Automating the construction of higher order data representations from heterogeneous biodiversity datasets.

A thesis submitted for the degree of Doctor of Philosophy

by:
Nicky NICOLSON

Supervisor:
Dr. Allan TUCKER

Department of Computer Science, Brunel University, London

November 2019

BRUNEL UNIVERSITY, LONDON

Abstract

Department of Computer Science

Doctor of Philosophy

Automating the construction of higher order data representations from heterogeneous biodiversity datasets.

by Nicky NICOLSON

Datasets created from large-scale specimen digitisation drive biodiversity research, but these are often heterogeneous: incomplete and fragmented. As aggregated data volumes increase, there have been calls to develop a "biodiversity knowledge graph" to better interconnect the data and support meta-analysis, particularly relating to the process of species description. This work maps data concepts and inter-relationships, and aims to develop automated approaches to detect the entities required to support these kinds of meta-analyses.

An example is given using trends analysis on name publication events and their authors, which shows that despite implementation and widespread adoption of major changes to the process by which authors can publish new scientific names for plants, the data show no difference in the rates of publication. A novel data-mining process based on unsupervised learning is described, which detects specimen collectors and events preparatory to species description, allowing a larger set of data to be used in trends analysis. Record linkage techniques are applied to these two datasets to integrate data on authors and collectors to create a generalised agent entity, assessing specialisation and classifying working practices into separate categories. Recognising the role of agents (collectors, authors) in the processes (collection, publication) contributing to the recognition of new species, it is shown that features derived from data-mined aggregations can be used to build a classification model to predict which agent-initiated units of work are particularly valuable for species discovery. Finally, shared collector entities are used to integrate distributed specimen products of a single collection event across institutional boundaries, maximising impact of expert annotations. An inferred network of relationships between institutions based on specimen sharing relationships allows community analysis and the definition of optimal co-working relationships for efficient specimen digitisation and curation.

Acknowledgements

I would firstly like to thank my primary supervisor Allan Tucker, he has provided immense enthusiasm, support and guidance during this research work, and I am very grateful for the opportunity to work with him. Thanks to Abigail Barker for her support in embarking on this process and all her friendship and guidance, to Kathy Willis for securing Kew's support, and to Alan Paton for all his advice over the past years, and for his on-going supervision.

Thanks to all at the Centre for Intelligent Data Analysis at Brunel who have made it such a friendly and welcoming place.

The research presented in this thesis depends on data from the International Plant Names Index, Index Herbariorum and specimen collection data - collected from the field, curated and digitised by natural history museums and herbaria around the world and aggregated using data standards and tools designed and built by the Biodiversity Information Standards (TDWG) community and the Global Biodiversity Information Facility (GBIF). Thanks to all who have been part of these efforts.

Thanks also to Bill Baker (Royal Botanic Gardens, Kew) and Sandy Knapp (Natural History Museum, London) for providing application references, to Theary Ung (Centre National de la Recherche Scientifique, Muséum national d'Histoire naturelle, Paris) for advice regarding a mid-career PhD, to Quentin Groom (Meise Botanic Garden, Belgium), Elspeth Haston (Royal Botanic Garden, Edinburgh), Deb Paul (Florida State University, US) and James Macklin (Agriculture and Agri-Food Canada) for providing letters of support, and to Sarah Phillips (Kew) for advice about collections digitisation.

I am grateful to the Global Biodiversity Information Facility for their recognition of the work presented in this thesis through the Young Researcher Award (2019).

Thanks to my family and friends for their support and patience.

Outputs

Publications

The following publications have resulted from the research presented in this thesis:

1. N. Nicolson, K. Challis, A. Tucker, and S. Knapp, "Impact of e-publication changes in the International Code of Nomenclature for algae, fungi and plants (Melbourne Code, 2012) - did we need to 'run for our lives'?," BMC Evolutionary Biology, vol. 17, p. 116, May 2017.

2. N. Nicolson and A. Tucker, "Identifying Novel Features from Specimen Data for the Prediction of Valuable Collection Trips" presented at the International Symposium on Intelligent Data Analysis, 2017, pp. 235–246.

3. N. Nicolson, A. J. Paton, S. Phillips, and A. Tucker, "Specimens as research objects: reconciliation across distributed repositories to enable metadata propagation" in 2018 IEEE 14th International Conference on e-Science (e-Science), 2018, pp. 125-135.

Publication 1 results from preliminary work conducted for this thesis, presented in chapter 3. Publication 2 is an early output from work further developed in chapter 4 and 5, and publication 3 results from work presented in chapter 6. Complete versions of these papers are included in appendix B.

Grants

The following competitively funded projects further develop research areas first presented in this thesis:

1. SYNTHESYS+: 4 year European Union Horizon 2020 project re provision of physical and especially virtual access to natural science collections.

2. MOBILISE: 4 year European Cooperation in Science and Technology (COST) action: Mobilizing data, policy and experts in scientific collections.

Grant 1 references and proposes to further develop data-mining process outlined in chapter 4 to develop a "specimen data refinery" for use in the mass digitisation of biological specimens. Grant 2 will use entities and interrelationships resultant from data-mining in chapter 4 and specimen reconciliation in chapter 6 to inform an assessment of data standards, and

annotation flow and inter-institutional relationships resulting from chapter 6 will be further developed in a working group on "New concepts and standards for data management".

Further details are given in appendix D.

Other outputs: conference presentations

The following conference presentations and posters include work presented in this thesis:

1. N. Nicolson and A. Tucker "Clustering botanical collections data with a minimised set of features drawn from aggregated specimen data", Big Data Analysis Methods and Techniques as Applied to Biocollections, Biodiversity Information Standards 2016.

2. M. Collins, N. Nicolson, J. Poelen, A. Thompson, J. Hammock and A. Thessen "Building your own big data analysis infrastructure for biodiversity science", Using big data techniques to cross dataset boundaries - integration and analysis of multiple datasets, Biodiversity Information Standards 2017.

3. N. Nicolson and A. Tucker, "Interactive visualisation of field-collected botanical specimen metadata: supporting data mining process development", International Symposium on Intelligent Data Analysis 2018.

4. N. Nicolson, A. Paton, S.Phillips and A. Tucker, "Integrating collector and author roles in specimen and publication datasets", Biodiversity Next, 2019

5. Q.J. Groom, C. Besombes, J. Brown, S. Chagnoux, T. Georgiev, N. Kearney, A. Marcer, N. Nicolson, R. Page, S. Phillips, H. Rainer, G. Riccardi, D. Röpert, D.P. Shorthouse, P. Stoev and E.M. Haston, "Progress in Authority Management of People Names for Collections", Biodiversity Next, 2019

6. N. Nicolson, A. Paton, S.Phillips and A. Tucker, "Examining herbarium specimen citation: developing a literature based institutional impact measure", Biodiversity Next, 2019

The visualisation tool discussed in item 3 is presented in fuller form in appendix A. Abstracts for presentations are included in appendix C.

Contents

List of Figures

List of Tables

List of Algorithms

Chapter 1

Introduction

Biodiversity informatics is an emerging, cross-disciplinary science which is concerned with the digitisation, mobilisation and analysis of data derived from studies about species, either conducted in-field or through research using preserved specimen collections. This is a global scale activity, enabled by recent developments in communications, data management and imaging technology, alongside skills development in data handling and computational approaches. Integration and standardisation of the data to support a wide variety of downstream uses is recognised as a particular challenge.

Datasets derived from biological specimen collections drive biodiversity research - providing crucial "what / where / when" evidence - but these are often incomplete and fragmented, as specimen digitisation is an on-going process with many different participants. Initiatives such as the Global Biodiversity Information Facility (GBIF) harvest "occurrence" data (derived from field observations and specimen collections) from multiple data sources, represented in the DarwinCore data standard (Wieczorek et al. 2012). The aggregated product of this data harvesting is presented in a single data portal for use by researchers and policy makers. In addition to core scientific use cases, e.g. species distribution (what is found where) and habitat composition (which species interact) (Chapman 2005), these data could also support meta-analyses regarding the rates of species description and the participation of individuals in the species description process. Better understanding of the species description process is important because despite centuries of intensive effort, species discovery is not yet complete, with tens of thousands of new species of animals, plants and fungi described every year. Comprehensive meta-analyses are currently difficult to perform due to the level of standardisation of the data: analyses of the rates of publication of species description (Scotland and Wortley 2003), and estimates of progress towards a complete inventory of plant species have been reliant on expensive human-scale curation of source datasets, and focussed on only a subset of the activities which contribute to species discovery (Bebber et al. 2010) (Bebber et al. 2013).

Digitisation and information management are important to enable wider

access to the relevant data, as well as to enable new kinds of research via larger scale computational approaches. As the volume of data mobilised through wide-scale initiatives like GBIF increases, there have been calls to better interconnect the data to answer these new kinds of questions - including these meta-analyses - and to form a highly-connected "biodiversity knowledge graph" (R. D. Page 2016).

This research project investigates approaches to automate the construction of higher-order data representations, which can be used to interlink and reshape heterogeneous biodiversity datasets. It contrasts the scale and management approach between relatively small-scale, editorially managed datasets (primarily those covering scientific name publications), and larger scale, bulk-created datasets (derived from specimen and literature digitisation). It documents the kinds of data management that have been applied to smaller scale datasets using human intelligence, and seeks to develop automated approaches using techniques from the field of intelligent data analysis to apply similar processes to heterogeneous, distributed datasets at much larger scales. The aim is to automate the construction of higher order data representations, which can then be used to allow a wider range of data resources to participate in the modelling of the species discovery process. A concept map summarises the data entities and their interrelationships (figure 1.1), and is used to provide a visual representation of the context for each part of the research. The project is subdivided into four phases of work: (i) an initial demonstration of the trends analyses possible when the datasets are interlinked and key entities are formally recognised, (ii) development of a data-mining technique to enable wider scale trends analysis on specimen data, (iii) an agent integration process cross-mapping author and collector entities, and (iv) a specimen reconciliation process to explore data sharing between institutions. A custom visualisation toolkit aids data exploration and surfaces new research questions throughout the phases of the project.

The key contributions of this thesis are:

- A **novel data-mining technique to establish formalised entities representing the collector, collecting trip and sub-trip activity sequences** from an aggregated set of specimen metadata, drawn from a distributed set of sources
- An **automated process to abstract a generalised "agent" entity encompassing the multiple roles that a scientist performs throughout their career (collector and author)**, integrating data from editorially managed datasets, and data derived from larger-scale data-mining approaches

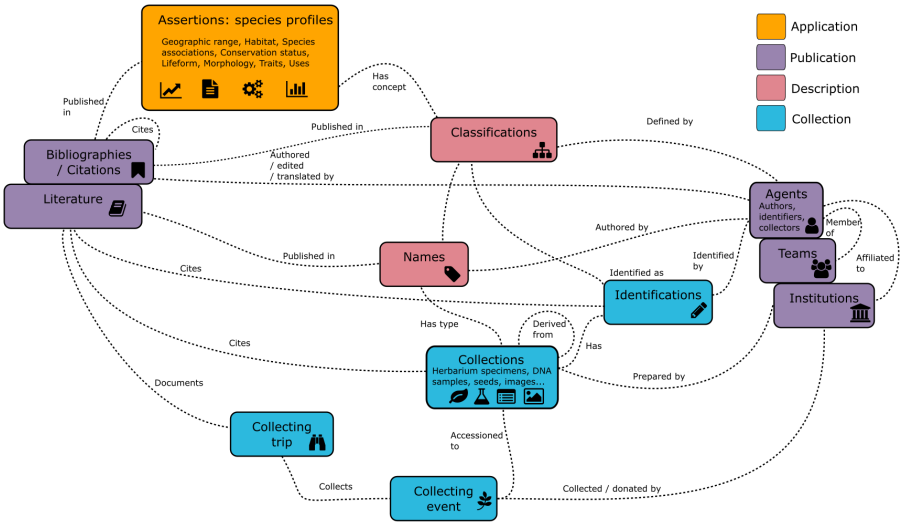


FIGURE 1.1: Concept map (read anti-clockwise from bottom left): researchers carry out field work in dedicated **collecting trips**, performing a sequence of **collecting events** which sample biological material from the field, along with field notes about the point of collection. The products of the collecting event (such as physical material, seeds, DNA samples) are lodged into multiple **specimen collections**. The process of accession into a specimen collection may include the distribution of duplicate samples to other **institutions** to form a distributed global collection. **Institutions**, **teams** and **individual researchers** are represented as **agents**. Agents label specimens with **identifications** – to determine the scientific **name** of the organism represented, and to place it into a taxonomic hierarchy using a **classification** system. The objective core of a species is the **name** in its formal nomenclatural sense, which has a special link through type citation to the "name-bearing type" specimen (a collection object). Scientific outputs such as classifications, names and phylogenies are published in **literature**, accessed via **bibliographic citations**. Once this groundwork is complete – i.e. species have been sampled, recognised, named and published - we can then document their characteristics to assemble **species profiles** - a diverse set of assertions including the morphological traits that a particular species displays, its chemical properties, human uses, conservation threats, legal (trade) statuses, interactions with other species - which are all evidenced with reference to published scientific literature or to physical voucher specimens (collection objects). These concept entities are coloured to indicate their fit with a set of high level activity stages: collection (blue), description and ordering (pink), publication (purple) and application (orange). This figure has been developed for this research project from a prototype version used to document shared entities between biodiversity information systems in a single research organisation.

- A **classification model to evaluate agent-initiated units-of-work (data-mined agent careers and collecting trips)** and assess their potential contribution to species discovery.
- A **reconciliation process to integrate the distributed specimen products of a single field collection event across institutional boundaries**.
- A **visualisation toolkit** designed to support the data-mining process and to surface new research questions

The structure of the remainder of the thesis is as follows:

A background chapter (chapter 2) further introduces the underlying scientific processes responsible for the generation of the source data, introduces the data used in the research, and the overall approach. This chapter develops a visual overview of the data entities and their potential interconnections (the concept map previewed in figure 1.1), sets the data in context using country level summaries, and lists the key research questions.

Chapter 3 has two purposes. Firstly, it presents a preliminary research analysis, demonstrating the kinds of trends analysis that are possible when the datasets are interlinked and key entities are formally recognised. This is an investigation of species name publication events in higher plants, focussing on data from a number of years pre and post governance rule changes designed to accelerate the process of species discovery. The aim of the research is to determine if these changes made an impact on the numbers of species described. A significant component of this part of the research project is to reshape the data to allow investigation of the authors participating in events relevant to species discovery. This theme of data reshaping to enable trends analysis on a non-primary data axis will be repeated throughout the research project. The second half of the chapter presents a range of machine learning techniques which can be applied to biodiversity data to facilitate these kinds of analyses.

Chapter 4 develops novel data-mining techniques using unsupervised learning, and applies these to a large aggregated dataset of specimen data to formalise data management of the collectors responsible for the field collection of the specimen material. Use of this collector entity reflects the data investigations conducted in the preceding chapter, which re-oriented name publication event data to promote the author of the publication event for more in-depth analysis. Here, specimen collection event data are re-oriented to promote the collector of the specimen for further analysis. Following recognition of the collector entity, further steps detect implicit aggregations - differentiating their work into the products of separate field collecting trips (using targeted clustering) and sub-trip units-of-work using state-transition analysis (Hidden Markov Models).

Chapter 5 recognises that the authors used in the analysis in chapter 3, and the collectors used in the analysis in chapter 4 frequently represent same person undertaking different activities at different stages of their careers. This part of the research abstracts authors and collectors to an underlying "agent" entity. Record linkage techniques are used to create links between these different data resources based on the underlying "agent"; features are then derived from collection and publication event histories and used to categorise the agents. Agent-initiated units-of-work are used to aggregate specimen data and assess contributions towards species discovery using classification models.

Chapter 6 uses the products of the agent data-mining processes first presented in chapter 4 to examine the level of data sharing between separate specimen-holding institutions, recognising the role of specimens as long-term research objects, which are subject to data annotation after accession into a reference collection. These data curation events can be viewed as research-grade units-of-work, generated by the agents recognised in chapter 5.

The final conclusions chapter (chapter 7) re-states the research objectives and context, and evaluates the activities undertaken. Potential criticisms of the work are anticipated and addressed, and potential revisions, generalisations to enable wider application, and suggestions for future related work are given.

Given the flow of results and techniques between the research activities described in this thesis, the inter-relationships between the separate research chapters are depicted graphically in figure 1.2.

A glossary provides definitions of terms from both the biodiversity informatics and computer science domains, and also provides background information for projects and initiatives referenced in the text.

Visualisation of concepts and data has been used as an enabling technique throughout the project, and software tools built to support this cross-cutting strand in the research project are presented in appendix A. Further appendices collate published outputs resulting from this research project: full journal articles and conference papers are presented in appendix B, and abstracts of conference presentations are in appendix C. Appendix D provides details of two competitively-awarded grant-funded projects which will further develop this work.

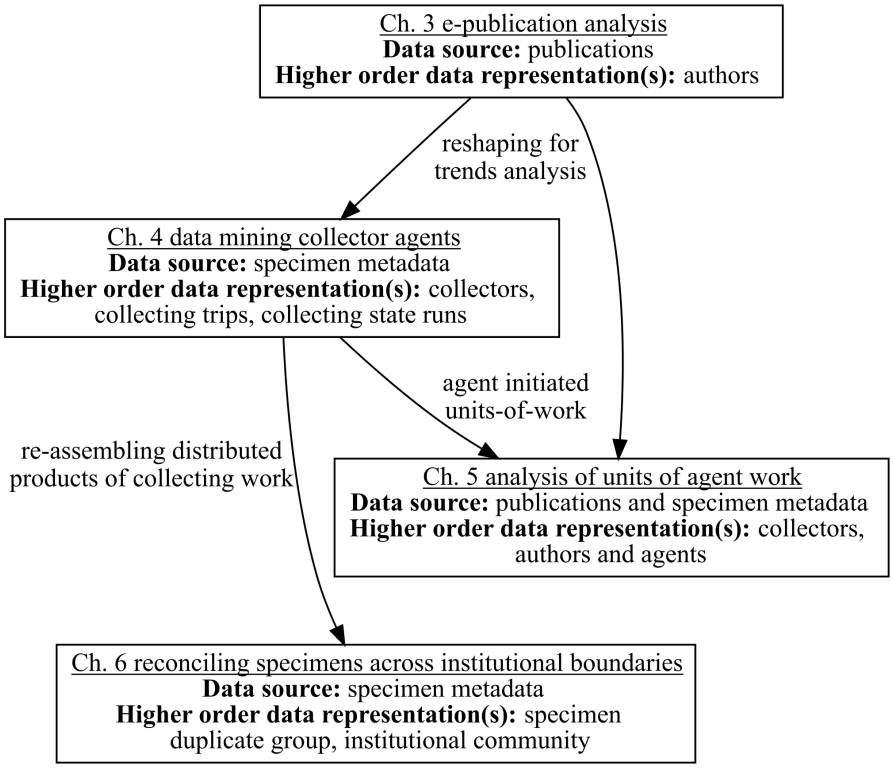


FIGURE 1.2: Relationships between the research chapters in this thesis

Chapter 2

Background

This chapter aims to provide (i) a background to the biodiversity informatics domain, its data generating processes and how its scientific and data resources are organised, (ii) a survey of existing work, and (iii) a proposal for new work, defining a problem statement and research questions which will be addressed in later chapters of this thesis.

2.1 Background to the biodiversity informatics domain

2.1.1 Definition of term, history and scope

Biodiversity informatics is an emerging science which aims to manage and learn from large volumes of data derived from studies about the natural world. Data are gathered via activities ranging from field work to laboratory based research using scientific specimens stored in natural history museums and herbaria worldwide - estimated at over 2-3 billion specimens globally (Chapman 2005) (Hardisty et al. 2013). Many uses have been documented for these primary biodiversity data (Chapman 2005) but the integration of biodiversity data is acknowledged as a particular challenge (Hardisty et al. 2013).

Biodiversity informatics is distinct from *bioinformatics* – the latter discipline is concerned with the application of informatics techniques to gene and genome level biology, whereas biodiversity informatics is mainly concerned with data gathered at the whole organism scale, using data from field observations and collections of physical specimens. The first cited use of the term dates from the late 1990s (Schalk 1998) (A. T. Peterson et al. 2010). Table 2.1 provides a scale comparison for a range of natural informatics

disciplines.

TABLE 2.1: Comparison of the scales of different informatics disciplines: bioinformatics (Hogeweg 2011), biodiversity informatics (Schalk 1998), evolutionary informatics (Parr et al. 2012) and ecoinformatics (W. K. Michener and M. B. Jones 2012)

Scale	Bioinformatics	Biodiversity informatics	Evolutionary informatics	Ecoinformatics
Natural processes				✔
Ecosystem				✔
Species		✔	✔	
Organism		✔	✔	
Genome	✔		✔	
Gene	✔		✔	

Biodiversity informatics has emerged and become established throughout the computational era, but the processes and data upon which it operates have been gathered over a much longer timescale. The primary underlying scientific activity is the science of systematics, described as composed of six factors (C. D. Michener et al. 1970):

> [Systematics] is the field that (a) provides names for organisms, (b) describes them, (c) preserves collections of them, (d) provides classifications for the organisms, keys for their identification, and data on their distributions, (e) investigates their evolutionary histories, and (f) considers their environmental adaptations

In the next section these user-oriented services provided by systematics will be re-ordered to present the underlying data generation processes in temporal order, from creation to application.

2.1.2 Systematics processes

The data entities generated by systematics processes, and their potential interconnections are illustrated in figure 1.1.

The initial data generation process in systematics is field **collection** of material from its natural habitat. Scientists often conduct intensive periods of field work (particularly if working in remote or difficult to access regions) by participation in **collecting trips**. They gather unstructured data about the environmental conditions, habitats and species observed in field notebooks and conduct a series of **collecting events**, gathering physical samples of living material for accession into reference **collections** for later study, and cross referencing these to field notes. Specimen reference collections are stored in **institutions** such as natural history museums and herbaria (plant collections). Institutions are linked through shared activities to form a global network: exchanging both specimens and associated data and expertise.

Specimen collections are managed by institutions for long term consultation by scientists (and researchers in other disciplines, including

geography and history), as a data resource for new research and an aid to interpret existing research. As specimens are used in research contexts, they are annotated with extra data. One key annotation is **identification** - the labelling of a specimen with a scientific **name**, to indicate that it represents an existing or new species, and to place it into a predictive hierarchy using a **classification** system. The process by which new species names are created is governed by a set of rules (the nomenclatural code), which specify that scientific names must interpretable (by referencing a physical type specimen) and must be introduced via a formal **publication**.

The outputs of the systematic process are formally published scientific names and classifications. These products can then be used to organise a set of allied data to create **species profiles** - to make assertions about the morphological traits that a particular species displays, its chemical properties, human uses, conservation threats, legal (trade) statuses, interactions with other species etc. These assertions can be evidenced with reference to published scientific literature or to the physical specimens themselves.

2.1.3 Distribution of data and expertise

As scientific collections have been assembled over historic time periods, globally they are unevenly distributed, often reflecting colonial history. The difficulty experienced by researchers in accessing accurate information about species, and the necessary materials (including specimen collections and scholarly publications) with which to conduct research is termed the "taxonomic impediment" (Convention on Biological Diversity 2007). International efforts to overcome this include the Global Taxonomy Initiative, part of the United Nations Convention of Biological Diversity, which aims to address the uneven access to skills and resources at a global scale (Secretariat of the Convention on Biological Diversity 2010).

The difference between the physical distribution of scientific specimens and the physical distribution of species globally can be represented visually, using current-day political countries as units (figure 2.1). Two metrics are generated for each country: the number of plant species estimated to be found in the country from a compilation of country level species counts (generated as an intermediate product in an analysis estimating the size of the worlds threatened flora (Pitman and Jørgensen 2002)) - and the number of plant specimens estimated to be held in the country, from the Index Herbariorum resource, which lists the location of specimen collections and estimates of their size (Thiers continuously updated). Each country point has been coloured to indicate its continent. This shows that the European countries hold very high numbers of specimens but contain relatively little species diversity and conversely that the "tropical hotspot" countries have

very high species diversity but hold proportionally few of the necessary materials (specimen collections) required for systematics research.

2.1.4 Informatics developments: digitisation and data aggregation

Since the recognition of the taxonomic impediment and the definition of the field of biodiversity informatics, several large scale regional and global data mobilisation efforts have been initiated, including the creation of the Global Biodiversity Information Facility (GBIF) - a response to the mega-science forum within the Organisation for Economic Cooperation and Development (*Final Report of the OECD Megascience Forum Working Group on Biological Informatics* 1999). GBIF has evolved to focus primarily on the mobilisation of occurrence data - that drawn from specimens and field observations - and the data standards needed to efficiently organise these. As the scope of GBIF is now more focussed than the wide-ranging vision laid out in the 1999 report, other global-scale initiatives have taken on informatics activities in some defined areas (Hobern et al. 2019). Effectively the high level activity stages depicted in figure 1.1 each have their own dedicated global data aggregator - GBIF for occurrence data (specimens and observations), Catalogue of Life for taxonomic data, the Biodiversity Heritage Library for literature, and Encyclopaedia of Life for species profiles. Where these data are drawn from physical material via digitisation, there is often a gap between the amount of data digitally available and the number of physical resources which must be digitised in order to mobilise the data. Totalling the estimated numbers of specimens given for each collection listed in Index Herbariorum gives 390.48M specimens in 4,073 collections Thiers (continuously updated), but GBIF contains 77.46M[1] digital records derived from botanical specimens, showing that there is still a considerable effort required to make legacy data digitally accessible and computable.

GBIF coordinated the production of the Global Biodiversity Informatics Outlook document which aims to identify activities and coordinate efforts and funding (Hobern et al. 2012). This initiative has recently been revised and updated with a call to form "an alliance for biodiversity knowledge" which will work on coordination and collaboration models (Hobern et al. 2019). This is intended to align digitisation and data aggregation projects to maximise utility and to gain efficiencies of scale.

2.1.5 Scientific vision

Complementing these technical advances, there have been discussions to define a shared scientific vision and an optimal approach to realising its aims.

