. Further thanks are due to Prof. Ian Gregory and Dr. Patricia Murrieta-Flores for their insight and guidance.

For Peer Review

Funding

This work was supported by the EU-China Research and Innovations Partnership (ECRIP) [2014/348-010], as part of the project Addressing food Security, Environmental stress and Water by promoting multidisciplinary Research EU And China Partnerships in science and business (Sew reap); the Leverhulme Trust [RPG-2015-230], as part of the Geospatial Innovation in the

Digital Humanities project; and the Arts and Humanities Research Council (AHRC) [AH/S012893/1], via the North West Consortium Doctoral Training Partnership (NWCDTP), as part of the project Augmented Humanity: Does the Human Enhance the Machine or the Machine Enhance the Human?

For Peer Review

Appendix

List of texts in corpus

Author	Title	Year of Publications	Volume Number
Ray, John	*Historia Plantarum*	1686	1
Ray, John	*Historia Plantarum*	1688	2
Ray, John	*Synopsis methodica stirpium Britannicarum*	1690	
Ray, John	*Historia Plantarum*	1704	3
Wilson, John	*A synopsis of British Plants*	1744	
Hudson, William	*Flora Anglica*	1762	
Nicholson and Burns	*History and Antiquities of the Counties of Westmorland and Cumberland*	1777	1
Nicholson and Burns	*History and Antiquities of the Counties of Westmorland and Cumberland*	1777	2
	Memoirs and proceedings of the Manchester Literary and Philosophical Society	1789	1
Pulteney, Richard	*Historical and Biographical Sketches of the Progress of Botany in England*	1790	1
Pulteney, Richard	*Historical and Biographical Sketches of the Progress of Botany in England*	1790	2
	Transactions of Linnaean Society of London	1791	1
Hutchinson, W.	*History of the County of Cumberland*	1794	1
Hutchinson, W.	*History of the County of Cumberland*	1794	2
	Transactions of Linnaean Society of London	1797	3

For Peer Review

Withering, William	*A Botanical Arrangement of British Plants*	1801	1
Withering, William	*A Botanical Arrangement of British Plants*	1801	2
Withering, William	*A Botanical Arrangement of British Plants*	1801	3
Withering, William	*A Botanical Arrangement of British Plants*	1801	4
Turner, D. and Dillwyn, L.	*Botanist's Guide through England and Wales*	1805	1
Turner, D. and Dillwyn, L.	*Botanist's Guide through England and Wales*	1805	2
	Memoirs and proceedings of the Manchester Literary and Philosophical Society, Second series	1805	1
Dugdale, J. T.	*A new British Traveller*	1819	1
Dugdale, J. T.	*A new British Traveller*	1819	2
Dugdale, J. T.	*A new British Traveller*	1819	3
Dugdale, J. T.	*A new British Traveller*	1819	4
Smith, James, Edward,	*English Flora*	1824	1
Smith, James, Edward,	*English Flora*	1824	2
Smith, James, Edward,	*English Flora*	1825	3
Winch, Nathaniel John	*Essays on the Geographical Distribution on Plants… of Northumberland, Cumberland and Durham*	1825	
Smith, James, Edward,	*English Flora*	1828	4
	Magazine of natural history	1829	1
Otley, J.	*A concise description of the English lakes and adjacent*	1830	

For Peer Review

	mountain		
Stokes, J	*Botanical Commentaries*	1830	
Hooker, William Jackson, Sir,	*The British flora; comprising the Phaenogamous, or flowering plants, and the ferns*	1831	
Smith P. eds	*Memoir and correspondence of the late Sir James Edward Smith*	1832	1
Smith P. eds	*Memoir and correspondence of the late Sir James Edward Smith*	1832	2
	Magazine of Natural History	1834	7
	Companion of the Botanical Magazine	1835	1
Murray, A.	*Northern Flora: or a Descriptions of Wild Plants belonging to the North and East of Scotland*	1836	1
Watson, Hewett Cottrell	*New Botanist's Guide to the Localities of the Rarer Plants of Britain*	1837	2
	Proceedings of the Botanical Society of London	1839	1
	Magazine of natural history, New Series	1840	4
Otley, J.	*A descriptive guide to the English lakes and adjacent mountains*	1842	
Babington, Charles Cardale	*Manual of British Botany: Containing the Flowering Plants and Ferns*	1843	
Hudson, J.	*Guide to Lakes*	1843	
Newman, Edward	*History of British Ferns and Allied Plants*	1844	
	The Phytologist	1844	1
	Gardeners Chronicle	1845	
	The Phytologist	1845	2
Atkinson, T.	*Hand-Book to the English Lakes*	1847	