[1] Numbers calculated from GBIF API call executed on 2019-11-05

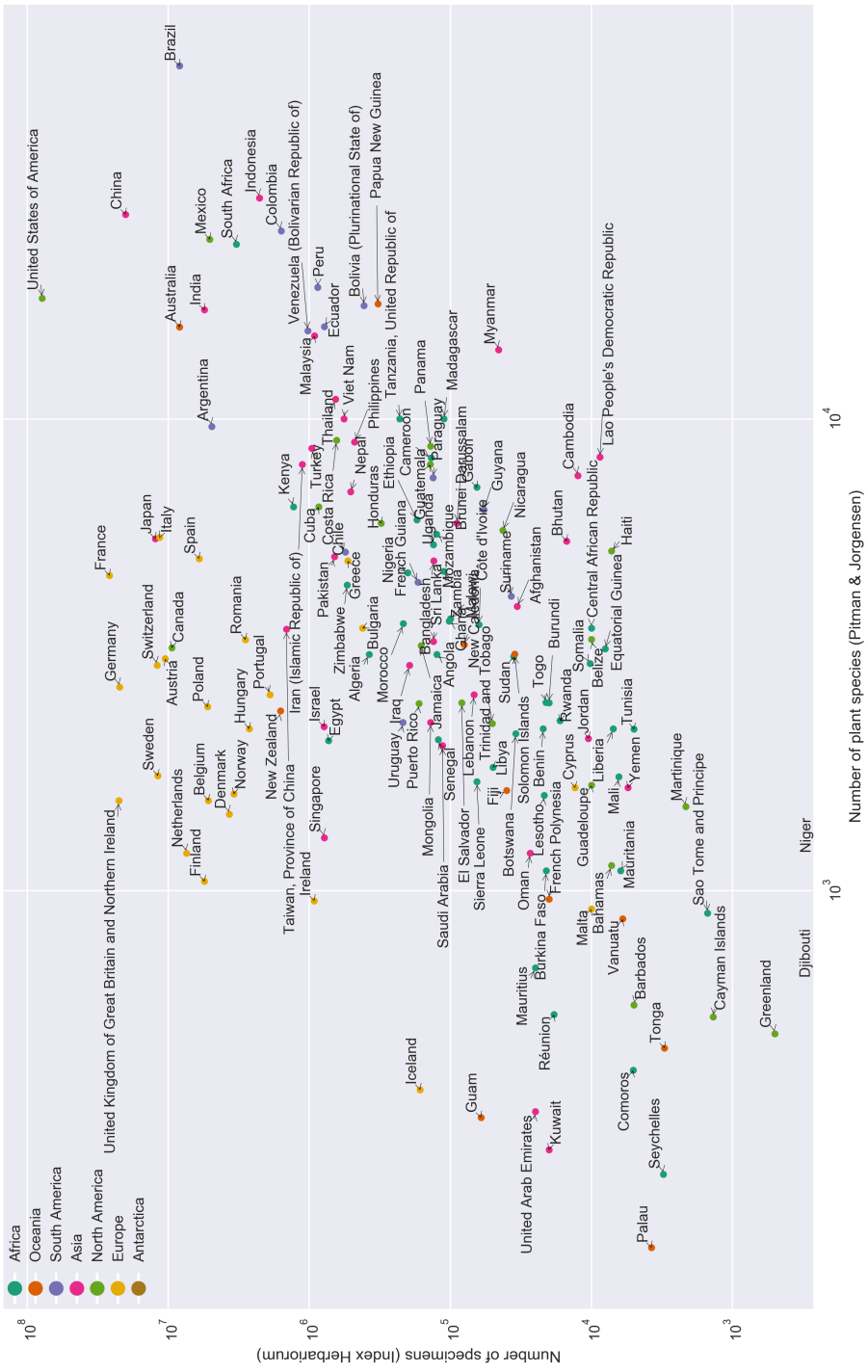


FIGURE 2.1: Per-country comparison of number of species found in-country (Pitman and Jørgensen 2002) against number of specimens held in-country (Thiers continuously updated)

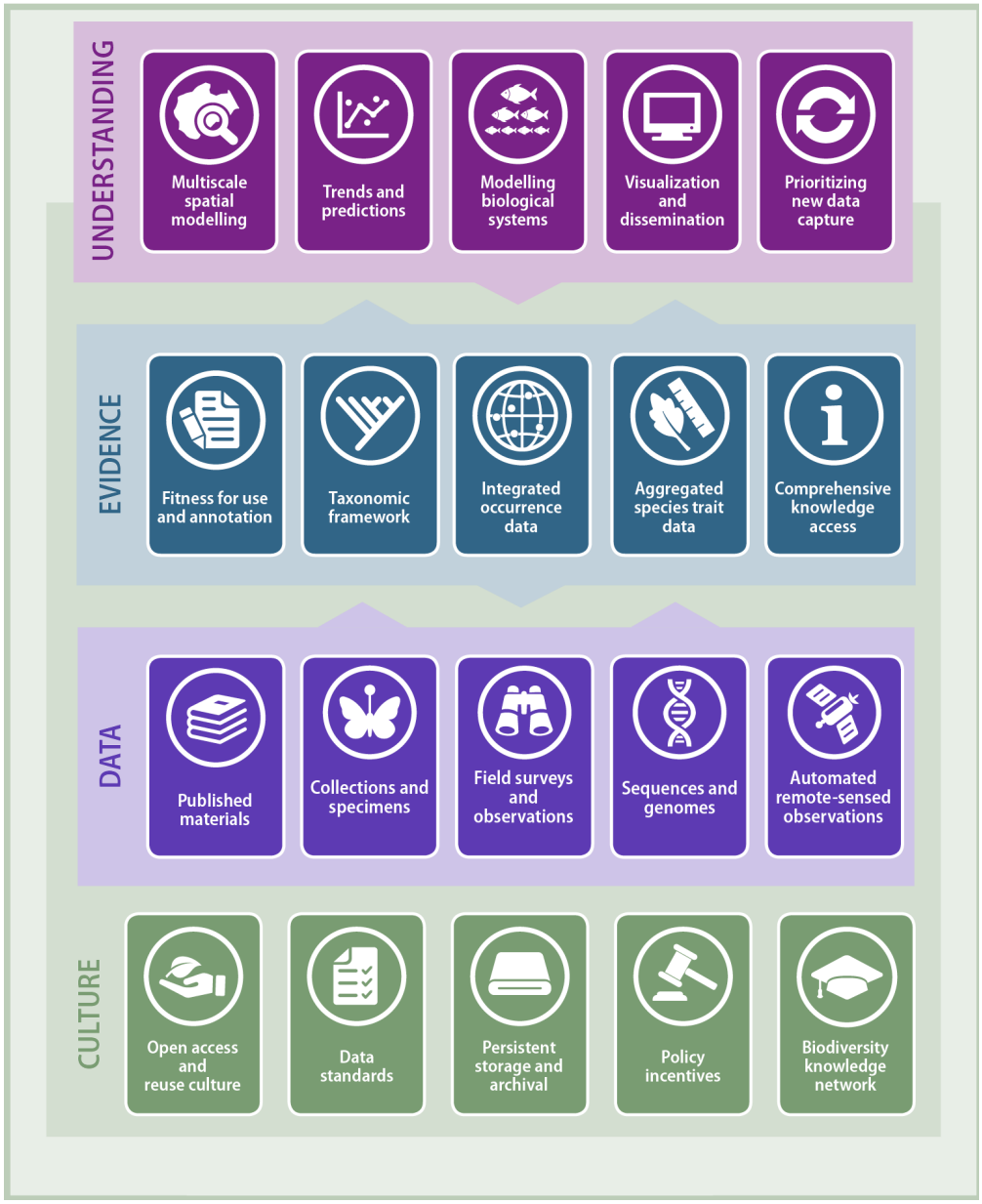


FIGURE 2.2: Global Biodiversity Informatics Outlook (GBIO) framework: "At the root lies the culture focus area which puts in place the necessary elements to turn biodiversity information into a common and connected resource – stable and persistent storage, pooled expertise, the culture and policies to support sharing, and common data standards. Building on those foundations, the data focus area aims to accelerate the mobilization of data from all sources, unlocking the knowledge held in our collections and literature, improving data quality and filling in gaps, and bringing observations and data from all sources from satellites to genomes online. The evidence focus area deals with refining, structuring and evaluating the data, to improve quality and place it within a taxonomic framework that organizes all known information about any species. Finally, the understanding focus area enables a broader synthesis, providing the modelling tools to enable us to look at whole ecosystems, make better policy decisions and react to any changes. The diagram shows how the focus areas interconnect, and breaks them down into their individual components." (Hobern et al. 2012)

A workshop convened in New York in 2010 documented the following ambition for biodiversity informatics (Wheeler et al. 2012):

> Our goal is no less than a full knowledge-base of the biological diversity on our planet, by which we mean: knowledge of all Earth's species, and how they resemble and differ from each other (i.e. all their characters from detailed morphology to as much genomic information as is feasible to collect); a predictive classification of all these species, based on their interrelationships as inferred from all these characters; knowledge of all the places at which each of these species has been found with as much ecological data as are available from specimens in the world's collections (e.g. host data, microhabitat data, phenology, etc.); and cyberinfrastructure to enable the identification of newly found specimens (including automated identification systems based on images and genomic information), the efficient description of species, and open access to data, information and knowledge of all species by anyone, amateur or professional, anywhere, any time.

The report recognises that the biodiversity data assembled to date can be viewed as a starting point, rather than an end in itself: from this starting point we may discover trends and make future predictions. The adoption of different techniques to analyse biodiversity data have been proposed (Kelling et al. 2009), mindful of the large volumes of observational data used as a start point, contrasting the traditional hypothesis-driven scientific method with a data intensive process where information emerges from the data. Data availability and technology have driven advances in the field to date (A. T. Peterson et al. 2010) – this work recognises that the advances have been opportunistic and calls for a conceptual framework to place biodiversity data in a broader context, to enable the modelling of the biosphere, from genes to ecosystems using a data intensive approach (Wheeler et al. 2012) (Purves et al. 2013). A "biodiversity knowledge graph" has been proposed (R. D. Page 2013) (figure 2.3) which is intended to develop a richly interconnected set of data in order to answer wider ranging questions.

2.2 Existing work: aggregating and inter-linking biodiversity data

The previous section presented three high level graphical overviews of the biodiversity landscape (figures 1.1, 2.2 and 2.3), which map high level data concepts and their interrelationships, showing that there is broad consensus

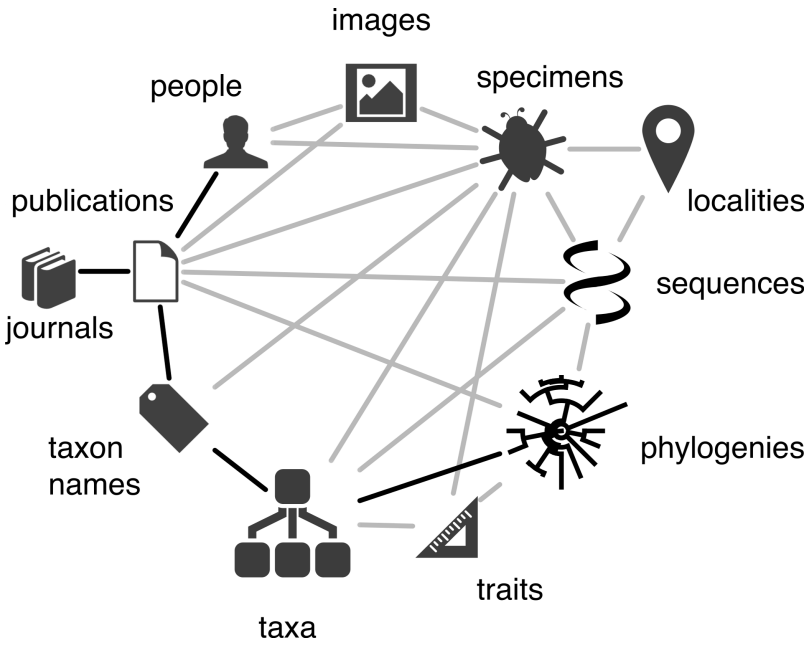


FIGURE 2.3: Biodiversity knowledge graph (R. D. Page 2013)

on the concepts recognised in the domain. A cross mapping of concepts from the separately published overviews is shown in table 2.2.

TABLE 2.2: Cross-mapping between depictions of concepts in the biodiversity informatics landscape

Concept map (figure 1.1)	GBIO (figure 2.2)	Biodiversity knowledge graph (figure 2.3)
Agents / teams / institutions	Culture: biodiversity knowledge network	People
Collecting trips / events	Data: field surveys & observations	Localities
Specimens	Data: collections & specimens	Specimens
Classifications	Data: sequences & genomes	Sequences / Phylogenies
Citations & literature	Data: published materials	Publications
Identifications	Evidence: fitness for use & annotation	-
Names	Evidence: taxonomic framework	Names
Classifications	Evidence: taxonomic framework	Taxa
Species profiles	Evidence: aggregated trait data	Traits
-	-	Images

2.2.1 Data representation, identification and aggregation

Given recognition of a common set of concepts, it is necessary to define shared data representations (vocabularies and data standards) in order to allow data sharing and tools development. Commonly used vocabularies include initiatives to encode the names of people who have authored scientific names (Brummitt and Powell 1992), literature relevant to the species description process (Stafleu and Cowan 1976) and institutions which hold specimen collections (Thiers continuously updated). Data standards define a data structure which may be used for interchange, and may utilise

vocabularies. An early data standard for the sharing of specimen data was the Herbarium Information System and Protocols for Interchange of Data (HISPID) protocol developed to share information between Australian institutions in order to prevent repeated data entry tasks. Since the definition of this specimen data standard there has been much work on the representation of specimen and scientific name data in data standards to permit data sharing and tools development.

A standards body for biodiversity data was created in 1985 as the Taxonomic Databases Working Group (TDWG), the organisation is now named Biodiversity Information Standards (TDWG). TDWG aims to facilitate standards development, and many of the vocabularies and data standards introduced above were adopted as "prior standards". This adoption recognises their utility to the scientific community, but also acknowledges that the "prior standards" were not created through a formal standards development process. Other data standards developed include DarwinCore (Darwin Core Task Group 2009) and Access to Biological Collections Data (Access to Biological Collections Data task group 2007) to represent specimens, Taxonomic Concept transfer Schema (Taxonomic Names and Concepts Interest Group 2006) to represent scientific names, and the Description Language for Taxonomy (Dallwitz 2006) and Structured Descriptive Data (Hagedorn et al. 2005) represent descriptive data and species profiles.

Alongside and complementary to data standards development, there has been work on persistent identifier schemes to allow the referencing of data elements via a resolvable identifier. An early trial implemented the Life Sciences Identifier scheme (LSID) (Clark et al. 2004) to scientific names (Richards 2010) (R. D. Page 2008), further work extended the persistent identification concept to other data elements recognised in the domain (Cryer et al. 2009) and recent work has focussed on specimens, using an HyperText Transfer Protocol (HTTP) Uniform Resource Identifier (URI) as a resolvable identifier rather than LSID (Hyam et al. 2012) (Guralnick et al. 2014) (Güntsch et al. 2017).

Larger scale data mobilisation was initially implemented as a federated search system, data providers installed middleware "wrapper" software to map their own data holdings to a shared data standard, and to respond to federated search requests. Two federated search systems were developed, one European-based using the Access to Biological Collections (ABCD) data standard and the BioCASE wrapper software and one using the Darwin Core data standard and the DIGIR wrapper software. Federated search became less reliable as more data providers participated in the networks, and the maintenance costs of local installations of middleware increased, and data aggregators have now moved to a data harvesting model, where wrapper

software is simplified to generate a package of data which is harvested by a process from the aggregator, and all queries run on the aggregation side, rather than being farmed out to all participants (Robertson et al. 2014). A current topic of discussion is to harmonise the data harvesting and post-harvest data processing to share effort between different aggregation networks. This is proposed as an exemplar project in the "alliance for biodiversity knowledge" (Hobern et al. 2019).

2.2.2 Data digitisation

Data concepts depicted in the high-level overviews of the biodiversity informatics landscape span those which are "born-digital" and those which are translated into computable form via the digitisation of physical specimens and literature. This section examines the different kinds of source data and the kinds of techniques which have been applied in order to populate and interlink these resources to form the connections shown in the representations of the proposed "biodiversity knowledge graph" (figures 2.3 and 1.1).

Specimens

Specimens are physical objects, digitisation processes applied to specimens can generate structured metadata (usually represented using a data standard as described above). A sample image of a herbarium specimen is given in figure 2.4 - showing the data held on the specimen that can be translated into structured metadata. Specimen digitisation may also generate images, image data can be considered as both structured (metadata about the creation of the image and regions within it) and unstructured (if different regions of the image are not identified).

Texts

Texts are another source of biodiversity data, which are digitised from reports and published materials. Descriptions of species, habitats and interactions have been published as texts in scientific literature over a timescale of hundreds of years. These are being digitised and made available as unstructured text through optical character recognition processes on page images in projects like the Biodiversity Heritage Library.

Named entity recognition is a set of techniques to recognise and extract information units (such as the names of people, places and things) from surrounding unstructured text. The field has evolved from the use of dictionaries to drive discovery of entities towards more probabilistic approaches (Nadeau and Sekine 2007). The use of named entity recognition in biodiversity informatics encompasses the use of dictionaries (lists of scientific names, gazetteers which list geographic names), rules based

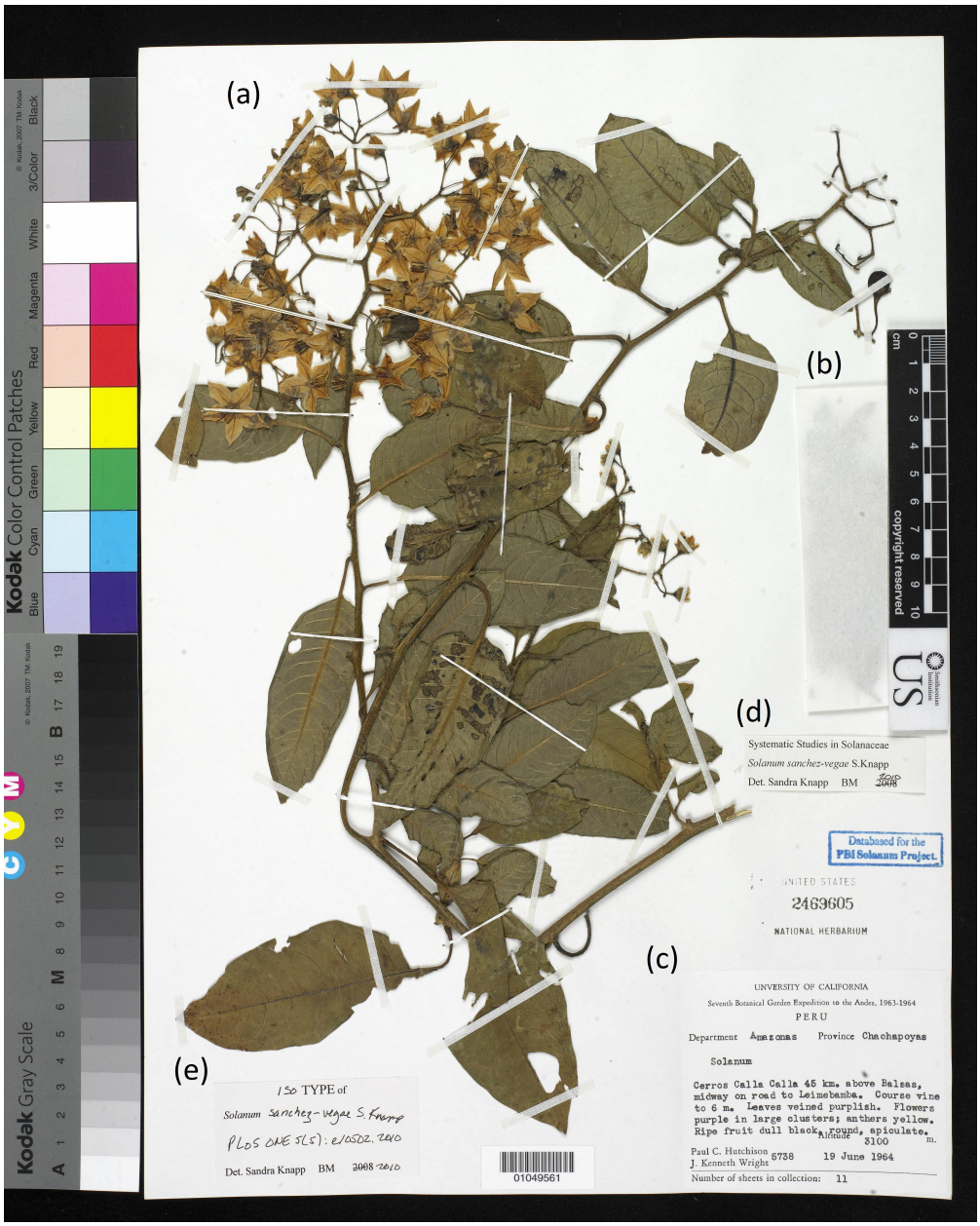


FIGURE 2.4: Example herbarium specimen (Orrell 2019). Along with the physical material that forms the specimen: the pressed, dried plant (a) and paper envelope of seeds (b), the specimen sheet holds collection metadata on the label in the lower right hand corner (c), and two research annotations: an identification statement (d) and a type citation - a note that the specimen was referenced in an academic work (e).

approaches which define word morphology and machine learning approaches which use probabilistic and statistical techniques. A dictionary based approach is used in TaxonFinder (Leary et al. 2007), a hybrid dictionary and rules based approach is used in TaxonGrab (Koning et al. 2005) and statistical and probabilistic techniques are used in NetiNeti (Akella et al. 2012).

Some publications include regular formatting to display species descriptions, the text derived from these can be subject to parsing to generate more structured representations suitable for computational use, representing a change of emphasis from unstructured text aimed at human users to structured data aimed at machine use at scale. Efforts straddling these two aims when creating new information have been seen with the rise of *semantic publishing*, where a journal article (aimed at human readers) also has underlying structured mark-up to permit machine use: a single data resource supporting two differing but complementary use cases (Penev et al. 2010).

Various options exist in terms of extracting data and making it computationally available, these range from "bag-of-words" approaches which are not concerned with the order of terms in a text (Tucker and Kirkup 2014) to full-scale linguistic analysis in which language and sentence structures are used to help the extraction of information. However, traditional natural language processing tools are difficult to apply to biological literature given the abbreviated sentence structure employed in descriptive passages (sometime known as *telegraphic sub-language*). Parsing strategies can be regarded as *supervised* or *semi-supervised* if aided by dictionaries and vocabularies to help the parsing process, as *active-learning-based* if supported by expert verification, or as *unsupervised* if the parsing process operates completely unseen in a bootstrap fashion. The latter approach is shown in work on the Charaparser software system (Cui 2012).

Dictionaries and ontologies

One strategy to achieve named entity recognition is to use a dictionary of terms. A data standard defines a structure for data, to aid interoperability between separate data repositories. Darwin Core (Wieczorek et al. 2012) is such a data standard that is widely used in biodiversity informatics, most noticeably to share data with the Global Biodiversity Information Facility (GBIF). A data standard may define controlled vocabularies (lists of permitted terms) for some of its fields. An ontology is a richer data representation tool than a dictionary, used to define terms, place them into a hierarchy and also define the inter-relationships between them (Walls et al. 2014). Multiple ontologies exist that are relevant to the biodiversity informatics landscape and the botanical domain within it, including the Plant Ontology (Jaiswal et al. 2005), the Biological Collections Ontology, the

Environment Ontology and the Population and Community Ontology (Walls et al. 2014).

Traditionally, data standards and ontologies tend to have been constructed by committees of experts in a *top-down* fashion. Due to increasing availability of data, an emergent, *bottom-up* approach can be added to augment existing ontologies and controlled vocabularies with new terms uncovered in text-mining. This augmentation is effectively an active learning approach whereby an expert user is presented with a set of suggested additions to an existing ontology – this is as described in the augmentation of the Hymenoptera Anatomy Ontology with terms drawn from selected literature extracted from the Biodiversity Heritage Library (Seltmann et al. 2013). This work also aimed to help understand the trends in term usage over time – reflecting the distinction between *data management* (building the ontology) and *data understanding* (conducting research with the data).

A further example of ontology augmentation is that using terms gathered from digitised collection object labels (using biological specimens as the source data rather than published literature). This has been demonstrated using habitat terms digitised with collection label data from the iDigBio specimen digitisation and aggregation project, to propose new entries in the ENVO ontology (Buttigieg 2015).

Record-linkage

Two major entity types in biodiversity informatics are the scientific names of organisms and the physical specimens (upon which the scientific names are based). When represented as *structured (meta)data* these are good candidates for *record-linkage*, as many different data sources exist which can be cross linked via these entities – GBIF mobilises specimen and occurrence data in 47,800 datasets from 1,961 different data providers,[2] the Catalogue of Life integrates taxonomic data from 130 different data providers (Roskov et al. n.d.).

Record-linkage exercises in biodiversity informatics have tended to be deterministic, focussing on rules based techniques. TAXAMATCH is a deterministic, rules based scientific names reconciliation utilising phonetic and edit distance calculations (Rees 2014). It is used in several projects including iPlant, Atlas of Living Australia and the Interim Register of Marine & Non-Marine Genera, but so far has tended to be used in user facing software rather than bulk behind-the-scenes data integration.

Botanical specimens in particular are good candidates for specimen to specimen record linkage as the standard working practice in botany is to collect duplicate specimens from an individual in the field and to distribute

[2]Numbers calculated from GBIF API call executed on 2019-11-05

these to separate institutions. An estimate for the level of duplication in US herbaria (with an estimated total of c 95 million specimens) is that half of the as yet undigitised portion are duplicated at least once. (Macklin et al. 2006). The FilteredPush annotation sharing framework is based around the principle of sharing data and annotations between linked specimen duplicates, but the literature describing the project focusses on the annotation sharing rather than the development of record links (Wang et al. 2009). It is referenced in work on improving the efficiency of specimen data capture by drawing information from remote sources (Tulig et al. 2012) which also includes reference to the Scatter-Gather-Reconcile function in the scientific collections management software system Specify. This software feature operates on data entry and searches out amongst other software users to retrieve information to suggest potential record links, but based on deterministic matches rather than probabilistic techniques. As per the scientific name matching software TAXAMATCH, Scatter-Gather-Reconcile is a user-facing, human-scale operation rather than seeking to apply a high-scale approach to the problem.

Image analysis

Specimen digitisation can also generate images of specimens (*unstructured image data*), which when associated with *structured metadata* about the specimen subject can be used as training data in classification and image analysis research. Herbarium specimen images have been scored to assess insect damage for long term studies on herbivory (Meineke et al. 2019), used to train deep learning applications to predict the taxon of the subject and to detect specimen management characteristics (Schuettpelz et al. 2017). More recent research has started to use image classification to generate datasets for phenological research, incorporating the structured metadata from the specimen digitisation process (to gather the location and date of collection) along with classification results to assert phenological state (leaf-out, flowering, fruiting) (Willis et al. 2017).

2.3 Proposed work: Automating the construction of higher-order data representations from heterogeneous biodiversity datasets

This introduction to the field of biodiversity informatics and the survey of existing work show that the domain includes a rich set of data resources, ranging from hand-curated datasets to large auto-generated un-standardised datasets. These are currently loosely connected, but would form a rich resource for large-scale scientific analysis if better interconnected. These

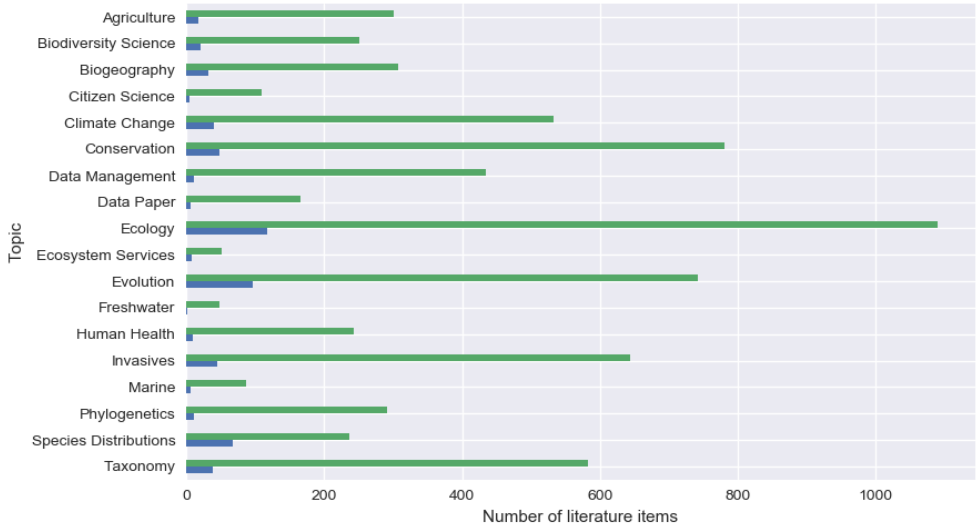


FIGURE 2.5: Categories of use of GBIF mobilised data. For each literature topic, the first bar shows the total number of literature items, the second bar shows the number of literature items that directly cite a data package mobilised by GBIF

resources could also support meta-analysis to understand trends and participation in the systematics process, which is an area of concern on a global scale. The digitisation and record-linkage task is considerable: automated approaches are likely to be needed as human-scale editorial data management and manual record-linkage will not practically scale to this size of problem.

Whilst considerable effort has been invested in data standards and techniques for representing, referencing and mobilising data, there is little comparable effort into the techniques for large-scale creation of references between data items. Data mobilised through the GBIF network and subsequently used in published analyses are monitored via a dataset citation process and literature tracking program. This means that it is possible to summarise the high-level research topics which cite GBIF mobilised data (figure 2.5). The breakdown of these research areas show that there is comparatively little reflective use of the aggregated dataset in data management applications.

This project uses botany as a content focus, due to the availability of standardised hand curated datasets and the wealth of digitised specimen material available. The primary aim of this research is to explore techniques to generate higher-order data representations to enable trends analysis.

A theme throughout the research will be to learn from the kinds of management techniques and research analyses applied smaller-scale hand-curated data, and to develop methods to apply these to larger-scale aggregated datasets. This will help to understand if the aggregation efforts have reached a critical mass, such that although incomplete, the aggregated

data can be used both to improve the data itself and the processes by which resources are made digital.

This research area is broken down into four sub-questions:

- Can data-mining tools be developed to uncover entities to allow reshaping of heterogeneous data to support trends analysis?
- What novel techniques can be designed to help correlate data-mined data entities with existing editorially-created entities?
- Can data-mined data entities help assess species discovery value, and help determine institutional and individual impact?
- Is it possible to re-integrate fragmented data across institutional boundaries, what data management efficiencies can be gained?

2.3.1 Approach

This section outlines the practical approach used to manage the research project.

The project is largely concerned with the identification and linkage of data entities in a particular scientific domain. To aid understanding and assessment of the contributions of each part of the research project to the overall picture of the domain, a visual context diagram (based on the concept map shown in figure 1.1) will be used to introduce each research chapter. Each visual context section highlights the relevant data entities and inter-relationships participating in the project. In addition to this conceptual, context-setting use of visualisation, a toolkit of practical, interactive visualisations was developed to aid data exploration and gather expert input at each stage of the project. As the use of this toolkit cross-cuts the individual research chapters, a summarised overview of the toolkit is included as an appendix (appendix A).

It is hoped that in addition to answering the specific research questions, this research project can help generate practical tools and techniques to aid exploration, management and understanding of biological specimens and literature data. With this aim in mind, the work has built on freely available open-source software, and the research and development process uses tools and techniques widely taught to scientific researchers working in software and data-intensive disciplines. Code has been developed in the *Python* language (Van Rossum and Drake Jr 1995), making extensive use of *scikit learn* for machine learning (Pedregosa et al. 2011), *pandas* for tabular data structures (McKinney 2010), *networkx* for graph data structures (Hagberg et al. 2008), *numpy* for numerical computing (T. E. Oliphant 2006), *scipy* for scientific computing (E. Jones et al. 2001–), *pomegranate* for probabilistic

programming (Schreiber 2018), *jupyter notebook* for literate programming (Kluyver et al. 2016), and *matplotlib* (Hunter 2007) and *seaborn* (Waskom et al. 2014) for charting and visualisation. All software and documentary outputs are managed using a revision control system (*git*), multi-step processes use explicit dependency management (*make*) and research and development activities have been managed in iterative development cycles.

Chapter 3

Preliminary analysis & overview of techniques

This preliminary analytical chapter is presented in two parts. The first demonstrates the kinds of trends analyses that are possible when interlinked data with formally-managed related entities are available for use. This is followed by an overview of the kinds of techniques available from the fields of machine learning and intelligent data analysis which may support the scaling up of this approach to larger scale datasets. A version of the research outlined in this chapter was published as a journal paper which investigated the correlations between name publication trends and recent rule changes to the processes by which scientists formally name plants. (The published version is available in appendix B.1.) Its presentation here uses a longer time frame than that used in the journal article and focusses on the use of the associated data entities - authors, the creators of new scientific names and publications, the containers for new scientific names - for trends analysis. An important component of this part of the research project is to reshape the data to allow investigation of the authors participating in events relevant to species discovery.

Later chapters will use these machine learning techniques to develop the analyses and to apply them in a wider context and at a larger scale, covering earlier stages in the systematic process: collection and annotation of specimens, in preparation for formal publication of scientific names.

3.1 Preliminary analysis

3.1.1 Visual context

The analysis included in this chapter utilises publication event data relating to the scientific names of higher plants, which are explicitly linked (via human-scale editorial effort) to formalised entities for the containing publication and the author of the work.

Figure 3.1 provides a visual context for the scope of this chapter - showing that the trends analysis here uses data related to the **publication** of scientific

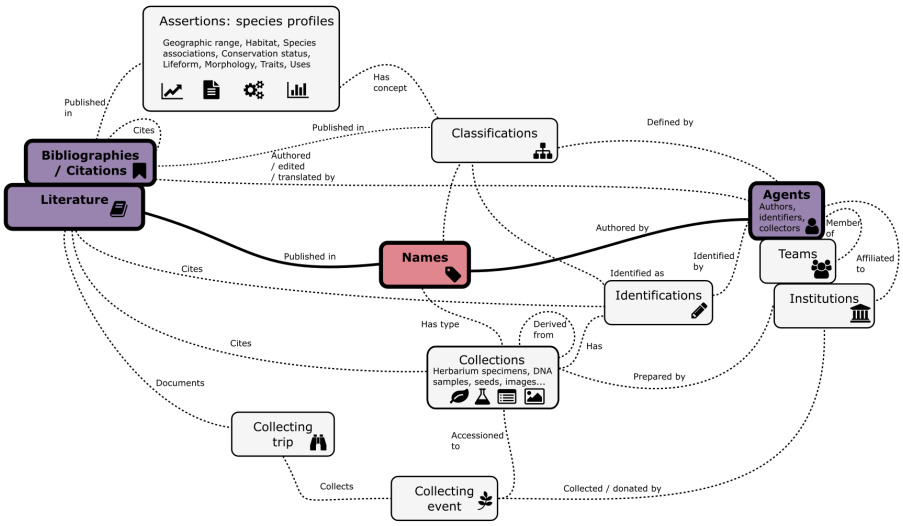


FIGURE 3.1: Visual context: e-publication

names, as conducted by the authors (**agents**) responsible for their definition and publication.

3.1.2 Introduction

Importance of scientific publication, particularly in nomenclature

Publication of results is one of the cornerstones of the scientific endeavour. Differences between scientific and general publishing were first articulated by Henry Oldenburg, who as Secretary of the Royal Society, established the first English-language scientific journal, Philosophical Transactions of the Royal Society (Knapp and Wright 2010). Oldenberg defined a number of functions for scientific publication: dissemination, registration, certification and archiving (which he named the "Minutes of Science") recognising the role of scientific publishing in informing the present and providing a record for future generations. Since Oldenberg's time, scientific (scholarly) publication has seen great change driven in part by increased interconnectivity of research communities, massive increases in funding for research and development since the middle of the 20th century, and key technological advances such as the Internet and digital publishing. These drivers have been described as having as big an effect as the replacement of parchment by paper, or the advent of mass printing technologies (Guédon 2001). The move away from print on paper to electronic-only publishing mechanisms fits with the move of much of scientific activity on-line. The pace of change in this area of scientific publishing is increasing, with more and more journals converting to on-line only publishing (including those relevant to the publication of

botanical research and new scientific plant names e.g., Evolution, New Phytologist, Biological Journal of the Linnean Society).

The naming of organisms is governed by the codes of nomenclature, which are managed separately for animals (the International Code of Zoological Nomenclature) and algae, fungi and plants (The International Code of Nomenclature for algae, fungi and plants). Discussions about publication changes in the electronic publishing era were therefore separate, although many of the drivers (increased speed of description) and issues raised (access to archival publications) were similar. This analysis focusses on the publication rule changes as relevant to botany (and governed under the International Code of Nomenclature for algae, fungi and plants - ICNafp). Decisions about updates to this code of nomenclature are made at Nomenclature Sections of International Botanical Congresses (IBC) held every six years (Brummitt 2006), (Turland 2013).

Establishment of an index to aid navigating the published scientific literature

The botanical community expressed mixed feelings about the advent of electronic publication of names, the record of discussion at the botanical congress documented these as "hopes and fears" (Flann et al. 2014). For many groups of organisms, it is difficult to track the effects of changes in scientific practice, but vascular plants can be analysed using data from the International Plant Names Index (IPNI, www.ipni.org) which records the publication events for scientific names.

The IPNI project began as "Index Kewensis", and was originally funded with a £250 legacy from Charles Darwin in his will for the "establishment of an index of all plants''(Croft et al. 1999), (Lughadha 2004). It was conceived in a time when it was feasible for a scientist to own all the relevant literature for their field, but it was even then necessary to have a bibliographic index to avoid repeated reference to scattered primary sources. The Index captured the name, authorship and basic bibliographic details of published plant names. Its first output was published in 1893, covering the names published from 1753 (the year of publication of Linnaeus' Species Plantarum and the start date of botanical nomenclature) till the start of the Index Kewensis indexing effort (1885). Three further original volumes and a number of five-yearly supplements were created and published, and in 1983 the data were digitised to an electronic database format. In the late 1990s Index Kewensis was amalgamated with the Gray Card Index (GCI) maintained by the Harvard University Herbaria and the Australian Plant Names Index (APNI) to form the International Plant Names Index (IPNI, www.ipni.org) see (Croft et al. 1999)). This dataset is accessible online and is continuously updated by a dedicated editorial team as new names are published;

approximately 8,000 new name records are added each year. The dataset is a valuable resource for trends analysis regarding the time, location and method of publication of new plant names.

This analysis looks at participation rates for authors and publishers in botanical nomenclature governed by the ICNafp (McNeil et al. 2012) using data from IPNI to examine whether the hopes - increased participation, increased rate of description - or fears - avalanche of sloppy nomenclature, proliferation of new on-line journals - have been realised.

3.1.3 Methods and materials

Data

The IPNI database contains basic bibliographic information about the place of first publication of vascular plant names (ferns and fern allies, conifers, cycads and flowering plants). Nomenclatural acts are recorded by a editorial team, who read literature, and record details of the name, its authorship and the date of effective publication into the database system. The authorship of the nomenclatural act is standardised using the principles laid out in Authors of Plant Names (Brummitt and Powell 1992) (also referenced under recommendation 46A of the ICNafp (McNeil et al. 2012)). Publication titles are also standardised by linking to an authoritative list.

Members of the editorial team apply the rules of the ICNafp and exercise nomenclatural judgement about the nomenclatural acts recorded. Annotations to indicate if an act contravenes the code (i.e. it is illegitimate, not effectively published, or not validly published) are added to the nomenclatural act record.

The database records are fully versioned, with date of application of each edit recorded.

Selection of data subset for analysis

Activity and emergence trends were calculated using a data subset of scientific name publication acts published between 1900 and 2015. This represents a timeframe with widespread social changes with impacts across science - world wars, the emergence of mass affordable travel and electronic communication - as well as some changes particularly relevant to the practices of systematics: the establishment of the UN Convention on Biological Diversity (in particular the Global Taxonomic Initiative in 1998) and changes in the nomenclatural codes to permit the use of English language descriptions and electronic publishing.

The most recent data included in the dataset are several years old - this ensures that more obscure titles have had a chance to be seen by the IPNI editorial team for the recording of nomenclatural acts. The lag time for some types of publications (e.g., small print-run journals and some books) can be

up to a year or more. Three years after the implementation of nomenclatural rule changes should give us enough of a range of samples to make an initial assessment of the impact of the most recent event (nomenclatural rule changes, implemented in 2012).

Analyses

The following analyses were conducted:

Authors - number active: the unique number of authors specified as members of the publishing author team in nomenclatural acts between 1900 and 2015 were counted, broken down by year.

Authors - number emergent: for all the authors active in the selected period (1900-2015), their date of emergence was calculated - this is the date when they were first recorded as a member of the publishing team of a nomenclatural act. This dataset was grouped by year of emergence to give a count for each year.

Publications - number active: as per the analysis for authors described above, this is the unique number of serial publications recorded as containing nomenclatural acts published between 1900 and 2015 were counted, broken down by year. A serial publication is defined as a multi-volume work.

Publications - number emergent: (as per the analysis for emergent authors described above) - for all serial publications active in the selected period, their date of emergence was calculated - this is when they were first recorded as containing a nomenclatural act. This dataset was grouped by year of emergence to give a count for each year of the study.

3.1.4 Results

The emergence and participation trends for authors and publications are shown in figure 3.2. These appear to show dips in activity during the two world wars in the first half of the twentieth century. Once activity levels regained, the next change is circa 1970 - recognised as the start of the era of affordable mass-travel, and post 2000 - just after the implementation (in 1998) of the Global Taxonomy Initiative and also when the internet became more widely available.

The data show no sudden difference in the emergence or participation of either authors or serials after the starting date for e-publication in 2012 (indicated by the vertical line in the plot). The apparent dramatic dip in the last year of the sample is likely due to the lag in discovery of nomenclatural acts published in less accessible media (e.g., small print-run local journals or books).

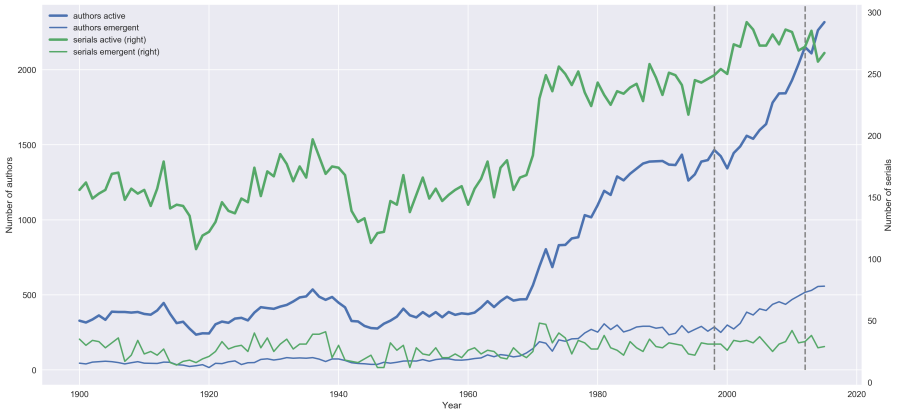


FIGURE 3.2: Number of authors and serials active and emergent / year (1900-2015)

3.1.5 Discussion

Recent discussions about the participation rates in systematics have recognised the "taxonomic impediment" (as defined by the Convention on Biological Diversity (Convention on Biological Diversity 2007)) and the need to support the communities working in the field through adoption of e-dimensions to their work (Scoble 2008). In botany, the most recent attempt to overcome the impediment was the adoption of e-publication and the scrapping of the Latin language requirement, which were expected to encourage more people to participate in the naming of species.

This discussion section is structured around "hopes and fears" regarding participation and access as discussed at the 2011 botanical congress, when the most recent "impediment-challenging" change was made, focussing on those areas relevant to participation.

Hopes: inclusivity: The data for recent years pre and post the rule changes do not show any upward trends in the numbers of authors actively publishing nomenclatural acts, nor in the number of people involved in the authorship of botanical nomenclature. (The data do not show any decrease in these measures either). This short term trend is only a snapshot of the longer term trend seen (for a smaller plant related dataset) by previous authors (Joppa et al. 2011). Anecdotally more authors appear to be associated with plant names, but further analysis of these trends is required. Biographical data on the authors of nomenclatural acts is not routinely collected by IPNI, and new efforts will be needed to ascertain if the community is truly changing.

Fears: a flood of nomenclatural acts, lessening quality: The data do not show a flood of nomenclatural activity creating "bad taxonomy" since the acceptance of e-publication. The numbers of journals continuing to be active in the process of publishing botanical nomenclature has remained more or

less constant (see figure 3.2) and there has not been a dramatic upsurge in the establishment of new journals.

3.1.6 Conclusions from the example analysis

This analysis has used an editorially created dataset to conduct trends analysis on non-primary allied entities (authors and publications) related to the main content of the dataset (nomenclatural events relating to the creation of new species names). The trends show changes relating to major world events, including the diminishing output due to the effects of the second world war - but the effect of deliberate changes to scientific practice, such as recent nomenclatural code adjustments, are less discernible in this dataset. Publication is the end result of a (sometimes elongated) scientific process, and it is often used for trends analysis due to easy accessibility and dedicated management. The use of publication data in the assignment of scholarly credit also contributes to the effort taken in the management and recording of the data.

The name publication dataset used here has supported a number of meta-analyses about the practices of systematics, including the time gap between specimen collection and name publication (Bebber et al. 2010), the numbers of authors participating in names publication (Bebber et al. 2013) as well as examinations of wider social trends such as the participation of different genders in scientific publishing (Lindon et al. 2015). Many of these discussions acknowledge that scientists working in systematics conduct many different activities alongside the description and publishing of new species (Joppa et al. 2011), (McDade et al. 2011). The data associated with these different activities is much larger scale and much less formally organised. The next section will discuss the application of machine learning techniques to facilitate the use of this much larger dataset in similar trends analyses.

3.2 Overview of machine learning techniques

This section aims to give an overview of the kinds of techniques available from the fields of machine learning and intelligent data analysis (Hand 1997) (Berthold et al. 2010) which can help address the computational challenges inherent in the "scaling-up" of analytical capacity in biodiversity informatics. These challenges include the handling of source data which are larger scale, less complete and more heterogeneous than those data sets used to date; it is necessary to overcome these challenges to properly investigate trends in biodiversity science.

"Machine learning" encompasses statistical techniques which are used to discern patterns from data. Techniques are broadly split between supervised -

those which require labelled data - and unsupervised. Many of the techniques require domain expertise to optimise input and verify outputs - this expert input is the key to "intelligent data analysis" (Hand 1997).

3.2.1 Feature definition and encoding

The discussion in the previous chapter on the representation, harvesting and aggregation of biodiversity data showed that these activities were reliant on data standards, ensuring that data drawn from separate sources are represented in similar ways, enabling the use of multiple separate data-sources in a single analysis. This section builds on the data standards used in biodiversity informatics to define some key terms relating to the representation of data in machine learning. Feature types and examples from biodiversity informatics are outlined in table 3.1.

TABLE 3.1: Feature types, with examples from biodiversity informatics

Feature type	Description	Biodiversity example
Categorical	Values drawn from a fixed set of possible values	Country codes
Continuous	Numeric values which may range between two fixed points	Dimensions: length, height, weight etc
Discrete	Numeric values which only take whole numbers	Counts: number of specimens observed
Ordinal	Values which fit into an ordered scale	Conservation status: ranges from *least concern* to *extinct* (International Union for Conservation of Nature et al. 2001)
Textual	Free text	Descriptions of specimens and collection locations
Temporal	Date or datetime depending on granularity	Date of collection
Geospatial	Points, polygons	Coordinates of collection point, distribution of a species

Some machine learning algorithm implementations can only deal with numeric features, these require techniques to generate numeric features from other feature types. A common technique for categorical features is *one hot encoding*, which translates a categorical variable into multiple features represented as an array of bits (see table 3.2). Similar techniques are used when dealing with textual data, where a string of text is tokenised into a list of terms, and the term list is encoded as a bit matrix.

TABLE 3.2: Feature encoding, using one-hot encoding to represent categorical features

Sample id	Country	Country_BR	Country_EC	Country_PE
1	PE	0	0	1
2	PE	0	0	1
3	BR	1	0	0
4	EC	0	1	0

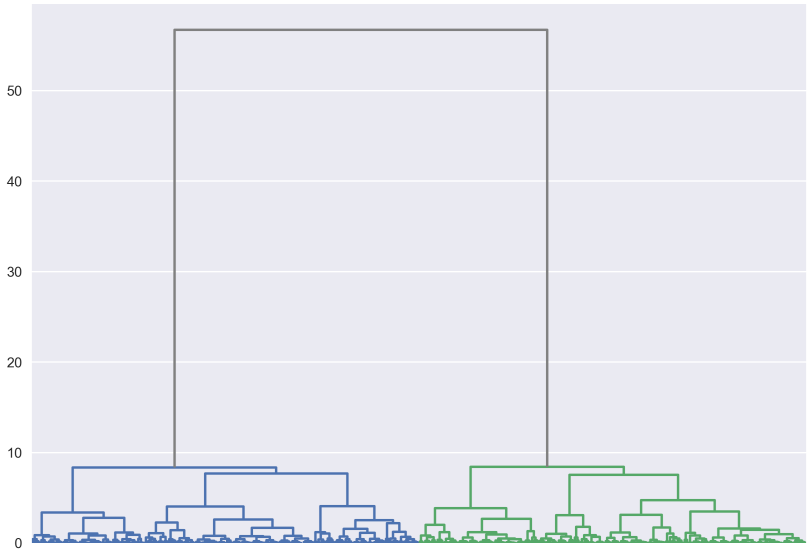


FIGURE 3.3: Hierarchical clustering - sample dendrogram, classifying the generated data shown in figure 3.4 (a). At the lowest level of the dendrogram are the leaves - where each sample is in its own class, at the top all samples are in a single class. The placement of a horizontal line through the dendrogram to separate the samples into a number of subsets is determined by finding the longest unbroken vertical connector. This shows that the data are clustered into two subsets.

3.2.2 Clustering

Clustering is a technique which takes a dataset of samples and uses their features to subdivide the samples into meaningful subclasses. As the input data is unlabelled, this is therefore an *unsupervised* learning problem. Here, three kinds of clustering approaches are introduced: hierarchical, prototype and density based:

Hierarchical clustering algorithms generate a dendrogram structure (figure 3.3) which represents the range of clusters in the dataset, ranging from the root of the tree (a single cluster encompassing all samples) to the leaves of the tree (a complete set of clusters, each containing a single sample). The direction of construction of the dendrogram is relevant, with some techniques working top down (*divisive*) and some working bottom up (*agglomerative*). Whilst hierarchical clustering does not require the user to specify a value for the parameter representing the target number of clusters (usually denoted as k), there is still a need to define a termination condition - the level at which the tree is judged to represent a meaningful clustering of the dataset (an example heuristic for this is outlined in figure 3.3). Hierarchical clustering is generally only tractable for smaller datasets.

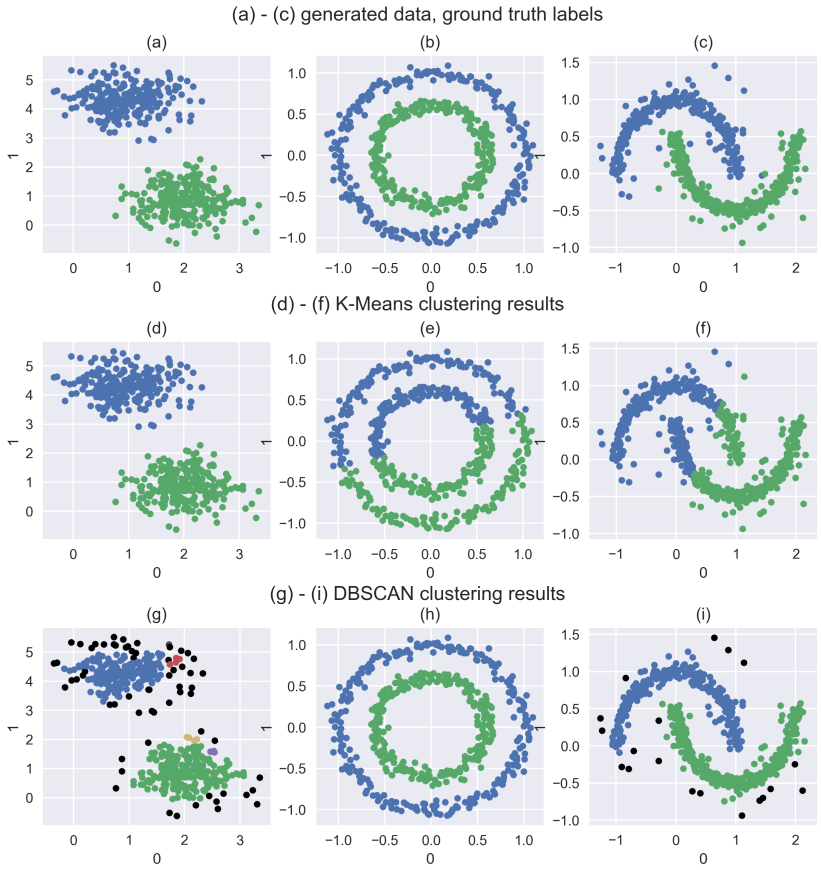


FIGURE 3.4: Prototype and density clustering on generated datasets of various shapes

Prototype clustering handles larger datasets better than hierarchical clustering, as a large number of sample points can be represented by a synthetic *prototype* point, allowing subsequent calculations to be generated using this new point. *k-means* is outlined as a sample prototype clustering algorithm in the section below. As prototype clustering methods do not produce a dendrogram (which can be examined to better understand cluster assignment), an analysis technique called *silhouette analysis* is also outlined, as a method by which the assignment of k-means cluster results can be visualised.

The main drawback of prototype clustering is that the use of prototype points favours clusters of regular shapes. *Density based* clustering is a technique which permits the identification of oddly shaped clusters (e.g. elongated traces), and can account for noise points in the input data set. *DBSCAN* is outlined as a sample density based clustering algorithm below, and the performance of k-means and DBSCAN on a range of generated datasets is shown in figure 3.4.