For Peer Review

Lancaster, E. ed.	*The Correspondence of John Ray*	1848	
Baker, John Gilbert	*A supplement to Baines' Flora of Yorkshire,*	1854	
	The Phytologist	1854	5
Martineau, Harriet	*Complete Guide to the English Lakes*	1855	
Walcott, M. E. C.	*Guide to the Mountains, Lakes and Northwest Coast of England*	1860	
Nicholson	Annals of Kendal	1861	
	Phytologist, New Series	1861	5
Baker, John Gilbert	*North Yorkshire: studies of its botany, geology, climate, and physical geography*	1863	
Linton, w. J.	*The Lake Country*	1864	
	The Naturalist	1864	1
Bentham, George	*Handbook of the British Flora*	1865	1
Bentham, George	*Handbook of the British Flora*	1865	2
Linton, W. J.	*The Ferns of the English Lake Country*	1865	
Lowe. E. J.	*Our Native Ferns*	1865	1
Lowe. E. J.	*Our Native Ferns*	1867	2
Watson, Hewett Cottrell	Compendium of the Cybele Britannica	1870	
	Journal of Botany	1871	9
Watson, Hewett Cottrell	*Topographical Botany*	1873	1
	Transactions of Cumberland Association for Advancement of Literature and Science	1876	1
	Transactions of Cumberland Association for Advancement of Literature and Science	1877	2

For Peer Review

	Transactions of Cumberland Association for Advancement of Literature and Science	1878	3
	The Naturalist	1878	4
	Transactions of Cumberland Association for Advancement of Literature and Science	1879	4
	Transactions of Cumberland Association for Advancement of Literature and Science	1880	5
	Transactions of Cumberland Association for Advancement of Literature and Science	1881	6
	Natural History Society of Glasgow	1881	5
	Transactions of Cumberland Association for Advancement of Literature and Science	1882	7
	Transactions of Cumberland Association for Advancement of Literature and Science	1883	8
	Transactions of Cumberland Association for Advancement of Literature and Science	1884	9
	The Naturalist	1883	8
	Journal of Botany	1883	21
Watson, Hewett Cottrell	*Topographical Botany*	1883	2
Baker, John Gilbert	*Flora of the English Lake District*	1885	
	Journal of Botany	1885	23
	The Naturalist	1886	
	The Naturalist	1888	
Malleson	*Holiday Studies of Wordsworth*	1890	
	The Naturalist	1890	

For Peer Review

	The Naturalist	1891
	The Naturalist	1893
Crump, W. B. and Crossland, C.	*Flora of the Parish of Halifax*	1904

For Peer Review

27

Colchicum autumnale. Near Darlington, Egleston, and Butterby, Durham.

Convallaria majalis. In Scotswood, Denton, and Castle Eden Denes, in Gibside Woods, and near Winch Bridge, Teesdale; also near Warden Mill.

Ornithogalum luteum. By the Tees at Wycliffe, Barnard-Castle, and Egleston, and the Wear at Butterby.

Juncus subverticillatus—Bicheno. By the Lakes of Cumberland and Westmorland.

Epilobium alsinifolium. On Cheviot and Cronkley Fell, and on the highest ridge of Foalfoot, at the head of Longsledale. It is the Epilobium alpinum of Ray and Curtis.

Pyrola rotundifolia. In Castle Eden and Hawthorn Denes.

Pyrolia media. In Scotswood, and East Common Wood, and by Roadley Lake, Northumberland. In Hounswood, and Blackstone Bank Wood, Durham, from 100 to 1,000 feet.

Pyrola minor. In Gibside Woods, and on Teesdale Forest, at Cocken, in Arngill, Cow Close, and Hindon Gills, and in Skullwood, near South Hamsterley, Durham. In East Common Wood, at Catcherside and Wallington, Northumberland.