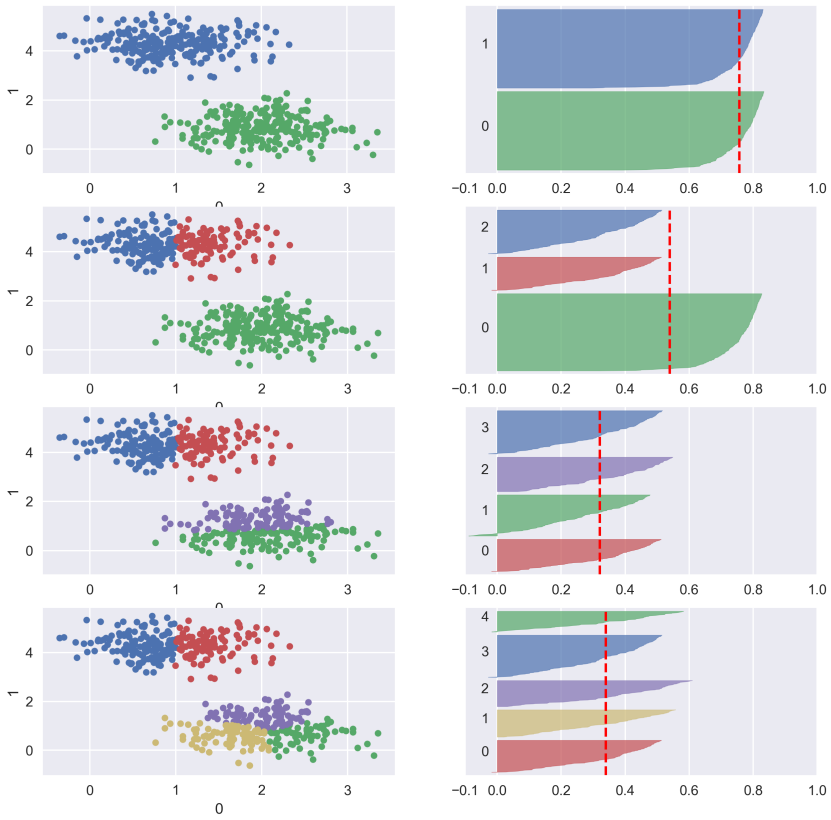


FIGURE 3.5: Silhouette plot for k-means cluster results, from top to bottom clustering with k values of 2, 3, 4 and 5

k-means

k-means clustering aims to separate the input dataset into k sets, minimising the variance in each cluster. The parameter k must be supplied as an input parameter. The algorithm uses two steps - assignment (points are assigned to a cluster) and update (new mean is calculated from the assigned points). The algorithm is defined in the listing 3.1. The results of a k-means clustering can be analysed graphically with a silhouette plot (Rousseeuw 1987), which represents the similarity of each sample to its own cluster compared to its similarity to the other clusters. The algorithm for the creation of silhouette scores is given in the listing 3.2, and a range of silhouette plots for k-means cluster results is shown in figure 3.5.

Algorithm 3.1: KMeans

Input : points, k, max_iterations
Output: labelled_points

```
Initialize cluster centroids
centroids = list(k)
labelled_points = list(len(points))
for centroid in centroids do
    centroid = select_random(points)
end
Compare each point to cluster centroids and allocate cluster ids
for point in points do
    Let point_index = index of point in points
    distances = list(k)
    for centroid in centroids do
        distances.append(dist(point, centroid))
    end
    Let min_dist_index = index of minimum distance
    labelled_points[point_index] = min_dist_index
end
changed = False
for iter in max_iterations do
    Update centroids: replace centroid with mean of assigned points
    for centroid in centroids do
        Gather all points assigned to this centroid
        Calculate mean of points
        Replace centroid with mean
    end
    for point in points do
        Let point_index = index of point in points
        distances = list(k)
        for centroid in centroids do
            distances.append(dist(point, centroid))
        end
        min_dist = min(distances)
        current_label = labelled_points[point_index]
        curr_dist = dist(point, centroids[current_label])
        if min_dist < curr_dist then
            Let min_dist_index = index of minimum distance
            labelled_points[point_index] = min_dist_index
            changed = True
        end
    end
    Check if the loop should terminate
    if changed == False then
        break
    end
    changed = False
end
return labelled_points
```

Algorithm 3.2: silhouette (Rousseeuw 1987)

Input : points, labels
Output: silhouette_values

```
silhouette_values = list(len(points))
cluster_labels = unique(labels)
num_points = len(points)
for point in points do
    Find assigned cluster
    assigned_cluster = labels[point_index]
    cluster_points = findPointsByClusterLabel(points, assigned_cluster)
    local_points = cluster_points.remove(point)
    Find distances to local points
    dists = list(len(local_points))
    for local_point in local_points do
        dists.append(dist(point, local_point))
    end
    Calculate mean local distance
    avg_local_dist = mean(dists)
    Calculate mean distance to each remote cluster
    cluster_dists = list(len(cluster_labels) − 1)
    for cluster_label in cluster_labels do
        if cluster_label <>point_label then
            remote_points = findPointsByClusterLabel(points, cluster_label)
            dists = list(len(remote_points))
            for remote_point in remote_points do
                dists.append(dist(point, remote_point))
            end
            avg_remote_dist = mean(dists)
            cluster_dists.append(avg_remote_dist)
        end
    end
    min_remote_cluster_dist = min(cluster_dists)
    Let min_remote_cluster_dist_index = index of min_remote_cluster_dist
    neighbour_i = cluster_labels[min_remote_cluster_dist_index]
    point_silhouette = (min_remote_cluster_dist − avg_local_dist) / max(avg_local_dist, min_remote_cluster_dist)
    silhouette_values.append(point_silhouette)
end
return silhouette_values
```

DBSCAN

DBSCAN (Ester et al. 1996) is a clustering technique originally developed for use with large spatial datasets. Several of the characteristics of the source datasets for which this algorithm was designed are equally applicable to the biodiversity domain. The authors of the DBSCAN algorithm note three requirements for clustering algorithms as applied to (large) spatial datasets (Ester et al. 1996):

1. Minimal requirements of domain knowledge to determine the input parameters, because appropriate values are often not known in advance when dealing with large databases.
2. Discovery of clusters with arbitrary shape, because the shape of clusters in spatial databases may be spherical, drawn-out, linear, elongated etc.
3. Good efficiency on large databases, i.e. on databases of significantly more than just a few thousand objects

The DBSCAN algorithm is based on the observation that *clusters* (irrespective of their shape) consist of fairly *dense* regions of points, and *noise* points are those found in areas of much lower density.

The following definitions and pseudocode are drawn from (Ester et al. 1996):

Given a database D of points of a k-dimensional space S, where a distance function for two points p and q is denoted by *dist(p,q)*, the following definitions can be made:

1. The *eps_neighbourhood* of point p is denoted by $N_{eps}(p)$ is defined by $N_{eps}(p) = \{q \in D \mid dist(p,q) \leq Eps\}$
2. A point p is *directly density reachable* from a point q wrt *MinPts* and *Eps* if:
    1. $p \in N_{Eps}(q)$
    2. $\mid N_{Eps}(q) \mid \geq MinPts$
3. A point p is *density reachable* from a point q wrt *MinPts* and *Eps* if there is a chain of points $p_1, \cdots, p_n, p_1 = q, p_n = p$ such that p_{i+1} is *directly density reachable* from p_i
4. A point p is *density-connected* to a point q wrt *Eps* and *MinPts* if there is a point o such that both p and q are *density reachable* from o wrt *Eps* and *MinPts*
5. **Cluster definition**: Let D be a database of points. A *cluster* C wrt *Eps* and *MinPts* is a non-empty subset of D satisfying the following conditions:

1. $\forall p, q$: if $p \in C$ and q is *density reachable* from p wrt *Eps* and *MinPts* then $q \in C$ (*Maximality*)
2. $\forall p, q \in C$: p is *density-connected* to q wrt *Eps* and *MinPts* (*Connectivity*)

6. **Noise definition**: Let $C_1 . \cdots, C_k$ be the clusters of the database D wrt parameters Eps_i and $MinPts_i, i = 1, \cdots, k$. *Noise* is defined as the set of points in the database D not belonging to any cluster C_i ie *noise* = $\{p \in D \mid \forall i : p \notin C_i\}$

The DBSCAN algorithm is defined in the listings 3.3 and 3.4 .

3.2.3 Graph analysis

The methods listed above have used data inputs in the form of matrices - data grids where each row contains a sample, and each sample is composed of several features of different types. This technique uses a different kind of data representation, which is useful for highly interconnected data: a graph. A graph is a data structure composed of *nodes* and *edges*. Both nodes and edges can hold features, and edges may have a special property to indicate the direction of the linkage between the nodes. The basic elements of a graph are shown in figure 3.6

Graphs are a natural representation for highly interconnected data, and support analyses at different levels of granularity. Many recent applications of graph techniques have been in social network analysis - researching the interconnections between people, teams and institutions. Scientific activities such as co-authorship and citation have been analysed with graph techniques.

As graph structures are easy to visualise, they are also useful as a data exploration technique. The use of graph data structures can also support unsupervised learning, particularly community detection - the partitioning of the graph into sub-regions of more densely interconnected nodes. When the graph is a weighted network, these partitions can be assessed using a modularity score (Blondel et al. 2008):

$$Q = \frac{1}{2m} \sum_{i,j} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j)$$

Where A_{ij} is the weight of the edge linking nodes i and j, $k_i = \sum_j A_{ij}$ is the sum of the weights of the edges attached to node i, c_i is the community to which node i is assigned, the δ function $\delta(u, v)$ is 1 if u=v and 0 otherwise and $m = \frac{1}{2} \sum_{ij} A_{ij}$.

This research project uses the Louvain community detection algorithm (Blondel et al. 2008) (defined in listing 3.5), which takes as its input a weighted graph. The algorithm consists of two phases, repeated iteratively:

Algorithm 3.3: DBSCAN (Ester et al. 1996)

Input : points, eps, minPts
Output: labelled points

```
noiseClusterId=-1
unclassified = None
clusterId = noiseClusterId + 1
for point in points do
    if point.clusterId = unclassified then
        if expandCluster(points, point, clusterId, eps, minPts) then
            clusterId = clusterId + 1
        end
    end
end
Each point in points is labelled with a cluster id (or noise marker)
return points
```

Algorithm 3.4: DBSCAN expand cluster (Ester et al. 1996)

Input : points, point, clusterId, eps, minPts
Output: True/False

```
noiseClusterId=-1
unclassified = None
seeds = findPointsInRegion(points, point, eps)
if len(seeds) <minPts then
    point.setClusterId(noiseClusterId)
    return False
end
else
    for seed in seeds do
        seed.setClusterId(clusterId)
        if seed == point then
            seeds.delete(point)
        end
    end
    while len(seeds) >0 do
        currentPoint = seeds.first()
        regionPoints = findPointsInRegion(points, currentPoint, eps)
        if len(regionPoints) >= minPts then
            for regionPoint in regionPoints do
                if regionPoint.clusterId IN (unclassified, noiseClusterId) then
                    if regionPoint.clusterId == unclassified then
                        seeds.append(regionPoint)
                    end
                    regionPoint.setClusterId(clusterId)
                end
            end
        end
        seeds.delete(currentPoint)
    end
    return True
end
```

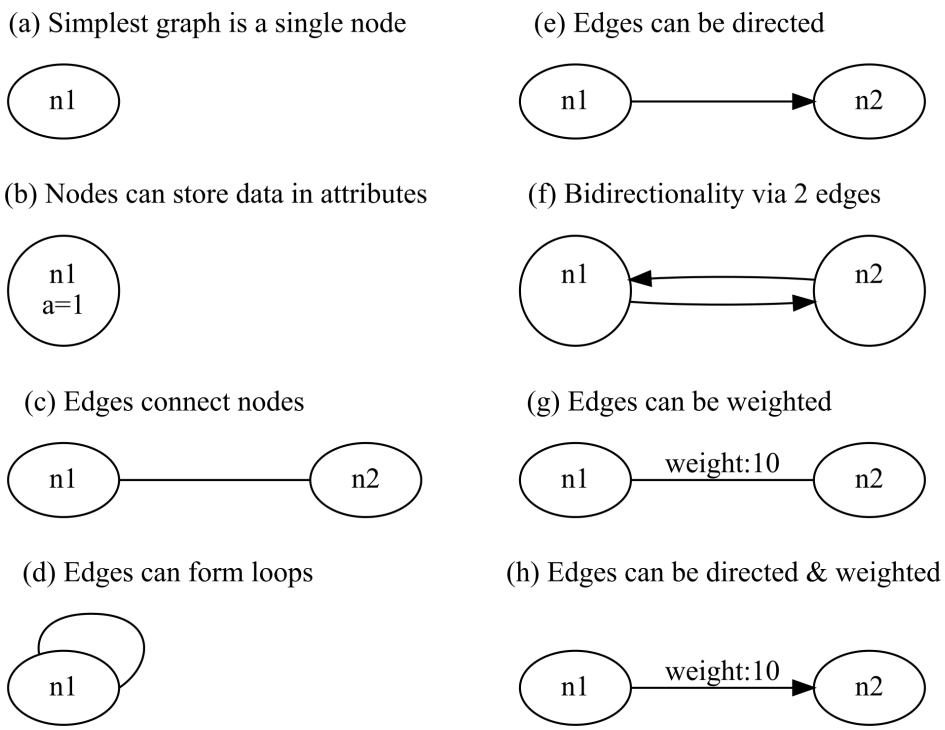


FIGURE 3.6: Graphs: basic elements (a - h)

first a community assignment phase, where communities are detected from a weighted graph (defined in the listing 3.6), then a rebuild phase, where a graph is re-built at community level (defined in the listing 3.7). The modified graph resulting from the rebuild phase is fed back into the initial (phase 1) community detection algorithm. The process terminates when no further modularity gain can be made. Pseudocode and definitions are reproduced from (Blondel et al. 2008).

Efficiency in the algorithm is given by calculating the modularity gain ΔQ given by moving an isolated node i into a community C by the following calculation:

$$\Delta Q = \left[\frac{\sum_{in} + k_{i,in}}{2m} - \left(\frac{\sum_{tot} + k_i}{2m} \right)^2 \right] - \left[\frac{\sum_{in}}{2m} - \left(\frac{\sum_{tot}}{2m} \right)^2 - \left(\frac{k_i}{2m} \right)^2 \right]$$

Where $\sum_{in}$ in the sum of the weights of the edges inside C, $\sum_{tot}$ is the sum of the weights of the links incident to nodes in C, k_i is the sum of the weights of the edges incident to node i, $k_{i,in}$ is the sum of the weights of the edges from nodes in i to nodes in community C, and m is the sum of all the weights in the network.

Algorithm 3.5: Louvain (Blondel et al. 2008)

Input : graph
Output: community_graph

```
comparison_modularity = 0
while True do
    community_graph = community_assignment(graph) //Algorithm 3.6
    community_graph = rebuild(community_graph) //Algorithm 3.7
    iteration_modularity = modularity(community_graph)
    if iteration_modularity ≤ comparison_modularity then
        break
    else
        comparison_modularity = iteration_modularity
end
return community_graph
```

Algorithm 3.6: Louvain community assignment (Blondel et al. 2008)

Input : graph
Output: graph

```
Initialise by placing each node in its own community
for i in graph.nodes() do
    for j in neighbours(i) do
        Test move this node (j) into community of (i)
        Calculate modularity delta and save in modularity_deltas
    end
    if max(modularity_deltas) >0 then
        Let j = index of maximised modularity change
        Execute actual move of node (j) into community of (i)
        actual_move(i,j)
end
return graph
```

Algorithm 3.7: Louvain rebuild (Blondel et al. 2008)

Input : graph
Output: community_graph

```
community_graph = new graph()
Create community nodes and their self-loops
for c in distinctCommunities(graph.nodes()) do
    c_nodes = graph.findNodesByCommunity(c)
    weight = sumInternalWeights(graph, c_nodes)
    c_node = c_graph.createNode()
    c_node.id = c
    community_graph.createEdge(c_node, c_node, weight)
end
Link community nodes, using data from original graph
for node in graph.nodes() do
    source = node.community
    source_nodes = graph.findNodesByCommunity(source)
    linked_nodes = graph.findLinkedNodes(node)
    for target in distinctCommunities(linked_nodes) do
        if source != target then
            target_nodes = graph.findNodesByCommunity(target)
            weight = sumEdgeWeights(graph, source_nodes, target_nodes)
            community_graph.createEdge(source, target, weight)
    end
end
return community_graph
```

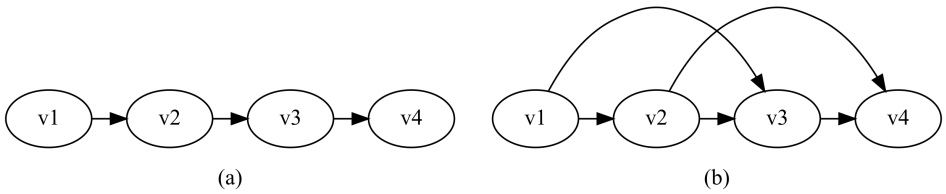


FIGURE 3.7: Markov chains: (a) first order, (b) second order

3.2.4 Temporal analyses: state-transition

Many of the concepts depicted in the biodiversity informatics concept map have a temporal component - name publication events have the date of publication, specimen collection events have the date of collecting. Temporally ordering data points to create time series allows the use of specialised time series modelling techniques, particularly Markov models (Barber 2012).

A time series can be defined as follows:

$$x_{a:b} \equiv x_a, x_{a+1}, ..., x_b$$

with $x_{a:b} = x_a$ for $b \leq a$

$$p(v1 : T) = \prod_{t=1}^{T} p(v_t | v_{1:t-1})$$

This describes a Markov chain: an ordered set of time periods, each of which can be in a particular discrete state, with probabilities of transition between different states.

$$p(v_t | v_1, ..., v_{t-1}) = p(v_t | v_{t-L}, ..., v_{t-1})$$

where $L \geq 1$ is the order of the Markov chain. The order of the Markov chain describes the contribution of previous states in the transition to the current state. In a first order Markov chain (figure 3.7a) only the previous state contributes to the transition. A second order Markov chain (figure 3.7b) receives contributions from the two previous states.

A hidden Markov model extends the Markov chain concept to represent a system which is composed of sequences of hidden and observable states. The model consists of the hidden states and the probabilities of transition between them (the Markov chain), and the observed states and their emission probabilities (the likelihood of witnessing the observed state, given a particular underlying hidden state). A hidden Markov model can be used to predict a number of different elements of the model: the present state (filtering), prediction of future states, inference of past states (smoothing) and the most likely hidden path (using the Viterbi algorithm). The structure of a

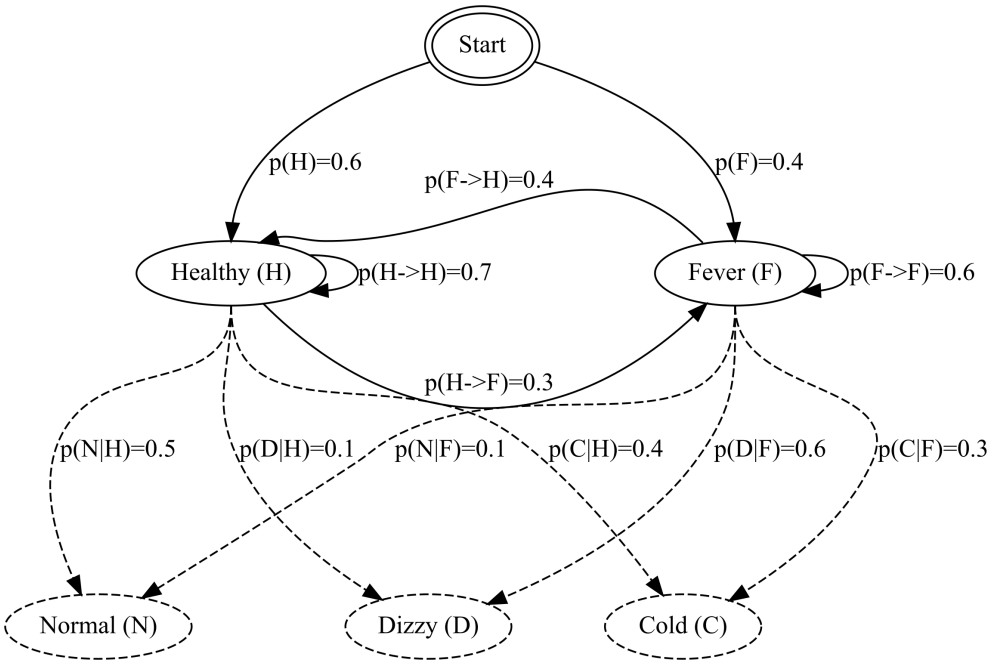


FIGURE 3.8: State-transitions and emissions for an example hidden Markov model. **Observable states**: normal, cold, dizzy; **Hidden states**: healthy, fever; **Start probabilities**: healthy: 0.6, fever: 0.4; **Transition probabilities**: healthy->healthy: 0.7, healthy->fever: 0.3, fever->healthy: 0.4, fever->fever: 0.6; **Emission probabilities**: healthy: normal: 0.5, cold: 0.4, dizzy: 0.1, fever: normal: 0.1, cold: 0.3, dizzy: 0.6

hidden Markov model (its transition and emission probabilities) can also be learned as an unsupervised task given a dataset of unlabelled states.

The Viterbi algorithm is used in unsupervised applications of hidden Markov models, where the model is constructed with estimated parameters for the start states, transition and emission probabilities. The algorithm outputs the most likely path of hidden states that would give rise to the observed states, and is defined in the listing 3.8.

An illustrated example uses a hidden Markov model and the Viterbi algorithm to model the diagnosis of an illness (an underlying hidden state) from reported symptoms (observable states). The components of the model are visualised in figure 3.8. The Viterbi algorithm uses a matrix data structure to calculate the path of states, having the dimensions K (number of states) x N (number of observations), which can be visualised as a trellis diagram (figure 3.9).

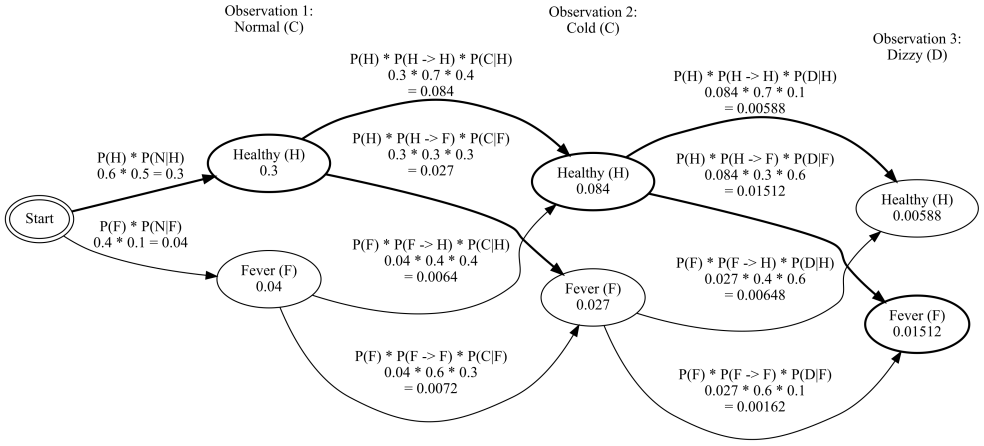


FIGURE 3.9: Trellis diagram for an example hidden Markov model, illustrating supervised use of the modelling technique. The state at each step is given by selecting the most likely outcome from the product of the following probabilities: P(state) * P(transition) * P(observation | state)

Algorithm 3.8: Viterbi

Input : Observation space O = $\{o_1, o_2, ..., o_N\}$
Input : State space S = $\{s_1, s_2, ..., s_K\}$
Input : Initial probabilities $\Pi = (\pi_1, \pi_2, ..., \pi_K)$
Input : Observation sequence Y = $(y_1, y_2, ..., y_T)$
Input : Transition probabilities A
Input : Emission probabilities B
Output: Most probable hidden state sequence X = $\{x_1, x_2, ..., x_T\}$

T_1 and T_2 initialised as K * T tabular data structures
X and Z initialised as arrays of length T
for $j \in \{1, 2, ..., K\}$ **do**
    $T_1[j, 1] = \pi_j * B_{jy_1}$
    $T_2[j, 1] = 0$
end
for *observation* $i \in \{2, 3, ..., T\}$ **do**
    for *state* $j \in \{1, 2, ..., K\}$ **do**
        $T_1[j, i] = \max_k(T_1[k, i-1] * A_{kj} * B_{jy_i})$
        $T_2[j, i] = \operatorname{argmax}_k(T_1[k, i-1] * A_{kj})$
    end
end
$Z_T = s_{Z_T}$
for $i \in \{T, T-1, ..., 2\}$ **do**
    $z_{i-1} = T_2[z_i, i]$
    $x_{i-1} = s_{z_{i-1}}$
end
return X

3.2.5 Classification

Classification is a *supervised* learning technique, which takes as input a labelled dataset and defines a process by which the features in the dataset can be used to predict class membership (as defined by the labels). When the problem domain has only two classes (which are usually represented as positive and negative) is known as *binary classification*. In contrast *multi-class classification* describes a problem domain where the number of possible class labels is greater than 2. *Multi-label classification* describes the situation where multiple class labels may be assigned to a single sample.

Supervised learning problems allow for the comparison of the predicted label with the correct label associated with the training data samples. This comparison is made by separating the labelled data into a training set (used to construct the classifier) and a test set (used to assess the classifier) - termed *cross-validation*. The most trivial form of cross validation simply retains a portion of the labelled data to act as the test set (known as *holdout*), more robust methods split the dataset into a number of subsets and repeatedly retain one subset as test data and train the classifier on the remainder, averaging the classification results across the multiple iterations of this process. This is known as *k-fold cross validation*, where k is the number of subsets into which the data is split, and therefore also the number of iterations of the train/test process. A refinement of k-fold cross validation attempts to ensure that the class balance is preserved across the folds, this is known as *stratified k-fold cross validation*.

Decision tree classification

A commonly used classification technique is a *decision tree*, a hierarchical structure which separates the input data based on feature states at each decision point. An induction algorithm for a decision tree structure is outlined in listing 3.9 (Berthold et al. 2010), and a graphical representation of a decision tree is shown in figure 3.10. This demonstrates that decision trees are easily interpretable by examination of their structure; ease of interpretation is often a factor in the selection of classification technique. The *Gini impurity metric* displayed in each node is the probability of a randomly selected sample from the set contained in the node being wrongly labelled, if the class label was randomly assigned according to the distribution of class labels in the subset.

Random forest classifiers

Random forest (Ho 1995) is an ensemble technique which generates multiple decision trees (a "forest" of trees) using subsets of the available features, and outputs as its prediction the most frequently occurring class prediction from

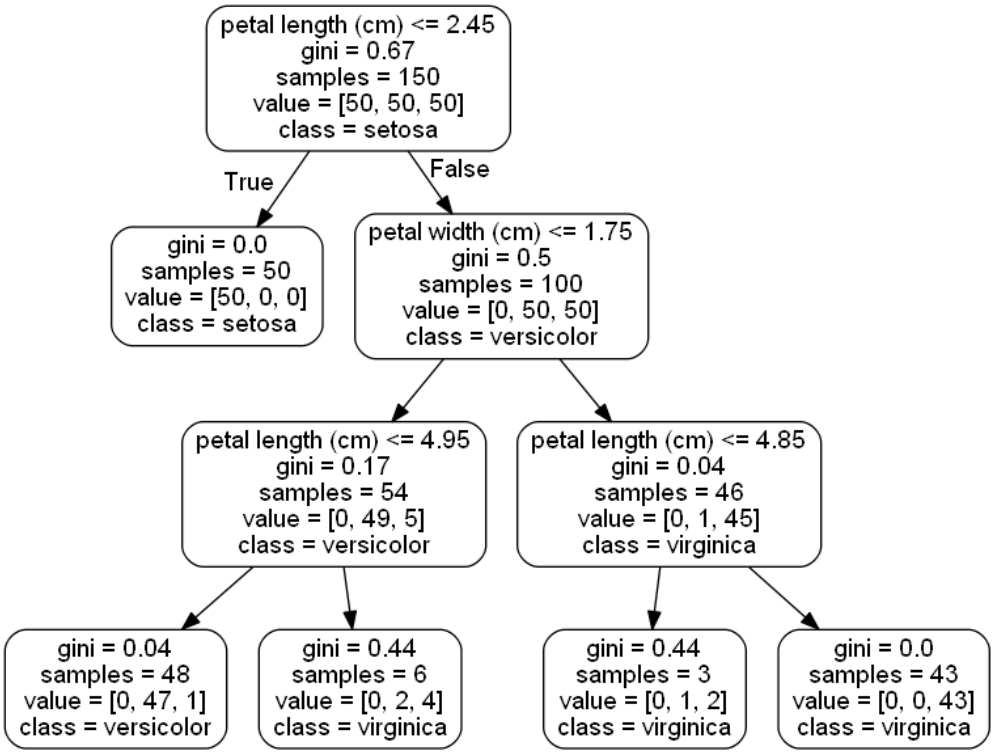


FIGURE 3.10: Graphical rendering of a decision tree classifier, built on the iris dataset and restricted to a maximum depth of 3. Root (the first) and internal (non-leaf) nodes display the condition used to make the split in the first line of the node caption. Leaf nodes do not split the data any further and therefore do not show the split condition. Gini (impurity) is the probability of a randomly selected sample from the set contained in the node being wrongly labelled, if the class label was randomly assigned according to the distribution of class labels in the subset. The number of samples, and the distribution of the number of samples across the possible classes is also given, along with the classification result.