Fig. 1. Excerpt from Winch's Flora of the Lake District (Winch, 1825, p. 27)

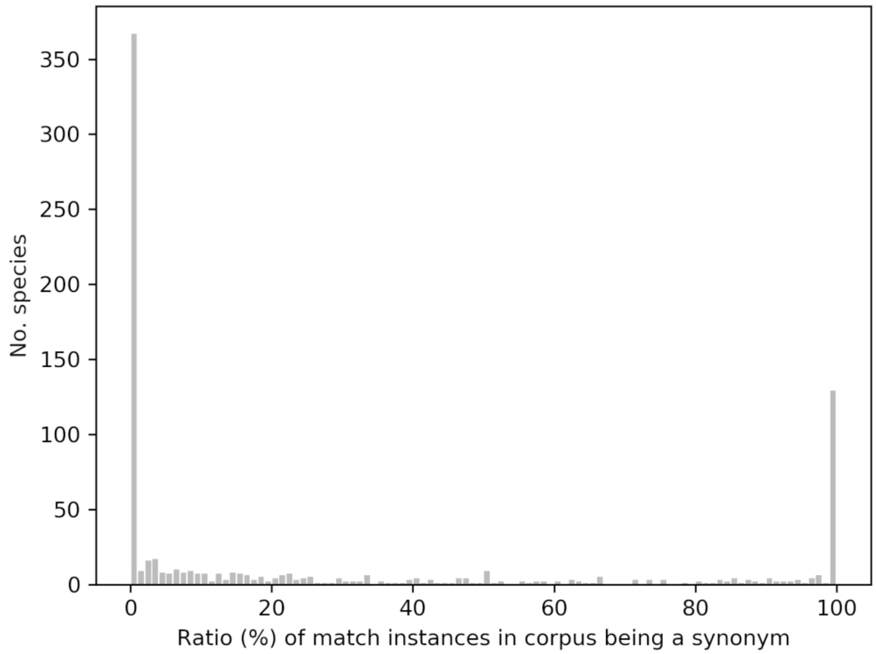

Fig. 2. Ratio (%) of plant species being matched in the corpus under synonym names

162x121mm (300 x 300 DPI)

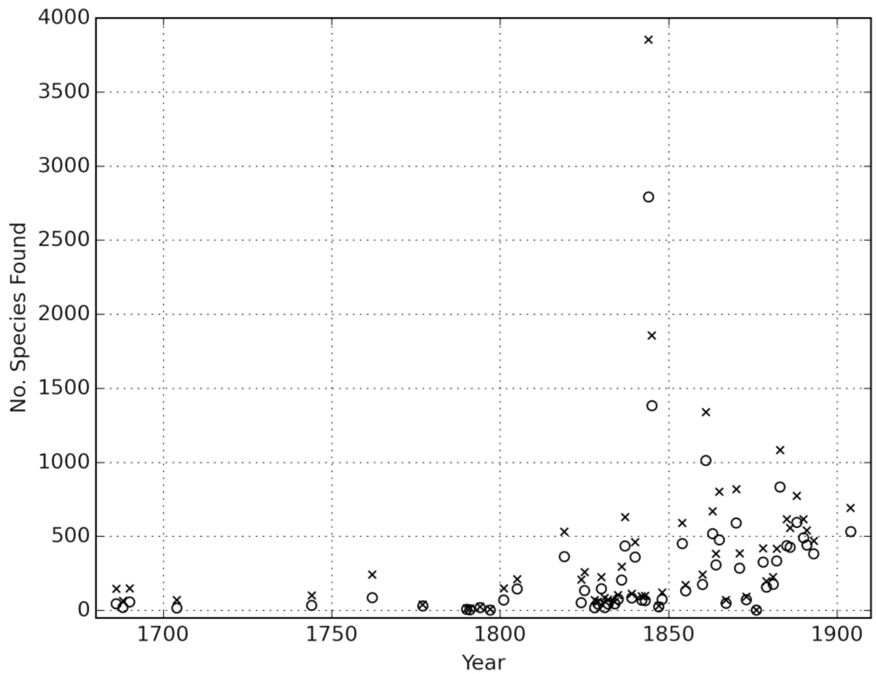

Fig. 3. Match rates across the corpus

203x152mm (300 x 300 DPI)