Algorithm 3.9: Decision tree induction (Berthold et al. 2010)

Input : training data $\mathcal{D}$
Input : set of available attributes $\mathcal{A}$
Output: decision tree matching $\mathcal{D}$ using all or subset of $\mathcal{A}$

```
if all elements in D belong to single class then
    return Node with corresponding class label
else if A = θ then
    return Node with majority class label in D
else
    Select new attribute A in A which best classifies D
    Create new node holding decision attribute A
    for each split v_A of A do
        Add new branch below with corresponding test for this split
        Create D(v_A) ⊂ D for which split condition holds
        if D(v_A) = θ then
            return Node with majority class label in D
        else
            Add subtree returned by recursive call:
              BuildDecisionTree(D(v_A), A \ {A})
        end
    end
    return node
end
```

the ensemble. As a very large number of decision tree structures can be constructed to form the ensemble, it is not possible to examine and interpret the structure (as was possible with a simple decision tree described above) although improved accuracy is possible. It is important to understand the trade-offs between accuracy and interpretability when selecting a classification technique for a particular task.

Naive Bayes classifier

A *Naive Bayes* classifier is derived from Bayes rule, which relates the *posterior*, the probability of the class given the observation $P(C_i|X_j)$, to the product of the *class conditional*, the probability of the observation given the class $P(X_j|C_i)$ and the *prior*, the probability of the class $P(C_i)$, divided by the probability of the observation $P(X_j)$. This denominator ensures that the probabilities sum to 1. This explanation of the naive Bayes classifier is drawn from the outline available in (Marsland 2014).

$$posterior = \frac{class\_conditional \times prior}{observation}$$

or

$$P(C_i|X_j) = \frac{P(X_j|C_i)P(C_i)}{P(X_j)}$$

The denominator can be calculated by normalising over the classes, given that any observation X_k has to belong to a class C_i:

$$P(X_k) = \sum_i P(X_k|C_i)P(C_i)$$

The observation X is a vector of multiple features, addressed from x_1 to x_n Using *joint probability*, the term $P(X_k|C_i) \times P(C_i)$ can be represented as:

$$P(C_i, x_1, ..., x_n)$$

And this can be calculated using the *chain rule*:

$$P(C_i, x_1, ..., x_n) = P(x_1, ..., x_n, C_i)$$
$$= P(x_1|, x_2, ..., x_n, C_i)P(x_2, ..., x_n, C_i)$$
$$= P(x_1|, x_2, ..., x_n, C_i)P(x_2|, x_3, ..., X_n, C_i)P(x_3, ..., x_n, C_i)$$
$$= P(x_1|, x_2, ..., x_n, C_i)P(x_2|, x_3, ..., X_n, C_i)...P(x_{n-1}|x_n, C_i)P(x_n|C_i)P(C_i)$$

However this is simplified by assuming independence between the different features (this is the "naive" part of the naive Bayes classifier), and means that the probability of the class given the observation vector is the product of the individual probabilities of each of the observation features given the class:

$$P(C_i|X) \propto P(C_i, x_1, ..., x_n)$$
$$= P(C_i)P(x_1|C_i)P(x_2|C_i)...P(x_n|C_i)$$
$$= P(C_i)\prod_{i=1}^{n} P(x_i|C_i)$$

The *maximum a posteriori* (MAP) hypothesis is used as a decision function, to assign an observation to a class. This selects the class C_i where:

$$P(C_i|X) > P(C_j|X) \forall i \neq j$$

In multi-class classification it can be important to determine if the sum of the competing class probabilities outweighs the probability of the class selected by the MAP hypothesis. This calculation is known as the *Bayes Optimal Classification*, and is defined as $1 - P(C_i)$.

Analysis of classification results

As described above, a classification model can be analysed by splitting the labelled data and comparing the predicted class membership generated by the classification model with the actual class membership provided by the

original labels. There are a number of metrics which can be derived from these comparisons, for simplicity these are illustrated with binary classification examples.

A confusion matrix compares the true labels derived from the input data with the predicted labels generated by the classifier (see table 3.3). Correctly classified samples will be counted in the categories on the leading diagonal (top-left to bottom-right). The sum of the four categories gives the total number of samples, and the sums of the first and second columns respectively give the numbers of samples with condition positive and condition negative.

For binary classification problems, the total population, column sums and sub-totals in each of these categories can be used to derive a number of metrics:

- Accuracy = $\frac{\sum \text{true positive} + \sum \text{true negative}}{\sum \text{total population}}$
- True positive rate (also known as sensitivity or recall) = $\frac{\sum \text{true positive}}{\sum \text{condition positive}}$
- False positive rate = $\frac{\sum \text{false positive}}{\sum \text{condition negative}}$
- Specificity = $\frac{\sum \text{true negative}}{\sum \text{condition negative}}$

The true positive and false positive rates can be combined to generate the *receiver operator characteristic (ROC)*. In addition to calculating this as a single summary statistic to describe the performance of a classifier, it is common to generate a sequence of true positive and false positive rates with a varying threshold value (controlling the division between the two classes). These can be plotted to generate a curve, with the area under the curve calculated as a summary statistic. Multi-class classifiers can be reframed as a binary classification problem by constructing a classifier for each class (termed one-versus-rest), which enables the calculation of these kinds of metrics in the multi-class case. Figure 3.11 shows a set of receiver operator curves generated from a 10-fold cross-validation run of a one-versus-rest decision tree classifier on the *iris* dataset.

TABLE 3.3: Confusion matrix illustrating type I and type II errors

	Actual positive	Actual negative
Predicted positive	True positive	False positive (Type I)
Predicted negative	False negative (Type II)	True negative

3.2.6 Feature selection

When datasets are composed of very many features, it can be important to use *feature selection* to derive a useful subset. Counter-intuitively, the addition of extra features can lead to degradation of model performance due to the

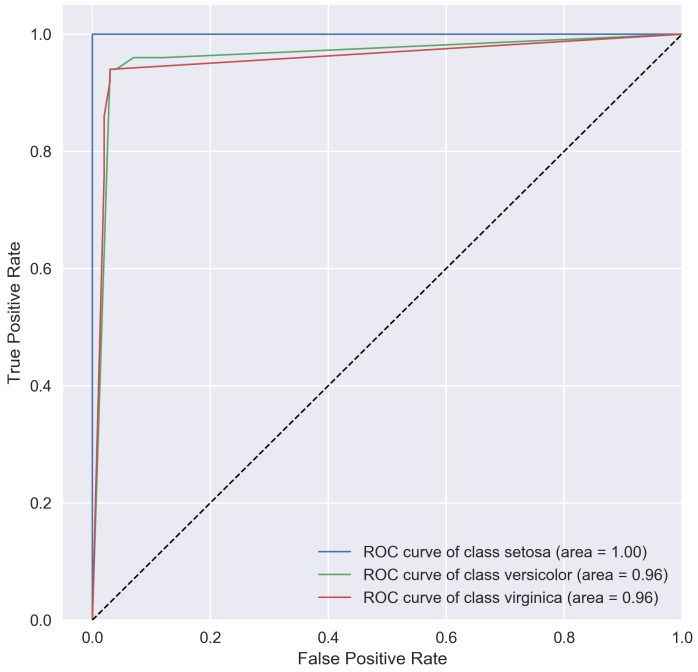


FIGURE 3.11: Receiver operator curve in the multiclass case (one versus rest), 10-fold cross-validation

curse of dimensionality: as the number of dimensions increases the ratio between the space occupied by the features and the containing space decreases. Feature selection can also guard against *overfitting* - generating a model which so closely matches the training data that it is unable to properly generalise to new samples. Using a dataset composed of fewer features can also be practically useful in terms of the computational requirements for training.

Univariate statistical tests such as *chi-squared* can be used to examine the relationship between each individual feature in a dataset and the target variable.

The feature dataset can be examined using *correlation analysis* to find pairs of correlated features. A feature which is highly correlated with another can be regarded as redundant, and eliminated from the feature set (as long as the correlated feature is retained). An example plot of pairs of features from the *iris* dataset is given in figure 3.12.

Tree-based methods (both simple decision trees and ensemble methods like random forests) can output *feature importance* based on Gini impurity (introduced in section 3.2.5).

Wrapper methods use a subset of features to build and evaluate a model, and use classification analysis metrics to assess its performance. These

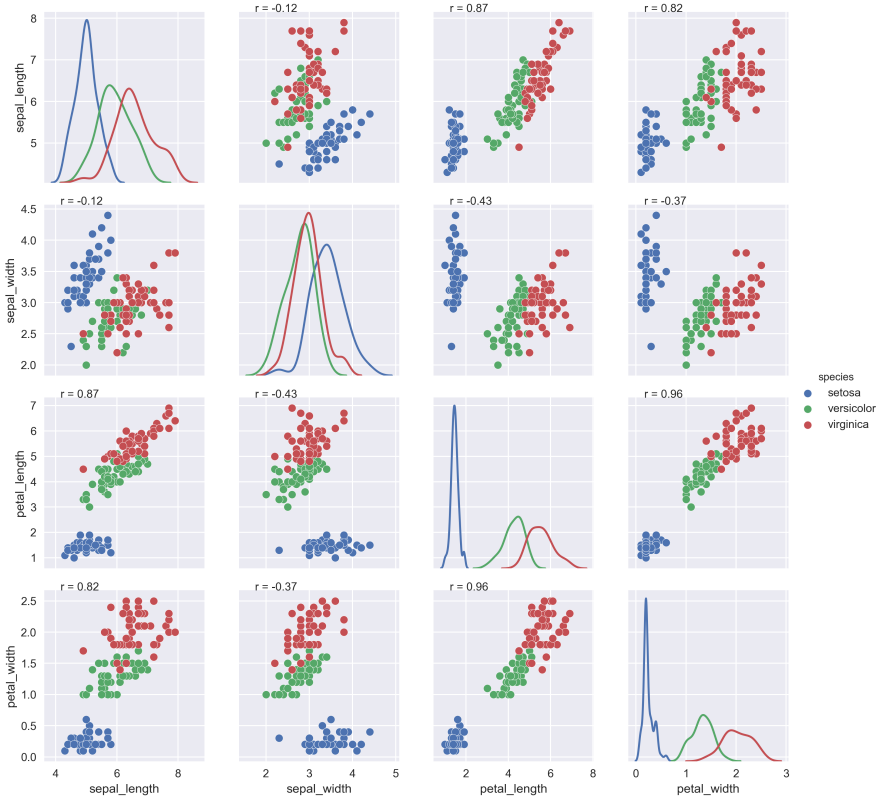


FIGURE 3.12: Correlation analysis of features in the iris dataset, with overall Pearson's correlation shown in each subplot

require search strategies to select feature subsets for evaluation, as an exhaustive analysis of all possible feature subsets will be computationally expensive for all but trivially-sized feature sets.

3.3 Conclusions and relevance to the next research chapter

This "preliminaries" chapter is included to demonstrate the use of species name publication data to determine trends. This analysis has been facilitated by the editorial management of the dataset to formally manage allied entities, which are used to reshape the data (for example to look at emergence trends for newly seen authors). Despite major streamlining changes to the publication process for new names, which were anticipated to expedite publication of results, no step change in output was detected. Why could this be? As outlined in the background chapter, the publication of names is the final - yet most visible - stage in the process of systematic research. The preceding stages comprise the collection of specimens, and the labelling of specimens with names to form hypotheses about their inter-relationships.

A logical expansion of this work would be to apply similar analyses to data relating to these earlier stages in the systematic process. The scope and

range of the data are much greater (an outline comparison is given in table 3.4). This change in scale, plus the lack of dedicated management mean that a similar analysis using collections data will be difficult due to the lack of standardisation in the larger collections dataset.

TABLE 3.4: Comparison of the scope and management of nomenclatural and specimen datasets

	Nomenclatural data	**Collections data (botany)**
Core entity	Scientific name	Specimen
Source	Editorially generated	Mass digitised opportunistically collected
Linkage	Date, Agent, Publication, Taxonomy, (Specimen)	Taxonomy, Georeference
Scale	c 1.6 million name records	c 60 million specimen records digitised; c 400 million physical specimens
Completeness	Complete at species level	Fractionally digitised
Management	Dedicated editorial team	Distributed
Standardisation	Single set of standards	Diverse set of standards

This chapter has shown that machine learning techniques may be applied to the data which has been collated in the biodiversity informatics domain. The next chapter will explore the use of these more automated methods to establish the entities and interlinks required for similar analyses on larger scale collections data, using a data-mining process composed of unsupervised learning steps (clustering and state-transition analysis) to help establish the necessary entities, rather than relying upon dedicated editorial management.

Chapter 4

Data-mining collectors and collecting trips from aggregated specimen data

This research chapter outlines the use of heterogeneous specimen collection data as a resource for data-mining higher order data representations. The data mining process developed here detects multiple new entities in the specimen dataset which allow the specimen data to be reshaped and repurposed for different kinds of trends analyses.

An early version of the research outlined in this chapter was published as a conference paper. (The published version is available in appendix B.2.)

4.1 Visual context

Figure 4.1 provides a visual context for the scope of this chapter - showing that the data-mining process developed here uses data related to collections to assert **agent**, **collecting event** and **collecting trip** entities. A comparison of the visual context of this chapter with that of the previous preliminaries chapter (in figure 3.1) shows that here the aim is to construct a similar set of entities (object: specimen, creator: collector agent and container: collecting trip - see table 4.1), but this work focusses on the larger scale and less formally managed physical specimen domain rather than the publication domain.

TABLE 4.1: Entity role comparison between name publication analysis and proposed specimen data-mining

Role	Name publication analysis	Specimen analysis
Object	Publication event	Collecting event
Creator	Author	Collector
Container (micro scale)	Publication	Collecting trip
Container (macro scale)	Career	Career

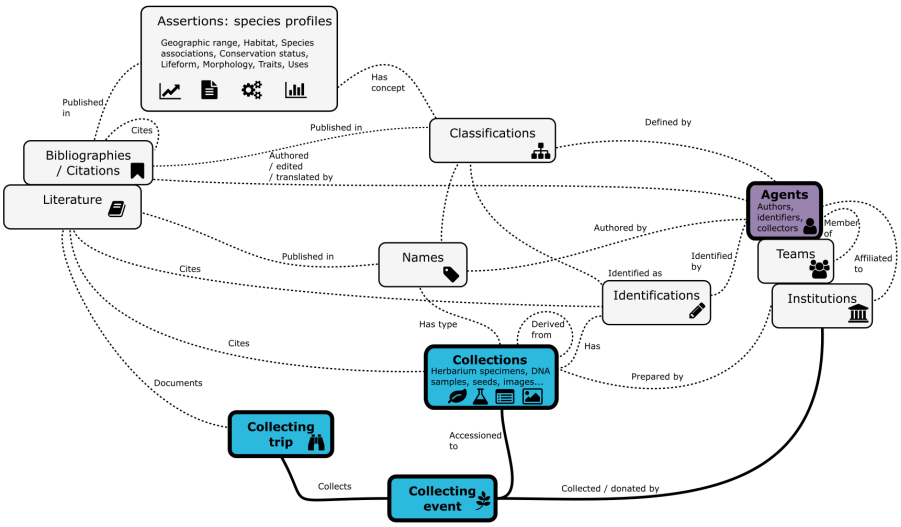


FIGURE 4.1: Visual context: data-mining

4.2 Introduction

Biological specimens collected over hundreds of years and held in natural history museums and herbaria are a rich reference source with which to understand the natural world, and to analyse its changes over time. Estimates of the total number of specimens vary between 2-3 billion specimens globally (Chapman 2005) (Hardisty et al. 2013). Only a small percentage have associated digital data. Aggregation initiatives such as the Global Biodiversity Informatics Facility (GBIF) harvest and mobilise digital specimen data: their data portal currently holds information on over 166.75M specimen records.[1] In order to aid the mobilization of the data, there has been an effort to develop standards regarding the representation of the data (Wieczorek et al. 2012), and references to it (Güntsch et al. 2017). These standards are important as due to the scale of the overall task, data have been digitised in a distributed fashion, at different rates and to different levels of completeness.

In addition to the structured data held on the specimens themselves, field collected specimens are often accompanied by a wealth of information about the collection site, habitat and associated species, logged in field books, which are also being digitised via literature digitisation initiatives.

Although plants are a comparatively well known group, and are well represented with digitised specimen data, species discovery is not yet complete, and approximately two thousand new species are described per year (*International Plant Names Index* n.d.). Not all species discovery is via field work: a sizable proportion of species discovery is conducted from pre-existing specimens already lodged in institutional collections (Bebber

[1] Numbers calculated from GBIF API call executed on 2019-11-05

et al. 2010). Estimates of the total number of plant species recognise the importance of species discovery from pre-existing collections and the use of collections data to plan species discovery in the field. An important component of systematic research is to understand the trends behind the rates of species description. Given the relationship between the collection of specimens, and the use of specimens in the description of new species, trends analysis of specimen collecting rates would contribute towards a better understanding of species discovery. Collection analyses to date have used closely defined data and institutional subsets - e.g. type specimens from a select set of institutions (Bebber et al. 2012) or have used summarised data resources covering collector activity rather than the actual specimen data (Penn et al. 2018).

The application of intelligent data analysis techniques could allow the use of the full specimen dataset. A more wide-ranging analysis would help meet two key aims: *data mobilisation* by better utilising and curating the existing data, and finding efficiencies that will help the digitisation process, and *data understanding* by uncovering patterns that will help plan future scientific effort as research is conducted with specimens or in the field.

The novel data-mining techniques demonstrated here detect new entities (*collector* and *collection trip*) from the duplicated, incomplete and variably transcribed specimen datasets, and are comparable to the data entities created in the editorial management of publication data. These can be used to draw together heterogeneous data, which has been recorded in different places, to different standards, in order to support and develop our understanding of a complex system - species discovery.

The remainder of the chapter is structured as follows: a background section further introduces the nature of the specimen data available by defining terms and outlining the specimen collection process, and a methods sections describes a data-mining process to detect collector and collection trip entities from raw specimen data. The data-mined entities are used to reshape the data and detect trends over time, these processes are also conducted with baseline groupings to provide a comparison. Results of the data-mining and trends analysis are shown, and ideas for further work are discussed.

4.2.1 Collecting practice

A *specimen* is a physical sample of biological material collected in the field. In botany, a collected sample may consist of multiple specimens, named *duplicates*. The *collecting team* is the team of collectors responsible for the *collecting event* (gathering and documenting the specimen), this team may include multiple collectors, referred to by personal name. The *primary collector* is the first listed member of the *collecting team*, and controls the *recordnumber* - a number given to the specimen in the field, usually sequential

and unique to the primary collector. Recordnumbers are locally managed, rather than centrally assigned. When duplicate specimens are collected, they are given the same recordnumber (Bridson 1998). A *collection trip* is a circumscribed period of specimen collecting activity - a sequence of collecting events conducted by a particular primary collector, focussed on a particular place and time. An *itinerary* is a list of the collecting localities visited by a primary collector in a collection trip, which may be documented in a *field book*, cross referenced to specimens via the recordnumber.

An *institution* is the holder of specimens for long term storage and reference consultation, usually natural history museums or herbaria (botanically focussed specimen collections). Institutions may distribute duplicate specimens to external partners, to form a globally distributed reference collection. *Digitisation* is the process of creating electronic records from the data held on the physical specimen, which may include *imaging* - the creation of a digital image of the specimen, and / or *geo-referencing* - the process of determining a latitude / longitude pair from a textual description of the collecting locality. This is necessary due to the historic nature of the specimen collection effort, which pre-dates the use of technologies such as hand-held global positioning systems (GPS) in the field. Duplicates are recognised as a source of data to speed the digitisation process (Tulig et al. 2012). The *collector name transcription* is the transcription of the collector names made when specimen data is read for digitisation. A single collector may have multiple varying collector name transcriptions, depending on the standards used in the different institutions, transcription errors and spelling mistakes. As the collector name transcription is necessary to identify duplicates (Tulig et al. 2012), variability in this data element impedes efficient use of the global specimen dataset. *Aggregation* is the collation of digitised specimen records from many institutions into a single data repository, represented using a structured *data standard*.

Primary biodiversity data derived from specimens have many applications in research (Chapman 2005) including species description and discovery. Specimen references in published literature are currently rather informal, and based on textual representations of the cited specimen, although there are moves to establish persistent identifiers to digitised specimens to improve traceability (Güntsch et al. 2017). Botanical collectors self-manage recordnumbers as a cross reference between physical material and information recorded in field notes / photographs etc. Recordnumbers are used as a component of an informal specimen identifier: as often seen in literature, specimen references are formed of collector name and recordnumber, sometimes also with year and institution code of holding institutions. The use of the personal name of the collector in this style of specimen identifier means that these are difficult to use at scale, as the

recording of personal names is very variable, with different abbreviation styles. When data are aggregated this problem is compounded - many different recording practices are seen, both at the individual name level (differing abbreviation styles) and the recording of multiple names which form a specimen team (different concatenation styles).

The collector - who makes decisions in preparatory planning and in the field about what to collect - is obviously a major contributor to species discovery (Bebber et al. 2012), and the collection trip has been recognised as a way to understand the accumulation of knowledge regarding the species found in a particular geographic area (Utteridge and de Kok 2006), as different collecting trips will have different motivations for field study. Investigating the characteristics of the collector has also been proposed (Utteridge and de Kok 2006) (Siracusa et al. 2018), including the differentiation between specialist (taxonomically focussed) and generalist collectors, and the application of these characteristics to the collecting trips conducted by the collector (Utteridge and de Kok 2006). Despite the scope for more advanced analyses of specimen data when differentiated and grouped by collector and / or collection trip, these entities are not formally managed - only the collecting team is a component of the main data standard used to share specimen data, which is supplied as a text transcription (Wieczorek et al. 2012) (GBIF.org 2018). Studies involving the grouping of specimen data by collector and or collection trip have had to use manual specimen record allocation (Bebber et al. 2012) to these groupings and / or expert knowledge (Utteridge and de Kok 2006), which limits scope. This means that the sequential nature of recordnumbers has been minimally exploited to date - but they have been used to create itineraries, by cross-referencing by hand between specimen data and field books (L. Smith and R. Smith 1967) as an aid to geo-referencing.

An example use of sequential recordnumber for a single collector is a test for a positive correlation - as a particular collector moves forward through time, their own personal sequential recordnumber increases (see figure 4.2(a)). Exploration of the data in this way can be useful to identify outliers (resulting from data transcription errors), but applications are limited due to the difficulty in initially identifying the set of specimens relating to a single primary collector, due to the variation in collector name transcriptions. Plotting a fuller corpus of specimen data (see figure 4.2(b)) - a sample of points from specimens collected in a single year), shows some visually distinguishable elongated "clusters", each of which correspond to the set of specimens collected by a particular primary collector and labelled with their own sequential recordnumber, which ascends over time.

This research uses the sequential recordnumber as a feature for clustering to detect the primary collector, thereby overcoming the variability

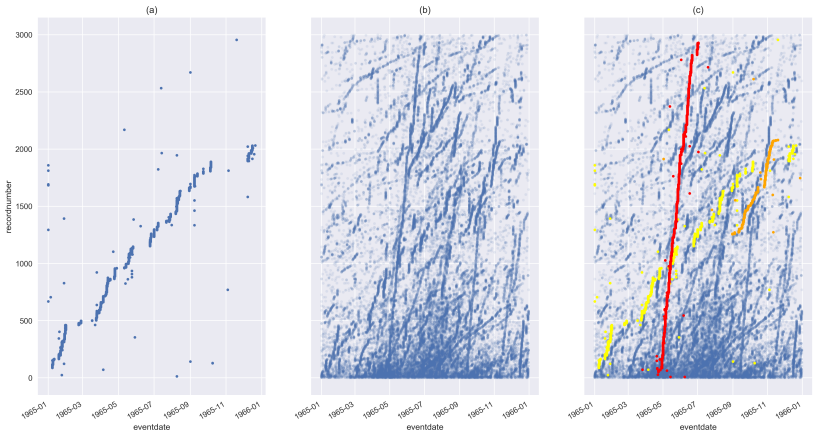


FIGURE 4.2: Use of allocation of recordnumber sequences over time to distinguish collectors. (a) plot of specimen datapoints where collector name includes the substring "Belem", (b) plot of all specimen datapoints in same region as Belem datapoints (year = 1965, recordnumber between 0 and 3000), (c) results of data-mining to distinguish individual collectors, top three most prolific collectors in this temporal and numeric subset shown, Belem cluster in yellow.

encountered when using the un-standardised transcription of personal names. The process employs a novel combination of data-mining techniques to detect these clusters, in order to identify higher order abstractions (collector, collection trip and collecting run) from an incomplete raw specimen dataset. These abstractions are recognised in the domain, but are absent from digital datasets. The recordnumber sequence is used to cluster specimens as they were gathered over time, resulting in a grouping by primary collector. The collector grouping is then used to detect the collection trips made by that primary collector. Finally, the collecting trips are subdivided into runs of days featuring intensive collecting activity, and runs of days when collecting activity is much reduced or absent. These three new abstractions are used to group the data and to define features at the grouped level, these features are used to examine trends in specimen collecting over time.

4.3 Methods and materials

4.3.1 Approach

The process described here was developed to allow visualisation of intermediate results at each stage, in order to allow an analyst to influence the design of the process. Visualisations were created as interactive scatter plots (as further described in appendix A), allowing the analyst to focus on

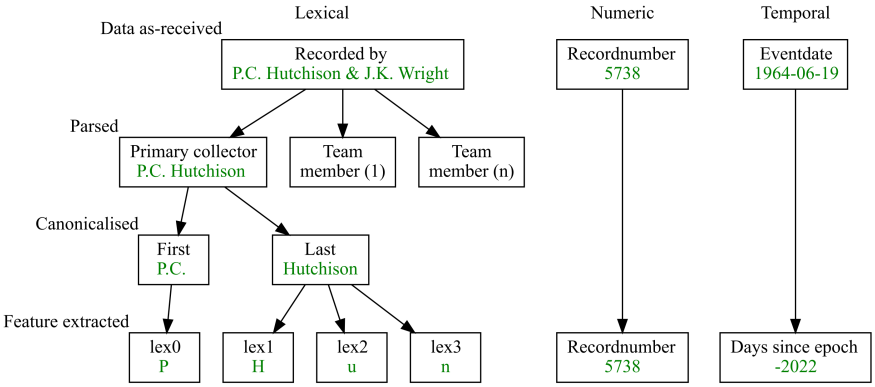


FIGURE 4.3: Data preparation process, to generate three classes of features: lexical (derived from the string transcription of the collecting team as found in the recordedby field), numeric (the numeric portion of the recordnumber, which is sometimes supplied with a alphabetic prefix) and temporal (the collecting eventdate represented as the number of days since 1970-01-01).

particular areas of the data, and to navigate to the GBIF data portal to examine the underlying specimen records.

4.3.2 Data

The main specimen dataset was downloaded from the Global Biodiversity Informatics Facility, encompassing data generated from botanical specimens (GBIF.org 2018). This large dataset (63.27M records) was used for exploratory data analysis with an analyst in order to design the process described here.

4.3.3 Data-mining process

Data-mining process steps (see algorithm listing 4.1):

- **Preparation for data-mining** - determining eligibility. Eligible records must have a precise eventdate (recorded to the day), a numeric recordnumber and a transcription of the primary collector name, from which lexical features can be extracted. See example preparation steps in figure 4.3.
- **Data-mining: find collectors**
  - DBSCAN clustering using lexical, numeric and temporal features. DBSCAN clustering was selected due to its ability to handle elongated cluster shapes, and to categorise some sample points as noise (see section 3.2.2). The features used were those shown in the figure 4.3: temporal, numeric and lexical features. Temporal and numeric features were used as-is, categorical lexical features were

Algorithm 4.1: detectSpecimenAggregations

```
Input  : specimens
Output: labelled_specimens
Preparation: determine eligibility and extract features
specimen_features=prepareSpecimens(specimens)
Data mining (1 of 3): find collectors
Initial clustering:
specimen_clusters = DBSCAN(specimen_features, eps_coll, min_size_coll)
Post-processing (1 of 2): break clusters based on lexical analysis
processed_clusters = new list()
for specimen_cluster in specimen_clusters do
    collectors_names = specimen.collectors_name for specimen in
      specimen_cluster
    if lexically_coherent(collectors_names) then
        processed_clusters.append(specimen_cluster)
    end
    else
        seeds = collectors_names
        separated_clusters = cluster(seeds)
        processed_clusters.append(separated_clusters)
    end
end
Post-processing (2 of 2): group clusters using shared features
Let grouping_strategies be a list of lists of features used to group specimens,
  arranged from narrow to wide scope
for grouping_strategy in grouping_strategies do
    Let groups be specimens grouped by grouping_strategy
    for group in groups do
        Let clusters be a list of the clusters represented in group
        Assess clusters for potential joins, if lexically coherent wrt
          collectors_names
    end
end
Label these processed_clusters as collectors
Data mining (2 of 3): find collecting trips
for collector in collectors do
    Let collector_specimens be all specimens in collector
    collecting_trip_clusters = DBSCAN(collector_specimens, eps_trip,
      min_size_trip)
    Label these collecting_trip_clusters as collecting_trips
end
Data mining (3 of 3): find collecting runs
for collecting_trip in collecting_trips do
    Let collecting_trip_specimens be all specimens in collecting_trip
    Use HMM state-transition analysis to subdivide collecting activity
end
Label specimens in contiguous collecting_run states as collecting_runs
return labelled_specimens
```

one-hot encoded to use 0 as absence, and 1000 as presence. eps was set to 300, and min_samples was set to 2.

– Post-process: break greedy clusters, examining canonicalised transcription of collector name.

  - Post process: block to join clusters, looking for candidate matches in neighbouring numeric / temporal / geographic space.
  - Assign collector identifier

- **Data-mining: find collecting trips**
  - Targeted DBSCAN: assert collecting trip identifier. For each collector detected in the previous step, pass all of their specimens into DBSCAN to subdivide their work into collecting trips (circumscribed periods of intense collecting activity). The features used as input to DBSCAN were rescaled: using weeks_since_epoch (rather than days_since_epoch as in the previous DBSCAN run) and the recordnumber was divided by 100. eps was set to 3, and min_samples was set to 5.
  - Assign collecting trip identifier

- **Data-mining: find collecting runs**
  - Targeted hidden Markov model: assign collecting / travelling state identifier. For each collecting trip detected in the previous step, use state transition analysis to detect collecting / travelling state for each day of the collecting trip, using the number of specimens collected per day as the observed state (see section 3.2.4).
  - Assign collecting state run identifier

4.3.4 Definition of baselines

The data-mining process described above results in a number of new groupings, which are used to analyse specimen datasets. In order to make an assessment of the utility of these new groupings, a corresponding baseline was defined for each, and similarly used to group and analyse the specimen data. The definitions of these baselines are shown in table 4.2

TABLE 4.2: Baseline definition for each data-mined entity

Entity	Data-mining process (each step builds on output of previous step)	Baseline definition (fields used to create grouping)
Collector	DBSCAN plus post-processing	Primary collector
Collecting trip	DBSCAN for each collector	Primary collector, year
Collecting state run	Hidden Markov Model for each collecting trip	Primary collector, day

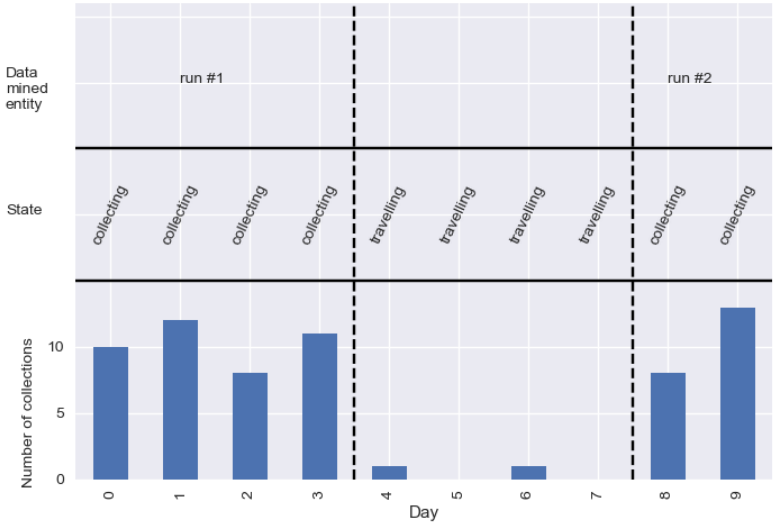


FIGURE 4.4: Run data-mining: example plot of the number of collections gathered per day from a single data-mined collector. This shows two runs of consecutive days with intensive collecting activity separated by a run of consecutive days of lessened collecting activity.

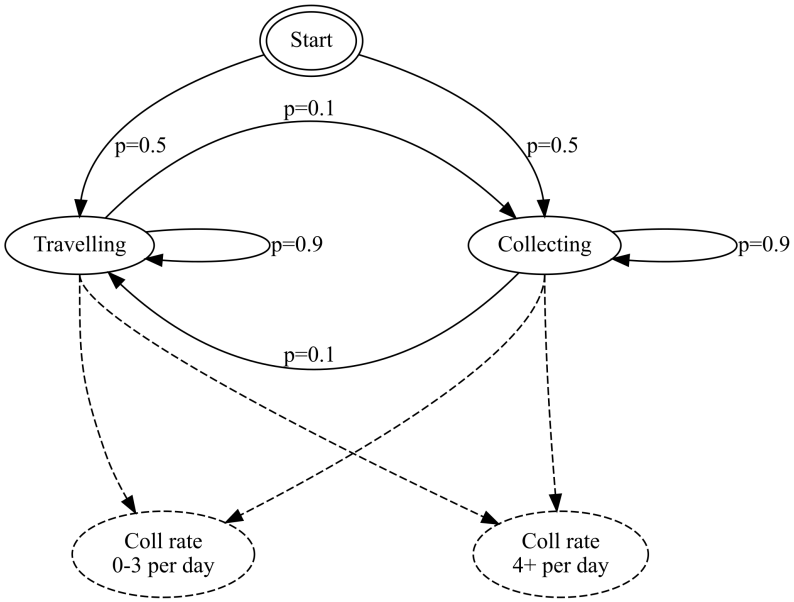


FIGURE 4.5: Run data-mining: hidden and observable states, and transition and emission probabilities for the hidden Markov model used to detect collecting run sequences

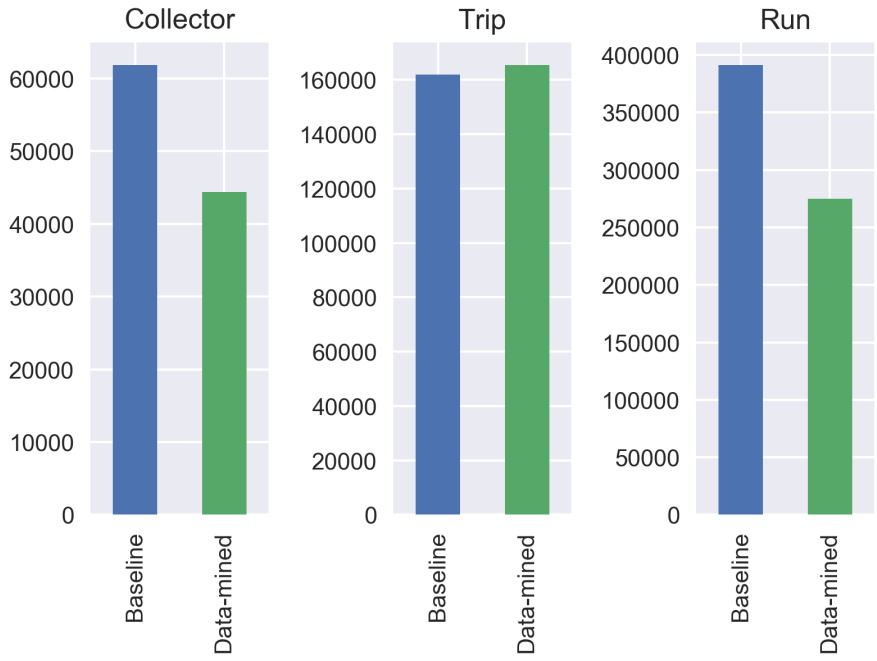


FIGURE 4.6: Data-mined aggregation counts compared with baseline

4.4 Results

4.4.1 Process results

An example output of the data-mining process is shown alongside the subset used to introduce the concept in figure 4.2(c) where recordnumber is plotted against eventdate to display elongated clusters hypothesised to represent the activity of individual collectors.

Number of data-mined entities compared with baselines

The data-mining process detected 44.34k collectors, 165.35k collecting trips and 274.8k collecting state runs. Comparisons of these numbers with baselines (61.8k baseline collectors, 161.89k baseline collecting trips and 391.03k baseline collecting state runs) are shown in fig. 4.6.

Showing extent of data variation

Figure 4.7 shows the data flow into data-mined entities, using a subset of data selected as those entities which include a last name of 'Hutchison' (as per the collector used in the data preparation process outline in figure 4.3).

Process participation

Eligibility criteria were established for participation in the data-mining process. A Sankey visualisation (figure 4.8) illustrates the participation in the

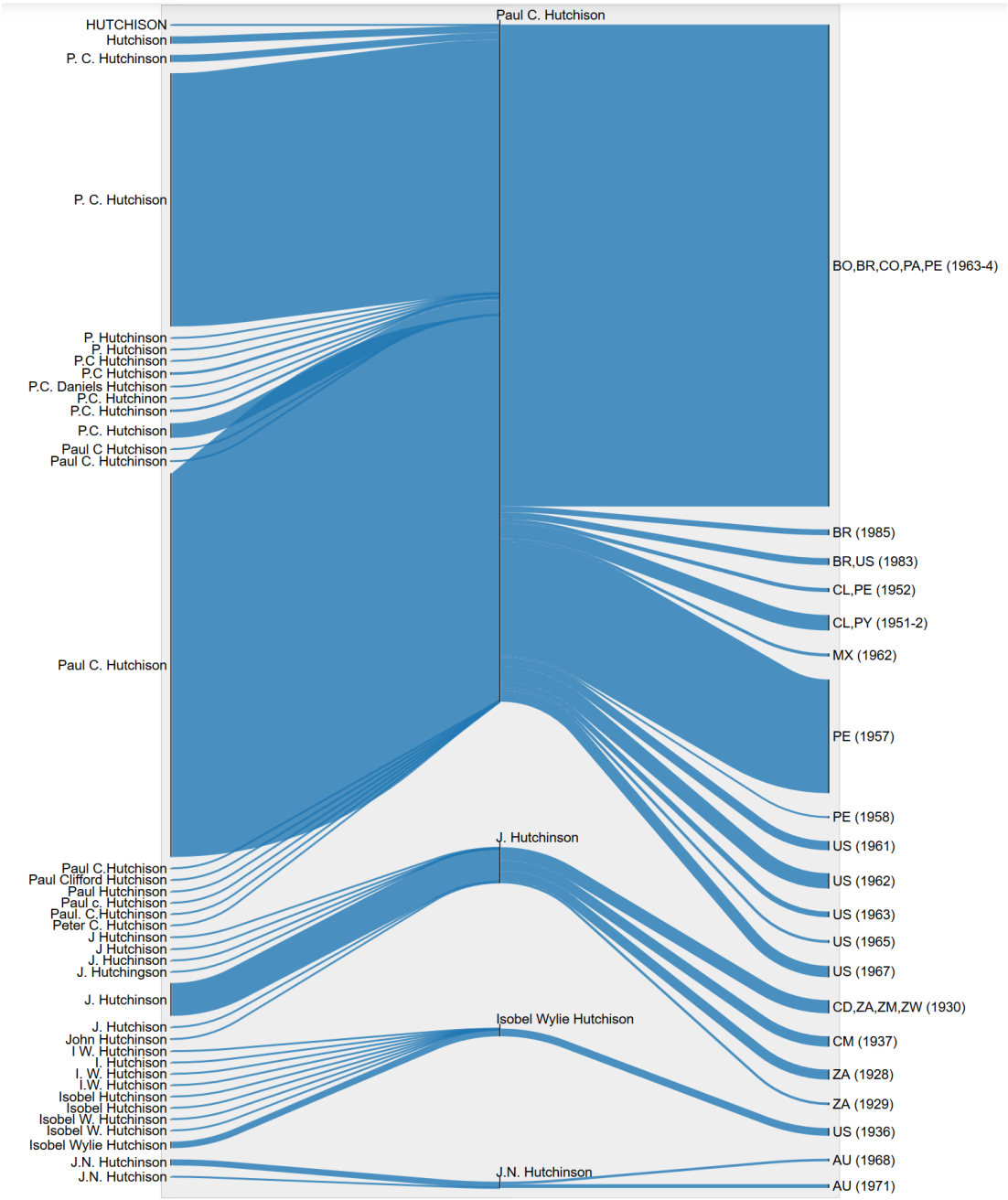


FIGURE 4.7: Sankey diagram to illustrate data flow into data-mined entities. The leftmost column shows the canonical recordedby value (which is also the baseline collector aggregation), the central column shows the data-mined collector entity, the rightmost column shows the data-mined trip (with countrycode and year range). This shows that the data-mining process detects distinct collector entities and that these have distinct temporal and spatial specialities - e.g. J.Hutchinson collecting in Africa in the 1920-30s and J.N.Hutchinson collecting in Australia in the 1960-70s. For both of these collectors "Hutchison" is a mis-spelling of their surname on at least one specimen record.

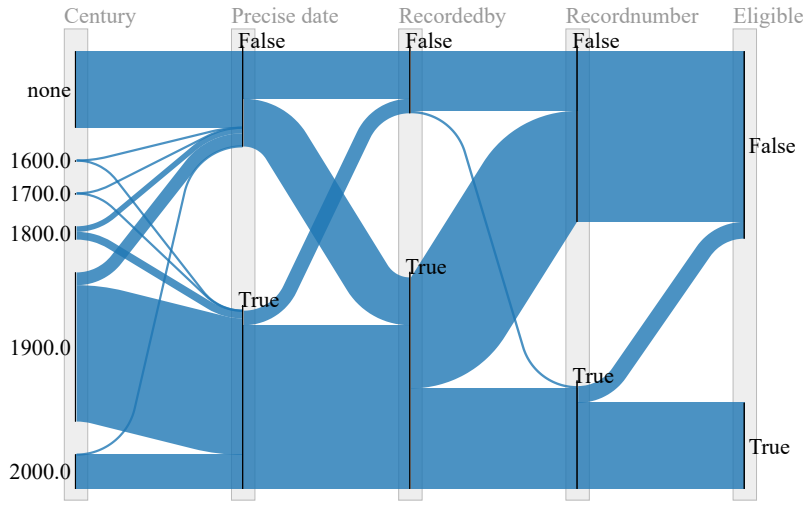


FIGURE 4.8: Sankey diagram to illustrate participation in the data-mining process. The preparation process defines an "eligibility" flag for each record, which is set to true if the record has a precise date, a collector name value in recordedby, and a numeric value in recordnumber. This visualisation shows the proportions of the dataset with these fields populated (flagged as true) or missing (flagged as false). The date of collection of the record is summarised to century (in the leftmost column). Reading left to right, subsequent columns show the proportions of the dataset that have the components of the eligibility flag available for use. The breakdown for the eligibility flag itself is shown in the rightmost column.

data-mining process and the characteristics of the ineligible portion of the dataset by showing the flow between different groupings (century, availability of precise date, recordedby and recordnumber and finally the eligibility status) across the complete dataset.

4.4.2 Trends analysis using data-mined entities

One objective of the data-mining process was to enable trends analysis on the specimen data. Using the data-mined aggregations (and baseline comparisons), the data were grouped and reshaped to conduct some simple trends analyses at collector, trip and collecting state run level. These trends analyses are summarised in the table 4.3.

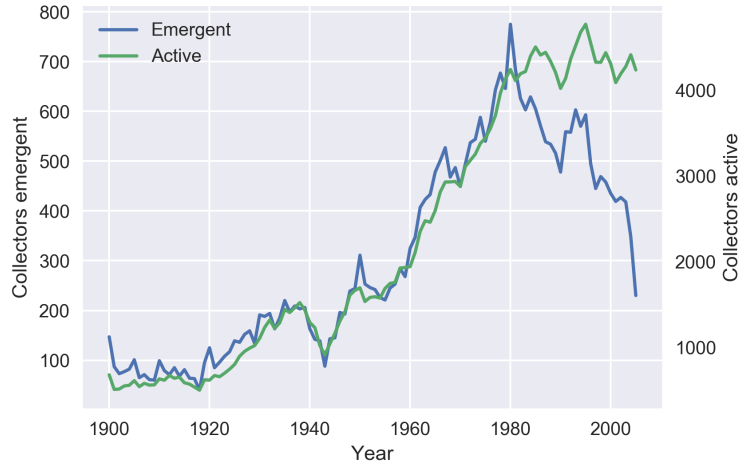


FIGURE 4.9: Numbers of collectors active and emergent (1900-2005)

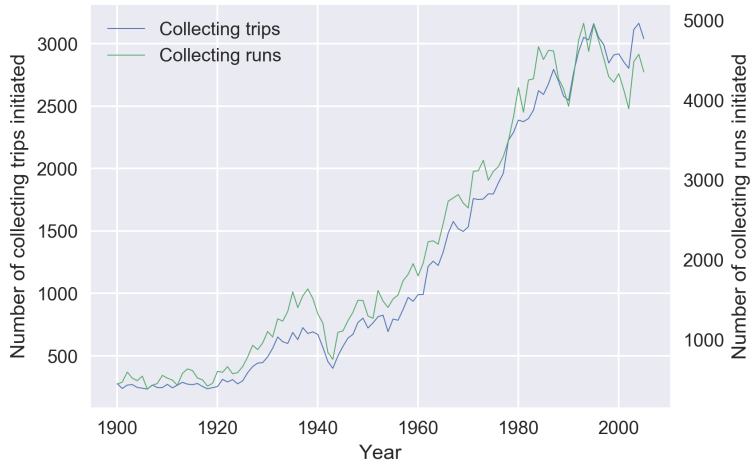


FIGURE 4.10: Numbers of collecting trips and collecting state runs per year

TABLE 4.3: Trends analyses conducted per data-mined entity

Role	Entity	Trends analyses
Object	Collecting event	
Creator	Collector	
Container (macro scale)	Career	Collectors active and emergent each year (figure 4.9)
Container (micro scale)	Collecting trip	Trips per year (figure 4.10)
		Active collecting days per year (figure 4.11)
		Mean duration of trips (figure 4.12)
	Collecting state run	Runs per year (figure 4.10)
		Mean number runs per trip (figure 4.13)
		Mean duration (figure 4.14)
		Prevalence of single run trips (figure 4.15)

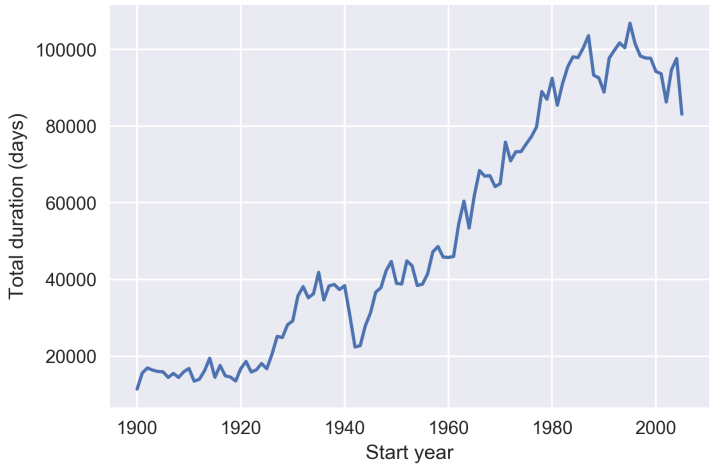


FIGURE 4.11: Collecting trip days per year

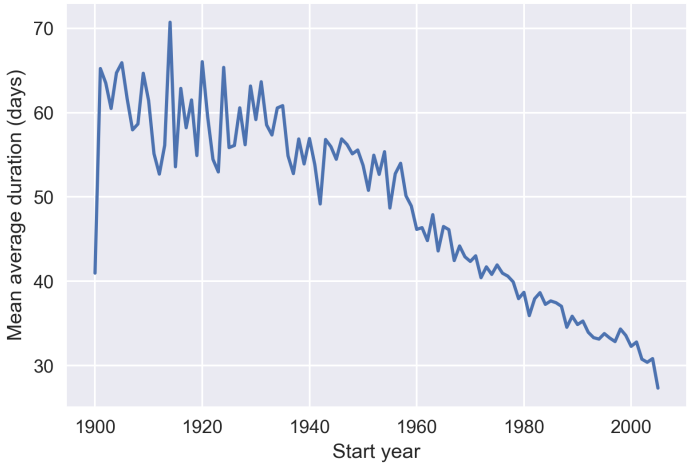


FIGURE 4.12: Collecting trip duration

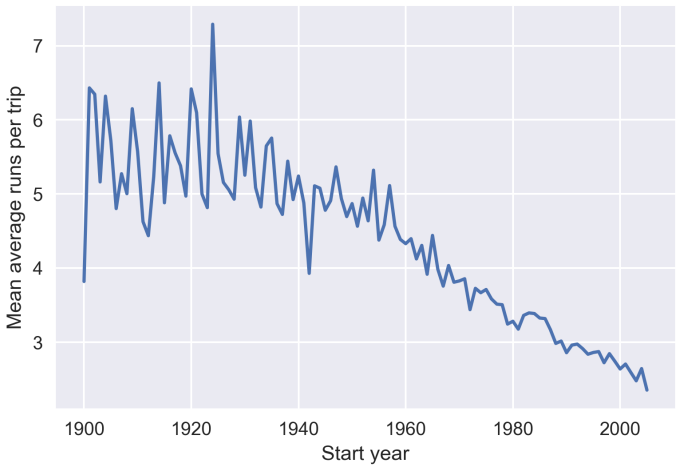


FIGURE 4.13: Collecting state runs per collecting trip

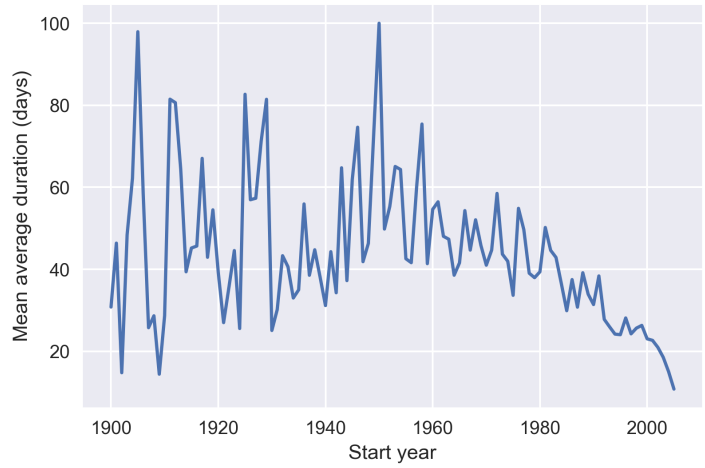


FIGURE 4.14: Collecting state run duration

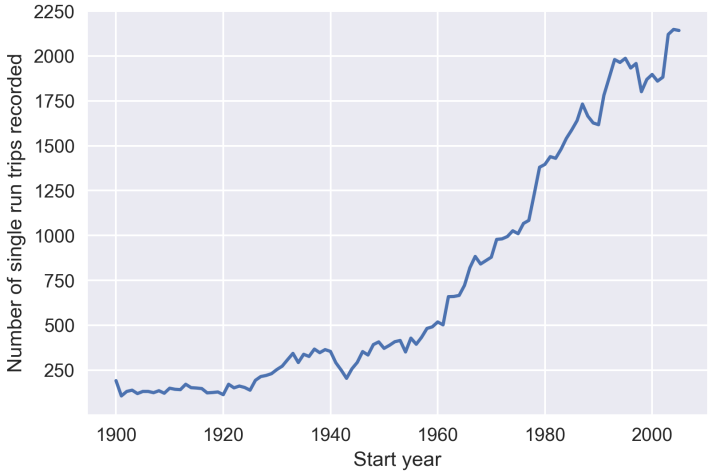


FIGURE 4.15: Prevalence of single state trips

4.5 Discussion

4.5.1 Process

Revision of the collector data-mining process

The data elements used as input into this data-mining process (encoded in DarwinCore terms as *recordedby*, *eventdate* and *recordnumber*) represent simple numeric and textual data elements routinely captured during specimen digitisation. The use of recordnumber in this data-mining process illustrates an opportunistic use of a data element that may previously have been considered of use only in a limited context (originating from a shorthand note to cross-reference physical material and notes or sketches). Best practices for field collection work are documented in handbooks (e.g. (Bridson 1998)), it may be worth expanding the recommendations regarding use of recordnumbers to highlight their potential utility in wider contexts like data-mining applications, and to ensure that future collections can participate in these kinds of analyses.