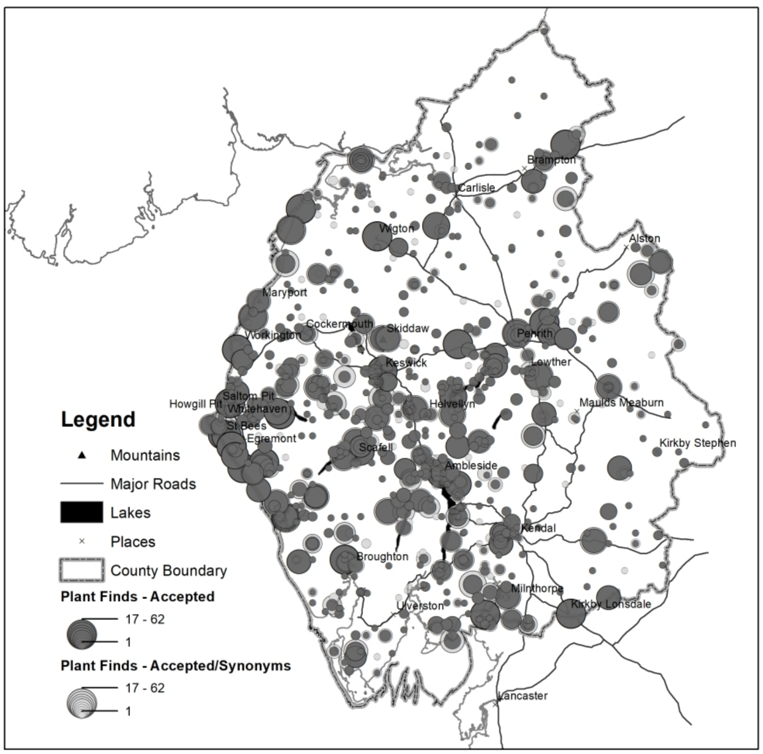

Fig. 4 Geographical distribution of plant species matched from across the corpus

210x297mm (300 x 300 DPI)

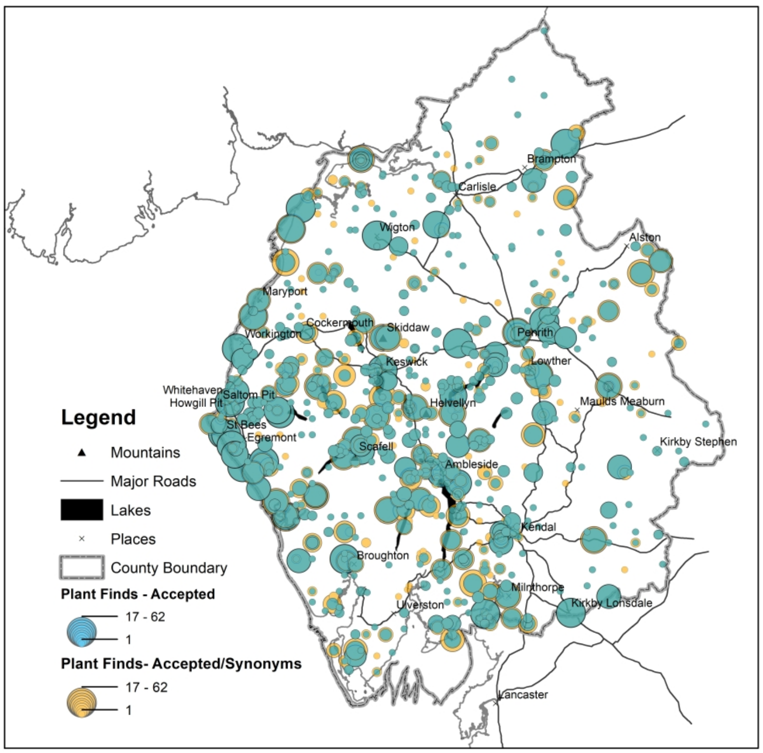

Fig. 4 Geographical distribution of plant species matched from across the corpus

210x297mm (600 x 600 DPI)