Analysis of the number of records eligible to participate in the data-mining process shows that availability of a numeric *recordnumber* filters out a lot of specimen records. Also, whilst recordnumbers are seen in some zoological collections, they do tend to be a more botanical field practice. The data mining process is based around clustering to detect traces which represent the activity of a particular collector, in this case as they move through numeric and temporal space. If the numeric component (recordnumber) was not available, could a similar data-mining process use a different data element from the specimen metadata? This would enable different records to be data-mined and could also allow the application of the data-mining process to non-botanical data, if the alternative data element were one also used in zoological collection practice. The required characteristics of an alternative data element are that it is numeric, and that it varies over time, with nearby values more likely to be recorded at a similar time. Many specimen records has been georeferenced - labelled with latitude and longitude coordinates following interpretation of the locality description (or were labelled at point of collection using hand held GPS technology). These numeric coordinates could be used in the data-mining process in place of the recordnumber - so that the data detection process is looking for traces of a collectors activity from their movement through physical space and time. An alternative data-mining process based on geographic coordinates, or a hybrid approach using recordnumber and geographic coordinates, depending on availability, would allow a greater amount of specimen data to participate.

A modified data-mining process employing traces through space and time would also allow the data-mining of non-primary members of the collecting team. The technique originally outlined here is only concerned with the primary member of the team (assumed to be the maintainer of the sequential recordnumber) - subsequent members of the team are discarded. The use of spatial features in preference to a recordnumber feature only available for the primary collector would allow a much fuller data-mining process and potentially lead to the development of a social network of relations between collectors based on co-participation in collecting teams. This would open up collecting team data to the kinds of analyses often directed at author team data, e.g. investigation of mentoring relationships and the "chaperone effect" (Sekara et al. 2018).

Revision of the collecting state detecting process

A Hidden Markov model was used to detect state transitions, representing the boundaries between days of intense collecting activity, and days of lesser collecting activity (likely to be used in travelling, preparation work, setting up camp etc). A fixed set of parameters was used for all input data (see figure 4.5), a potential further refinement of the process could be to use the forward-backward algorithm to establish different parameters for each collector data-set, this would enable a more precise modelling of different modes of collector behaviour.

4.5.2 Trends analysis

The results of the data-mining process show that a heterogeneous set of specimen collection data can be data-mined to establish entities which can be applied to the source data to facilitate trends analysis, similar to those conducted with a "clean" dataset of name publication data which is labelled with editorially managed entities.

As with the trends derived from publication events (figure 3.2), it is shown that more people are participating in specimen collection over time (figure 4.9), although the number of newly emergent collectors tails off. Similarly, the numbers of collecting trips and collecting runs increase over time (figure 4.10). However, duration of collecting trips decreases over time (figure 4.12), and the number of collecting state runs per trip also decreases over time (figure 4.13), with an increased prevalence of single run trips (figure 4.15).

4.6 Conclusions

The preliminaries chapter (chapter 3) showed that trends analyses were possible using name publication data, due to the investment in editorialised data management, which has resulted in a dataset of robust supporting entities. An example is the trends analysis on participation in name publication, which can be explored by reshaping the data to give greater prominence to the associated author entities. A similar editorial investment is not feasible at the scale of the global set of specimens, necessitating a different approach. The analysis presented in this chapter has shown that an automated data-mining approach can be used to establish allied entities in specimen data, representing the collector, collecting trip and collecting state run, and that these can be used to investigate participation trends.

The aim of this piece of research was to use data-mining on a heterogeneous specimen dataset in order to detect entities representing the *object*, *creator* and *container* as were available in a smaller scale dataset through editorial management (see comparison in table 4.1). The development of this data-mining process has refined our understanding of the "container" grouping - as this was detected as a *sequence of work* by exploiting an *agent managed sequence* (recordnumber) which ascends over time. It is possible that this approach could be generalised for application to similar problems in biodiversity or related domains - these were examined for other examples of data generated via a similar process. Many of the datasets generated in the digital age have recognised the need for shared persistent identifiers across distributed datasets (e.g. the use of Digital Object Identifiers [DOIs] in publishing) and by implementing these have sidestepped the need for this kind of analysis. The examples selected represent data generation via digitisation of historic information, that which pre-dates easily accessible shared identifiers. As seen in the previous chapter, scientific names for plants are referenced using micro-citations, page level bibliographic references, and there has been a manual effort to standardise the authorship for these to enable trends analysis. Page level microcitations can be seen as another representation of the object / creator / sequence of work data generation process: the object is a page-level microcitation, created by an author, within a bibliographic container (article or book) as a sequence of work. As page number is sequential and pages located in close proximity are likely to be authored by the same person, heterogeneous bibliographic datasets could also be candidates for a similar data-mining process using this technique.

4.7 Relevance to next chapter

Both of the components of the systematic process that we have investigated so far feature "agents", people performing a particular role: authors in the name publication analysis (chapter 3) and collectors in this specimen data-mining chapter. The next chapter will explore the overlap between these two agent datasets, to determine if these currently separate agent categories can be further abstracted to a generic scientist entity, which performs multiple roles (including collecting and authoring) throughout a career.

Chapter 5

Analysis of units of agent work

This research chapter defines a method to establish a generic scientist ("agent") with participation in multiple activity stages - collecting and name publication. This scientist agent is established by integrating the collector resultant from the data-mining process defined in chapter 4 with a dataset of editorially created authors (as utilised in chapter 3). Specimen collecting event and name publication event data are used to define features across these different activity stages. These features are used in a classification process to predict which agent-initiated units of work contribute towards new species discovery.

An early version of the research outlined in this chapter was published as a conference paper (the published version is available in appendix B.2), further developments are covered in a later conference abstract (listed in appendix C.6).

5.1 Visual context

Figure 5.1 provides a visual context for the scope of this chapter - showing that the agent definition process developed here uses data related to the **collecting trip**, **collecting event**, the **collections** data themselves, along with the data about the **agents** responsible for the collection of the specimens and their use in the recognition and formal publication of **scientific names**.

5.2 Introduction

The species description process can be characterised as composed of three stages:

1. **Collection** of specimens via fieldwork
2. **Determination** of specimens to form species delimitation hypotheses
3. **Publication** of new or revised species names in accordance with the nomenclatural code

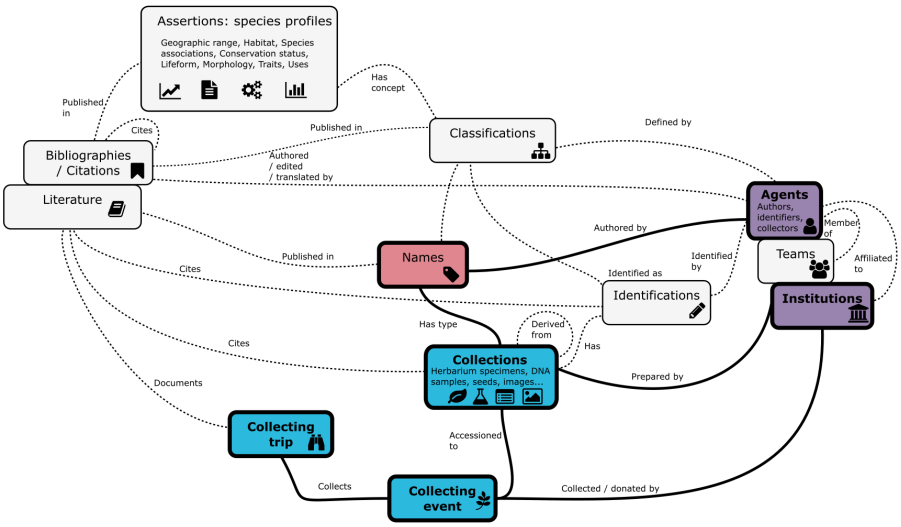


FIGURE 5.1: Visual context: agent analysis. This research chapter uses data relating to the **collecting trip** and **collecting event**, the **collections** data themselves, along with the data about the **agents** responsible for the collection of the specimens and their use in the recognition and formal publication of **scientific names**

Publication of scientific results is often used as a proxy for scientific output and used in career metrics. It has been suggested that in disciplines such as systematics, which are based on field work and specimen examination, the use of publication metrics to determine scientific output and career advancement may miss some crucial activity stages which generate research outputs (McDade et al. 2011). Attribution of scientists' effort in collecting and determination of specimens is an area of focus for a joint Research Data Alliance and Biodiversity Information Standards group who are working on data standards to represent attribution events (Thessen et al. 2016).

The aim of this piece of work is to develop a fuller profile of scientists' activity in the species description process, and to understand the contribution of particular units of work (collecting trips and complete collecting careers) to the species discovery process.

5.3 Methods and materials

5.3.1 Data

Two datasets are used in this analysis, the first results from the data-mining process described in the previous chapter (chapter 4) and is a dataset of collecting events labelled with collector identifiers, the second is an editorially-managed dataset of name publication events drawn from the

International Plant Names Index (*International Plant Names Index* n.d.), as used in the preliminary analysis chapter (chapter 3).

5.3.2 Integration of collector and author agents

The record linkage process is designed to make use of the connections inherent in the data to determine an area of focus defined by the taxonomic specialisation of the agent. This can narrow down the set of potential matches to a more select set of candidate matches. An **agent** acting as a **collector**, conducts a series of **collecting events**. Collecting events generate **specimens**, which are used in scientific research and labelled with scientific **names**. A component of a scientific **name** is the **author**, also an **agent**. (See figure 5.2). Rather than trying to directly establish links between the complete set of collector agents and author agents, the record linkage process only attempts to cross link collectors and authors who share participation in scientific names - either by a direct authorship relation (labelled *authored by*), or by an indirect collection event which gives rise to collection objects which are labelled with scientific names authored by a potential matching author agent (the multi-step route from the *collector* agent to the *name*, which forms the remainder of the diagram).

The candidate set of matches are filtered through lexical examination, using similarity methods developed in the previous chapter. Links proposed by this process are assessed against a gold standard dataset of links created by hand. The gold standard dataset was created by randomly selecting collector records resultant from the data-mining process in three "volume" categories (collectors responsible for the collection of 100 - 9,999 specimens, 10,000 - 19,9999 specimens and more than 20,000 specimens). 60 records in each volume category were manually linked to the author dataset, to give a gold standard dataset of 180 records. Metrics were calculated for each volume category, and for the dataset as a whole to determine the number of gold standard records that were linked by the automated record linkage process, and the proportion that were linked correctly (the positive predictive value).

5.3.3 Assessing balance of activities

After conducting record linkage on data-mined collector agents and editorially managed author agents, it is possible to examine the balance of activities between these two phases of work. A scatter plot of the total collection events against total publication events for each of the linked agents is given in figure 5.3. This shows that some agents appear to have a definite focus on a particular stage of the species description process.

The number of publication events and the number of collection events are used to derive a *relative difference* metric to represent the balance of activities.

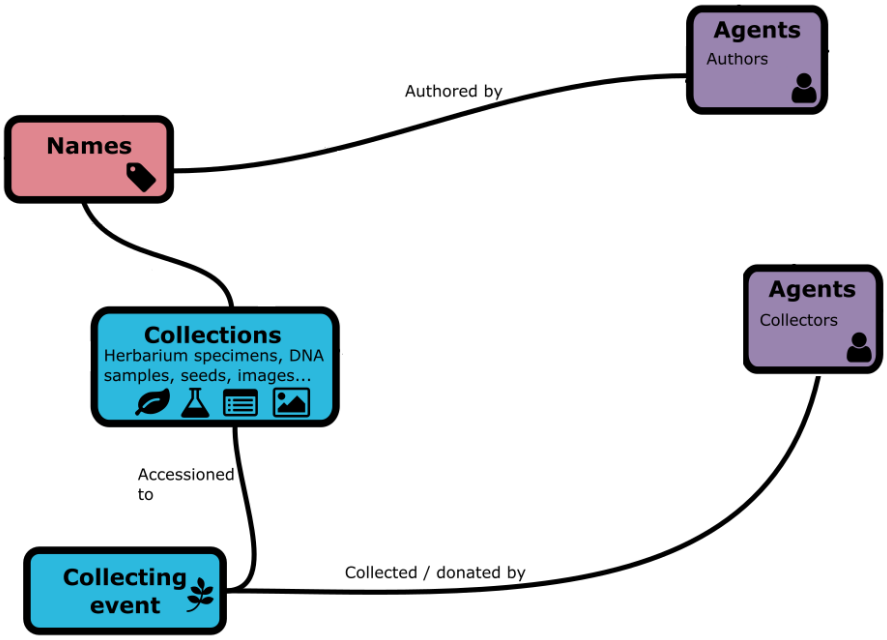


FIGURE 5.2: Record linkage principle: integration of author agents and collector agents, utilising connections inherent in the data. Instead of attempting to directly cross-link between author agents and collector agents, only those agents which share participation in scientific names are considered as potential matches

This is defined as a value varying between -1 (wholly collection event focussed) and +1 (wholly publication event focussed). In its simplest form, the metric assumes an equal balance of cost between the collecting and publication events. Publication is likely to be a higher cost activity than individual collecting events, so we can also define a weighted version of the metric to account for the difference in cost between the two activities:

$$abm = \frac{\sum_{p\_events} - \sum_{c\_events}}{max(\sum_{p\_events}, \sum_{c\_events})}$$

$$abm\_weighted = \frac{p\_cost \sum_{p\_events} - c\_cost \sum_{c\_events}}{max(p\_cost \sum_{p\_events}, c\_cost \sum_{c\_events})}$$

5.3.4 Feature definition

In addition to the activity balance metric, a number of features may be defined from the aggregations of specimen data enabled by the higher order data representations resulting from the data-mining process outlined in chapter 4. These are categorised as **temporal** (using minimum and maximum *eventdate*, *year* and *decade*), the **scale** of the aggregation (the *duration*, the *total specimens* included and the range of *recordnumbers* allocated). The **character** of the aggregation is defined from range of values in family, countrycode and continentcode. Where a single value in one of these fields accounts for more than 60% of the specimens in an aggregation, it is said to be *specialist*,

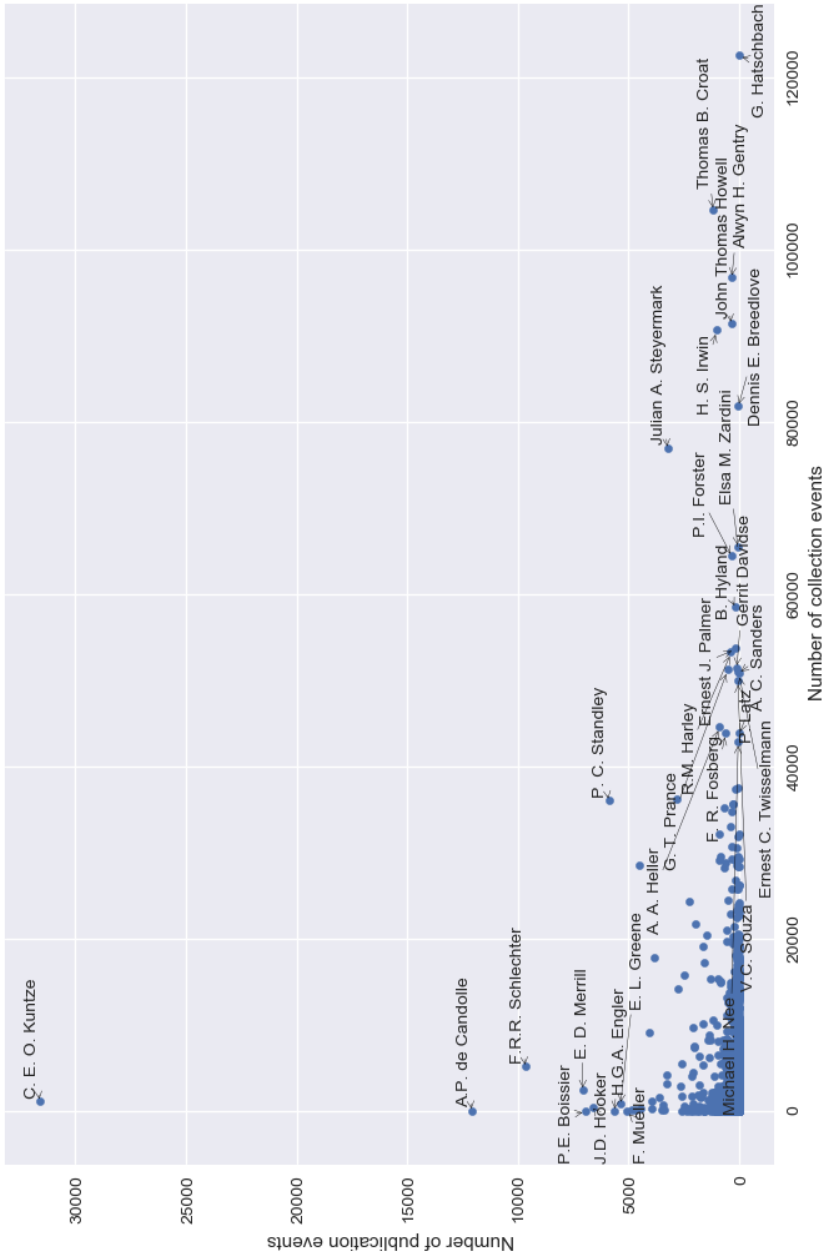


FIGURE 5.3: Activity balance scatter. For each of the linked agents, the total number of collection events is plotted against the total number of publication events, showing that some agents have a definite focus in one area of activity.

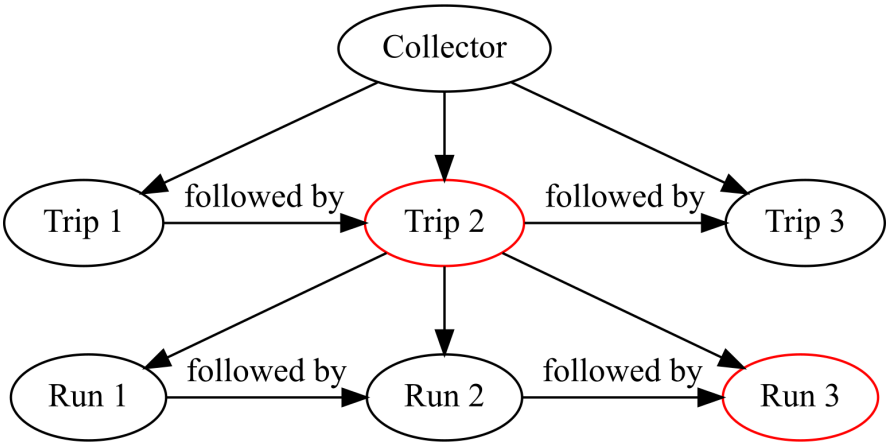


FIGURE 5.4: Explanation of creation of features from relationships between data-mined entities. The entities collector, trip and run can be thought of as forming a hierarchy: a collector conducts a number of trips, ordered in time, each trip is composed of one or more runs, also ordered in time. The features described as "child aggregation counts" count the aggregations at the next lowest level. The features described as "preceding sibling counts" count the aggregations at the same level and to the left of the aggregation in question. In the example the collector has a child aggregation count of 3 trips. The trip highlighted in red has a preceding sibling count of 1, and a child aggregation count of 3. The run highlighted in red has a preceding sibling count of 2.

conversely, if the most frequent value accounts for fewer than 30% of the specimens in an aggregation, it is said to be *generalist*. The range of *elevation* values is also used to assess the character of the aggregation. The **experience** of the collector at a point in time is assessed by creating features for the total number of previous specimens collected and the total number of previous collection trips made. The use of the inter-relationships between the data-mined aggregations to define new features is explained in figure 5.4.

Finally, a feature is defined that will later be used as the class variable in classifiers: this encodes the **species discovery value** of the aggregation, and is simply a Boolean flag indicating if the aggregation contains material that was later used as a type specimen, representing a contribution towards species discovery. Data files containing these features are constructed, these are used

as training data in the next step. These features are described in table 5.1.

TABLE 5.1: Definition of numeric and Boolean features on collector and collecting trip aggregations resulting from a data-mining process on a specimen dataset

Category	Field	Aggregation	Description	Career	Trip
Temporal	eventdate	min, max		✔	✔
	year	min, max, nunique		✔	✔
	decade	min, max, nunique, mode, mode_contrib, specialist, generalist		✔	
Scale	recordnumber	min, max, nunique		✔	✔
Character	elevation	min, max, median		✔	✔
	family	nunique, mode, mode_contrib, specialist, generalist		✔	✔
	countrycode	nunique, mode, mode_contrib, specialist, generalist		✔	✔
	continentcode	nunique, mode, mode_contrib, specialist, generalist		✔	✔
Rate	correlation_score			✔	✔
	slope			✔	✔
	14D_avg	career, active, ratio		✔	✔
	eventdate_week	nunique, career_diff		✔	
	weeks_active	percentage, per_year_median		✔	
	eventdate_month	nunique, career_diff		✔	
	months_active	percentage, per_year_median		✔	
	trips_per_year	min, max, median		✔	
Experience	trip_id	nunique	Child aggregation count	✔	
	trip_id	num_previous_trips	Preceding sibling count		✔
	run_id	nunique	Child aggregation count	✔	✔
	gbifid_count		Count of specimens	✔	✔
	author_id	Boolean	Set if collector linked to author record	✔	✔
	metric_w50		Activity balance metric (weighted)	✔	✔
Species discovery value	types	Boolean	Class variable, set if type material included	✔	✔

5.3.5 Classification and evaluation

The feature-sets generated by the data-mining process are used to train classifiers to predict the species discovery value of the grouping. These are compared to a baseline aggregation, derived without the data-mining process, as used for comparison purposes in the previous chapter. Both the baseline and the data-mined datasets were down-sampled to balance the binary class variable, as the samples for the positive class were seen less often.

A random forest classifier was trained on the down-sampled data, using 10-fold stratified cross-validation. As the class variable is binary, ROC-AUC curves were used to assess classifier performance. Feature selection was also conducted to examine which of the features defined were the most indicative.

5.4 Results

5.4.1 Integration of collector and author agents

44.34k data-mined collector agents and 52.63k editorially managed author agents were input into the process and 4174 links were made to assert a generic scientist identity. A gold standard dataset was used to assess the record linkage process, the results of this evaluation are shown in table 5.2.

TABLE 5.2: Agent record linkage evaluation using gold standard dataset of manually created links. Results are shown for each collecting volume batch (separating collectors responsible for different numbers of specimen collections - i.e. high volume collectors are differentiated from low volume collectors), along with a total assessment.

Volume	Number_gold_standard_links	Number_linked	Positive_predictive_value
100-9999	60	32	1
10000-19999	60	51	1
20000+	60	55	1
Total	180	138	1

5.4.2 Activity balance metric

Histograms of the values of the raw and weighted activity balance metric are shown in figure 5.5.

5.4.3 Classification and feature selection results

The number of samples participating in the classification process, and number retained after down-sampling are shown in table 5.3

TABLE 5.3: Participation in unit-of-work classification. Numbers of samples participating in the classification process following size check (minimum number of specimens per aggregation: 25) and down-sampling to balance the binary class variable.

Aggregation	Samples read	Samples retained after null and size check	Class variable: True	Class variable: False	Resampled to
Collector, data-mined	43679	19318	6371	12947	12742
Collector, baseline	61795	26848	9819	17029	19638
Trip, data-mined	164965	60259	15144	45115	30288
Trip, baseline	161892	62869	19058	43811	38116

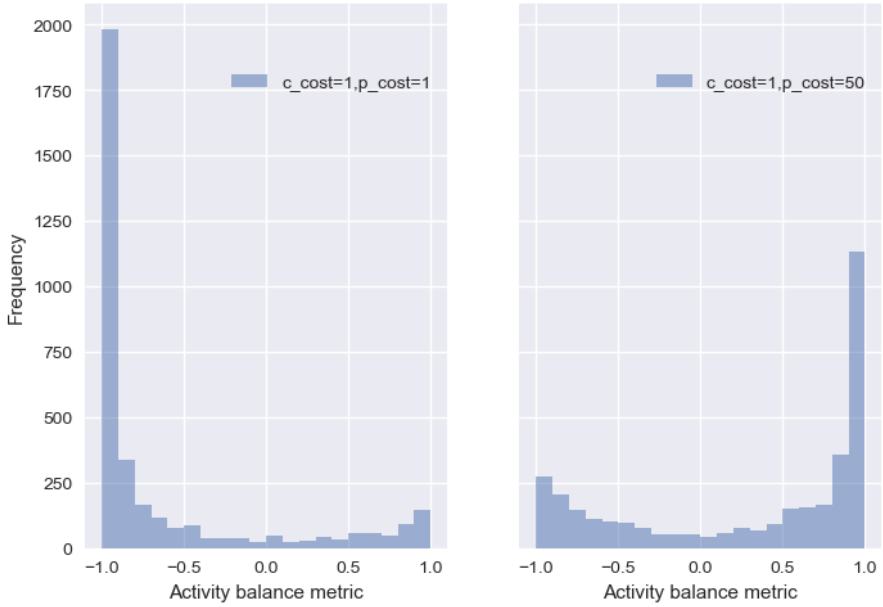


FIGURE 5.5: Activity balance metric histogram - raw (left), weighted (right)

The classification process was assessed by calculating the mean area under the receiver operator curve from the 10-fold cross-validated runs (see figure 5.6). Classification results from the collector aggregations show a slight performance increase from the trip aggregations. The datasets derived from data-mining were used to conduct feature selection using recursive feature elimination, scoring using the area under the receiver operator curve (ROC AUC) metric. The classifier results for feature sets of varying sizes are shown in figure 5.7, and the features ranked as most important in this process are listed in table 5.4.

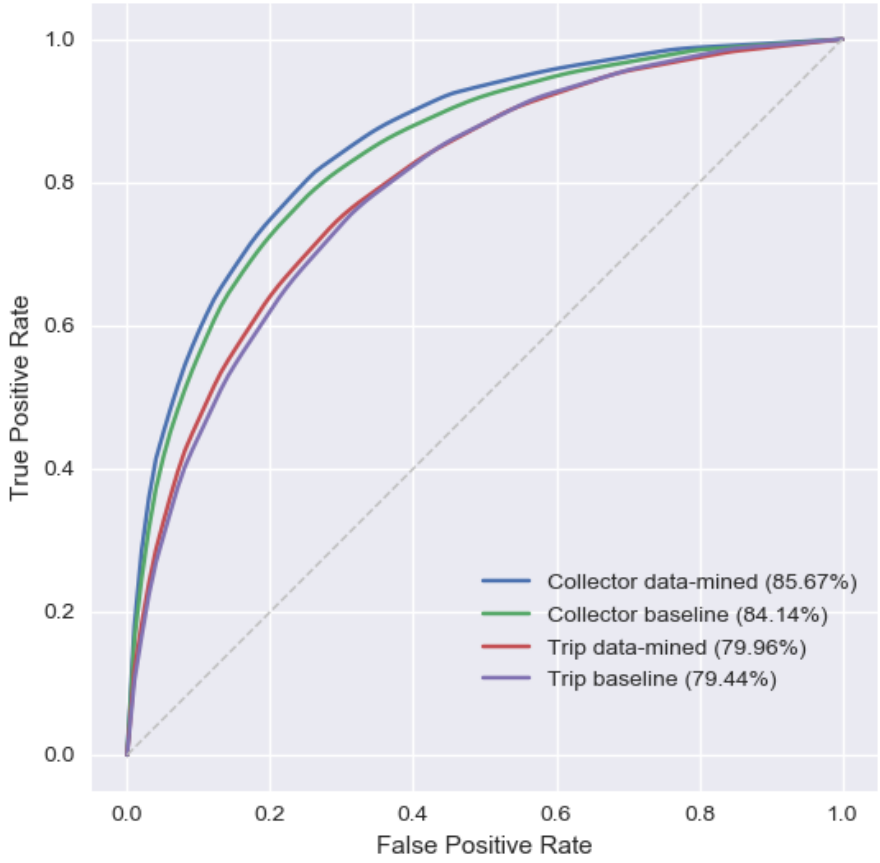


FIGURE 5.6: Assessment of random forest classifier performance for data-mined collector, baseline collector, data-mined trip and baseline trip using area under the receiver operator curve.

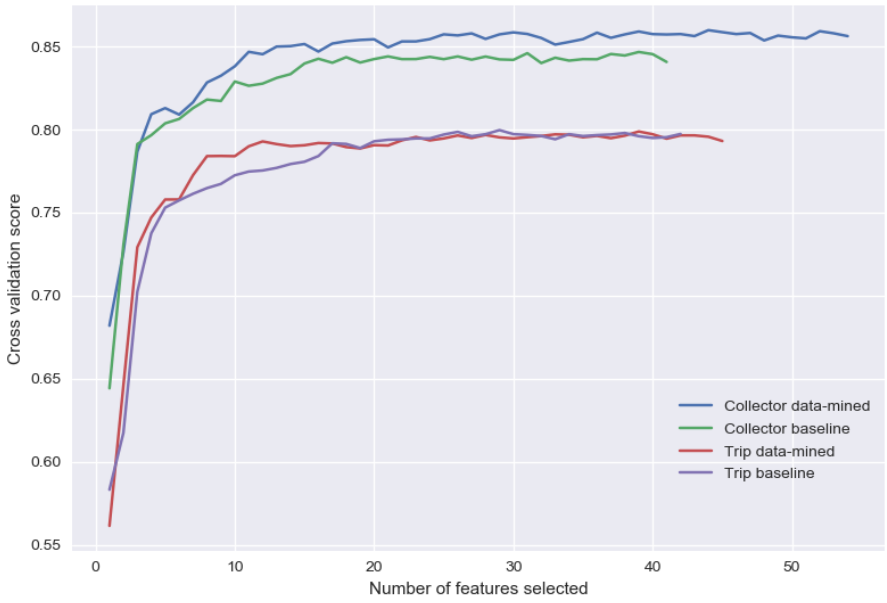


FIGURE 5.7: Classifier accuracy against number of features selected using recursive feature elimination for a random forest classifier on datasets of data-mined collector, baseline collector, data-mined trip and baseline trip

TABLE 5.4: Feature selection: top ranked features for each dataset, assessed using recursive feature elimination

Feature	Collector data-mined	Collector baseline	Trip data-mined	Trip baseline
year_min	✔	✔	✔	✔
year_max	✔	✔	✔	✔
year_nunique	✔	✔	✔	
decade_min	✔	✔	✔	✔
decade_max	✔	✔	✔	✔
recordnumber_min	✔	✔	✔	✔
recordnumber_max	✔	✔	✔	✔
specimens_collected	✔	✔	✔	✔
elevation_min	✔		✔	✔
elevation_max	✔	✔	✔	✔
elevation_median	✔	✔	✔	✔
family_nunique	✔	✔	✔	✔
countrycode_nunique	✔	✔	✔	✔
continentcode_nunique	✔	✔	✔	
gbifid_count	✔	✔	✔	✔
trip_id_global_nunique	✔			
run_id_global_nunique	✔		✔	
top_family_contrib	✔	✔	✔	✔
top_countrycode_contrib	✔	✔	✔	✔
top_continentcode_contrib	✔	✔	✔	✔
top_decade	✔	✔	✔	✔
top_decade_contrib	✔	✔	✔	
correlation_score	✔	✔	✔	✔
slope	✔	✔	✔	✔
14D_avg_career	✔	✔	✔	✔
14D_avg_active	✔	✔	✔	✔
14D_avg_ratio	✔	✔	✔	✔
eventdate_month_nunique	✔			
eventdate_month_career_diff	✔			
months_active_pc	✔			
months_active_per_year_median	✔			
eventdate_week_nunique	✔			
eventdate_week_career_diff	✔			
weeks_active_pc	✔			
weeks_active_per_year_median	✔			
trips_per_year_max	✔			
trips_per_year_median	✔			
metric_w50	✔		✔	
duration	✔	✔	✔	✔
recordnumber_diff	✔	✔	✔	✔
nomenclaturalist	✔	✔	✔	✔
start_year	✔	✔	✔	✔
num_previous_collections	✔	✔	✔	✔
decade_nunique		✔	✔	
trip_id_baseline_nunique		✔		
run_id_baseline_nunique		✔		✔
family_specialist		✔	✔	
family_generalist		✔	✔	
countrycode_specialist		✔	✔	
countrycode_generalist		✔		
continentcode_specialist		✔		
decade_specialist		✔		
num_previous_trips			✔	✔

5.5 Discussion

5.5.1 Record linkage

The record linkage process was evaluated with reference to a manually created set of gold standard links, which were randomly selected and batched to ensure coverage of a range of collectors with varying collecting volumes. The batch containing the highest volume collectors achieved the highest number of linkage results - as the linkage process utilises the connections in inherent in the data (see figure 5.2), higher volume collectors have more connections to use as a source of potential links. Future work could examine the potential for similar record linkage processes using different agent interest calculations - e.g. geographic focus - which are available in both the publication and collection datasets.

5.5.2 Activity balance and characterisation

An initial measure of activity balance between two phases of work has been proposed, with the ability to weight the metric to account for different levels of effort between different phases. It is possible that the relative effort weighting is not uniform across agents working in different areas of the world, as we have recognised that access to scientific resources is particular challenge for systematists, as seen in the global distribution of specimens in figure 2.1.

5.5.3 Classification

The construction of feature sets from the data-mined aggregations and the use of these to evaluate units of agent initiated work show that the aggregations can be used to assess contributions towards species discovery. The results from the feature selection process can also have a practical use in the prioritisation of data curation tasks - a specimen metadata record is composed of many data elements, but if a classification process can be defined to aid understanding of species description and feature selection identifies the most useful features for this task, the data elements from which these features are created can be prioritised for digitisation and data cleaning. When comparing the whole career collector aggregations against the more focussed collecting trips, it is shown that collector appears to perform better than trip; it is likely that trip is more susceptible to incomplete data than the whole-career aggregation. A more focussed study with a dataset which can be assessed for completeness would help to test this.

5.5.4 Generalisation of approach

The techniques presented here are based around the integration of editorially managed and data-mined datasets. These would potentially generalise to the integration of agents derived from separate data-mining activities, as would be found in domains like zoology, which lack a editorially managed nomenclatural resource covering species name publication events.

5.6 Conclusions and relevance to next chapter

This chapter has presented techniques which are used to integrate editorially managed datasets of author entities and collector entities derived from large scale heterogeneous aggregated datasets. The results of this integration process have been used to assess specialisation in distinct process stages through the definition of an activity balance metric, and to predict the contribution of particular units of work towards species discovery using classification techniques. The research presented so far has shown that specimens are the core of systematics research and that these are generated and used by scientists in the process of species discovery. The next chapter will propose techniques to reconcile the distributed products of a collecting event, to link specimens across institutional boundaries. The aim of this next piece of work is to maximise the impact of the expert annotations that scientists apply to specimens as they are used in research, and further develop a fuller picture of scientists' research activity with specimens.

Chapter 6

Reconciling specimens across institutional boundaries to enable metadata propagation

This third research chapter builds on the agent data-mining from chapter 4 to detect specimens generated from a common agent-initiated collection event. These specimens are often held and managed separately in distributed repositories, meaning that digital metadata created from the specimen (such as the transcription of the label, as shown in the sample specimen in figure 2.4) may vary across the specimen group. Reconciling these specimen groups enables the calculation of the number of metadata updates that may be propagated between institutions. The grouped specimens are also used to create a network representation of the relationships between institutions, which is used for community analysis.

The research outlined in this chapter was published as a conference paper (the published version is available in appendix B.3).

6.1 Visual context

Figure 6.1 provides a visual context for the scope of this chapter - showing that the specimen reconciliation process developed here uses data related to the **collecting event**, the **agent** responsible for the collection of the **specimen** and the **institutions** in which the specimens are lodged for long term storage and consultation, to aid the process of species discovery and the publication of new scientific **names**.

6.2 Introduction

Botanical specimens are core research objects in the science of taxonomy (the naming of biological organisms), stored for long term consultation in institutional repositories and referenced in academic works. Worldwide there are 4073 herbaria (botanical specimen repositories), containing 390.48M

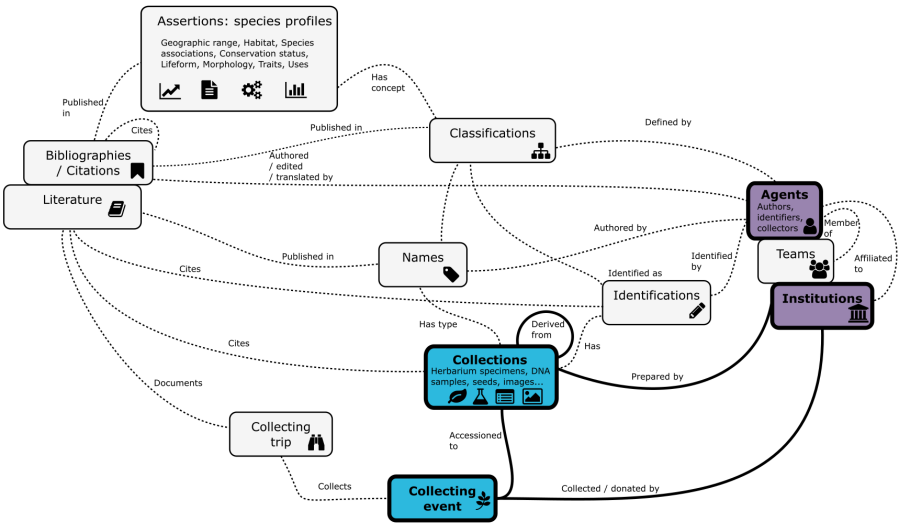


FIGURE 6.1: Visual context: specimen reconciliation

specimens - representing collections gathered over hundreds of years (Thiers continuously updated). Due to their physical characteristics (flattened, dried plant material is typically mounted on a large sheet of paper, stored inside a paper folder) and their management as a long term, consultable record, specimens act as vehicles for the communication of results and theories, as researchers annotate the paper sheet underlying the specimen. Annotations placed on specimen sheets are public and available for use by other researchers, this public yet potentially unpublished status is discussed in (Conn 2003).

Taxonomic researchers populate institutional repositories by conducting field-based collection events which generate multiple specimens. Recommended botanical practice is for a single collection event to generate five to six specimens, which will be deliberately distributed to separate institutional repositories. Physical distribution of specimens has three main goals: to *maximise access* - researchers working on their local flora should be able to consult the relevant specimens in their national herbarium, to provide *resilient storage* - duplicate specimens insure against disastrous loss of a single repository, and to ensure *efficient use of storage space* within repositories (Bridson 1998). Duplicate specimens are also used in genetic analyses: if the samples were collected from separate individuals, the duplicate set can be used to assess genetic diversity across the sampled population. Scientific theories regarding the recognition of species and their interrelationships are developed by researchers as they work with the specimens, which are traditionally accessed either by loan or by visits to institutions; more recently specimen digitisation initiatives have enabled online access to specimen metadata records and high quality images, this simplifies search and retrieval

of specimens and associated metadata, and allows some level of specimen examination to be conducted remotely. Independent creation and management of metadata for specimen duplicates can be inefficient (metadata creation is repeated unnecessarily), and inadvertently misleading (metadata diverges between different members of a specimen duplicate group).

One particular class of research annotation is the application of a scientific name to the specimen: this may be an existing name, or the researcher may recognise that the specimen represents a new species. Species description in plants is ongoing with circa two thousand new plant species described each year (*International Plant Names Index* n.d.). When a new species is described, one specimen is chosen as a physical representation of the otherwise abstract scientific name. Specimens which formally represent a scientific name are called *type specimens*; the selection of these is called *type citation*. When a specimen is cited as a type, all peers ("*duplicates*") which are generated from the same collection event - but which may be stored and managed remotely, in separate repositories - are also considered to have type status. New scientific names are created via a formal publication process governed by the International Code of Nomenclature for algae, fungi and plants (McNeil et al. 2012). The majority of new species are discovered from historic specimens already lodged within specimen repositories, being formally described years after collection (Bebber et al. 2010). The use of duplicate specimens as vehicles for the communication of results is illustrated by the historic use of "*exsiccatae*". These were uniform specimen sets with information displayed on printed labels, and were distributed to multiple herbaria as a kind of combined specimen and publication set. Until 1953 these were considered a valid publication mechanism for new scientific names (Triebel et al. 2011) (McNeil et al. 2012).

Taxonomists consider type specimens to be the most valuable specimens in a collection, and management reporting often includes both the total number of specimens held and the number of type specimens. The first major international digitisation effort in botany (JSTOR Global Plants Initiative) focussed on the digitisation of type specimens across more than 300 institutions in over 70 countries (ITHAKA 2015). In addition to reporting on the total numbers of specimens and types housed in an institutional repository (Bras et al. 2017), managers are also interested in the numbers of new type citations published each year as a metric of on-going research use of their specimens (Friis 2012). Some natural history institutions have experimented with bibliometrics to quantify use of their specimens in a publication context (Winker and Withrow 2013).

In addition to their core use in the science of taxonomy, specimens provide physical "what, where, when" evidence and are used for a wide

range of scientific applications such as species distribution modelling (Chapman 2005). Specimen exchange networks have also been used for historical social network analysis (Groom et al. 2014). These applications are generally dependent on aggregations of specimen metadata mapped to a common data standard and sourced from many different institutional repositories.

Problem statement Despite the widespread recognition that botanical specimens form a global collection, there is currently no comprehensive cross-institutional metadata flow between specimens generated from a common field collection event. Despite advances in the mobilisation and standardized representation of specimen metadata across the different specimen repositories, duplicate specimens have so far gone undetected, with metadata records for duplicates appearing unlinked in aggregated datasets. The main data element needed to assess specimens as potentially arising from a shared collection event - collector name, along with the collector's recordnumber and eventdate - are not formally managed. These missing links mean that valuable research annotations and type citations are not easily shared between institutions, and this impacts all downstream users of specimen data: taxonomic researchers working with individual specimens are unable to benefit from knowledge added elsewhere, leading to misinterpretation due to inaccurate and/or out of date naming, and users working with large aggregations of specimen data can find that specimen number estimates are overstated, as their datasets contain hidden duplicates.

This analysis builds on the data-mining process presented in chapter 4, which identified collectors from specimen data aggregated from multiple different institutions. Establishing a shared identity for collectors across institutional boundaries simplifies the reconciliation of the collecting events and the identification of the distributed sets of specimens generated - these can be detected and linked by simple grouping processes.

In contrast to existing work on annotation propagation - which has focussed on potential changes in working practices and tools and techniques to enable and incentivize this (Suhrbier et al. 2017) (Macklin et al. 2006) - this work applies these techniques to a dataset of existing digitally available specimen data in order to calculate the numbers of existing metadata elements and annotations which may be propagated between separate institutional collections.

The remainder of this chapter is structured as follows: a background section further introduces the problem domain with an explanation of the specimen life cycle and the kinds of annotations applied at each stage, and worked examples of distributed specimen sets whose members are independently managed at different institutions. Materials and methods describes the use of collector data-mining results to define a grouping process

to detect specimen duplicates in a dataset of specimen data from the Global Biodiversity Information Facility. Criteria for the identification and assessment of duplicate sets are proposed. The resulting specimen duplicate analysis is used to answer the following questions:

1. How many distributed, independently managed specimens can be reconciled across separate institutional repositories and linked as generated products of a common collection event?
2. How many metadata elements and research annotations can be propagated between institutional specimen repositories?
3. Can specimen duplicate linkages be used to infer network relationships between institutional repositories, which institutions are most frequently linked and do sub-communities or cliques exist in the inferred network?

Results are presented and ideas for expansion and future work are proposed.

6.3 Background

This section outlines the stages in the specimen life cycle, and indicates relevant projects at each stage.

Collection and storage: these activities represent standard practice across the specimen repositories

- **Collection**: material is gathered from the field and details of the collection locality (associated species, geology, habitat etc) are recorded in the collectors field notebook. The collectors recordnumber provides the cross-reference between the data recorded in the field notebook and the physical material collected, this is usually a sequential number managed individually by the collector.
- **Accessioning**: material is received by a specimen repository and prepared for long term storage, including mounting on a sheet of paper (for dried specimens).

Digitisation: due to the number of specimens held in the global collection, digitisation is incomplete, and is progressing through a variety of cross-cutting institutional, regional, international and thematic projects. The JSTOR Global Plants Initiative selected a particular class of specimens for digitisation (type specimens) across 300 institutions (ITHAKA 2015), other projects have been set up to digitise all specimens gathered from a particular country to enable data repatriation, as in the Brazilian REFLORA programme (REFLORA 2017) and to digitise specimens held within a particular country

as in the US National Science Foundation funded Advancing Digitisation of Biocollections programme (*Advancing Digitization of Biodiversity Collections | NSF - National Science Foundation* 2018). These latter projects show a trend of government funding for digitisation, recognising that these are part of the national scientific infrastructure (Bras et al. 2017) (L. M. Page et al. 2015).

- **Databasing**: details of the specimen (metadata) are added to an institutional data repository.
- **Aggregation**: databased records can be mapped to a data standard (e.g. Darwin Core (Wieczorek et al. 2012) and shared with aggregation projects. The Global Biodiversity Information Facility is an intergovernmental organisation that aggregates specimen-derived species occurrence records (alongside records from observations) to facilitate scientific research, iDigBio is a US based aggregator which focusses only on specimen derived data.
- **Georeferencing**: the metadata record in the institutional repository can have latitude and longitude added (this may be a costly step for historic records where the original collection locality is only a textual description of the place). Economies of scale are possible if records can be ordered so that similar places are georeferenced together (Hill et al. 2009) (Garcia-Milagros and Funk 2010).
- **Imaging**: the specimen is imaged and a reference to the image is added to the metadata. If the specimen metadata is shared with an aggregator the digital image may also be mobilised.

Depending on their range of holdings, some institutions are involved in multiple digitisation projects, others not at all. With technical advances in digitisation and the setup of high-throughput imaging facilities, some of these steps may be performed out of sequence - i.e. if the digitisation project is of a sufficient scale, it may be cost effective to rapidly image the specimens first and perform the metadata capture later, from a high quality digital image (van Oever and Gofferjé 2012) (Heerlien et al. 2015) (Sweeney et al. 2018).

Use as a research object: these steps outline the use of the specimen as a taxonomic research object. The use of specimens as a data source for computational applications such as species modelling is covered in the digitisation steps above, digitisation steps also facilitate discovery and access of specimens for taxonomic research. Annotation mobilisation work has focussed on tooling for the collection and propagation of newly generated annotations, including the projects AnnoSys (Suhrbier et al. 2017) and Filtered Push (Macklin et al. 2006) There has also been an effort to standardise the citation of specimens so that different repositories use a common HTTP

URI based naming convention by which their digital metadata records can be accessed (Güntsch et al. 2017). By convention, the citation of specimen records irrespective of digitisation status is made by stating the collector name, number and date, along with the herbarium code (Thiers 2018) in which the physical specimen may be found. These kinds of references can be found throughout the botanical literature, and examples are shown in the worked examples in the next section.

- **Determination**: the specimen is labelled with a scientific name, the date and the name of the researcher who made the determination are also added.
- **Citation**: the specimen is cited in a published academic work (e.g. to evidence the presence of a species in a geographic region).
- **Type citation**: the specimen is referenced as a type specimen in a published academic work to create a new species name.

The long term creation of a global network of specimen repositories, the more recent efforts to enable virtual access to specimens and their metadata, and the practice of sharing research annotations all fit well with the FAIR principles for scientific data management (Wilkinson et al. 2016): ensuring that the metadata and specimens on which scientific analyses are based are Findable, Accessible, Interoperable and Retrievable.

6.3.1 Worked examples

These examples are intended to illustrate the problem statement:

- specimen duplicates are *widely present* in distributed specimen repositories
- specimen duplicates are *unidentified* in data aggregations built by combining specimen datasets
- specimen metadata attached to derived specimens generated from a single source can *diverge* due to separate and independent data curation practices

Two examples have been selected, representing the two extremes of species description citing botanical specimens: species discovery in-field formalised by rapid publication just one year after collection, and species discovery in-repository with formalised description decades after field collection. A considerable proportion of new species are described from material already collected and stored in specimen repositories (Bebber et al.

2010). The second example shows a species description occurring 46 years after the field collection of the plant material on which is it based.

For each example a dataset of potential specimens is assembled, which is constructed as the superset of the specimens referenced in the literature (which may or may not be digitised) and the relevant specimen records found in digital form in a data aggregator. The metadata attached to the specimens is examined, showing where this has diverged due to independent management in separate institutional collections. These are shown in tables 6.1 and 6.2.

Rapid publication of species discovered in-field

This first example illustrates the specimens associated with the rapid publication of a species discovered via field work.

The publication data (displayed below) shows that there are at least 9 specimen duplicates, stored in different institutional repositories, indicated by the capitalised alphabetic herbarium codes (WTU, BH etc (Thiers 2018). The exclamation mark (!) after a code is a convention to indicate that the author has actually seen the specimen. In this case the author is also the collector of the specimen, so all are listed as having been seen.

> **Sedum citrinum** Zika, *sp. nov.* **Type**: UNITED STATES. California: Del Norte County, ridge 1.4 air km north of South Red Mountain, 1050 m, 9 June 2013, *P. F. Zika 26185* (holotype: WTU!; isotypes: BH!, CAS!, GH!, MO!, OSC!, RSA!, UC!, US!). (Zika 2014)

TABLE 6.1: Distributed curation of specimens arising from a common collection event, worked example (Zika 26185)

recordedBy	scientificName	held in	cited	digitised	type	georef'd	imaged
P. F. Zika	Sedum citrinum Zika	BH	✔	-	-	-	-
Zika, Peter F.	Sedum citrinum Zika	CAS	✔	✔	✔	✔	-
Peter F. Zika	Sedum citrinum Zika	CAS-BOT-BC	-	✔	-	-	-
P. F. Zika	Sedum citrinum Zika	CHSC	-	✔	-	✔	-
P. F. Zika	Sedum citrinum Zika	GH	✔	-	-	-	-
Zika, P.F.	Sedum citrinum Zika	K	-	✔	✔	-	✔
P. F. Zika	Sedum citrinum Zika	MO	✔	-	-	-	-
P. F. Zika	Sedum citrinum Zika	NY	-	✔	✔	✔	✔
P. F. Zika	Sedum citrinum Zika	OSC	✔	-	-	-	-
Peter F. Zika	Sedum citrinum Zika	RSA	✔	✔	-	✔	-
Peter F. Zika	Sedum citrinum Zika	UC	✔	✔	-	✔	-
P. F. Zika	Sedum citrinum Zika	US	✔	✔	✔	-	✔
P. F. Zika	Sedum citrinum Zika	WTU	✔	-	-	-	-

There are 8 digitally available records for this set of specimens, drawn from 8 separate institutional specimen repositories. (See table 6.1, table data source: gbif.org) These are independently managed and not interlinked. Despite being generated from the same collection event, the specimen

metadata show variation due to isolated management in separate repositories: 5 of the 8 are georeferenced, 4 of the 8 specify a type status and 3 of the 8 have an associated image. We can therefore calculate that the group contains propagable annotations for georeferences, typestatus and image (i.e. that for each annotation class, the group contains records with and without the annotation set, meaning that the annotation could be propagated from the specimens with the annotation to their peers without it). Of the digitised specimens in the group: 3 could receive a georeference, 4 could receive a type status annotation and 5 could be linked to an associated image. The identification of a specimen group could also make the initial creation of the specimen records for the currently undigitised members more efficient, by using existing data as a starting point rather than independently re-creating it.

Species discovery in-repository

This (second) example illustrates the specimens associated with the publication of a species discovered via work with existing specimens stored in institutional repositories.

The publication data (displayed below) shows that there are at least 6 specimen duplicates, stored in 5 different institutional repositories. The author has supplied a numeric identifier for some of the specimens (shown in square brackets), to help the reader locate the relevant records in the specimen repository and / or its associated metadata catalogue(s).

> **Solanum sanchez-vegae** S.Knapp, *sp. nov.* urn:lsid:ipni.org:names:77103635-1 **Type**: Peru. Amazonas: Prov. Chachapoyas, W side of Cerros Calla-Calla, 45 km above Balsas, mid-way on road to Leimebamba, 3100 m, 19 Jun 1964, *P.C. Hutchison & J.K. Wright 5738* (holotype, USM; isotypes, F [F-163831], K [K000545365], P [P00549320], US [US-246605], USM). (Knapp 2010)

TABLE 6.2: Distributed curation of specimens arising from a common collection event, worked example (Hutchison 5738)

recordedBy	scientificName	held in	cited	digitised	type	georef'd	imaged
P. C. Hutchison & J. K. Wright	Solanum sanchez-vegae S.Knapp	F	✔	✔	✔	-	✔
P. C. Hutchison & J. K. Wright	Solanum aligerum Schltdl.	F	-	✔	-	-	-
Hutchison, P.C.	Solanum sanchez-vegae S.Knapp	K	✔	✔	✔	✔	✔
Paul C. Hutchison | J. Kenneth Wright	Solanum cutervanum Zahlbr.	MO	-	✔	-	-	-
P. C. Hutchison	Solanum sanchez-vegae S.Knapp	NY	-	✔	✔	✔	✔
P. C. Hutchison	Solanum sanchez-vegae S.Knapp	NY	-	✔	✔	✔	✔
P.C. Hutchison & J.K. Wright	Solanum sanchez-vegae S.Knapp	P	✔	-	-	-	-
P. C. Hutchison & J. K. Wright	Solanum sanchez-vegae S.Knapp	US	✔	✔	✔	-	✔
P.C. Hutchison & J.K. Wright	Solanum sanchez-vegae S.Knapp	USM	✔	-	-	-	-

There are 7 digitally available records for this set of specimens, from 5 separate institutional specimen repositories. (See table 6.2, table data source: gbif.org) These are independently managed and not interlinked. As per the first example, despite being generated from the same collection event, the specimen metadata show variation due to isolated management in separate repositories, with all annotation categories holding inconsistent information: 3 of the 7 are georeferenced, 5 of the 7 specify a type status, 5 of the 7 have an associated image and 2 of the 7 have an outdated scientific name. We can therefore calculate that of the 7 digitised specimens in the group: 4 could receive a georeference, 2 could receive a type status annotation and 2 could be linked to an associated image.

These examples show that the separate specimen records held in different specimen repositories hold divergent metadata, and that there is the potential for metadata propagation between members of a specimen group. Specimen groups can be identified by grouping on the collector, their field-assigned record number and the eventdate, but this is non-trivial due to the variation in the recording style of the collecting team (shown in the recordedBy column), as duplicate records have been independently digitised to different data standards in different institutions and projects.

6.4 Methods and materials

6.4.1 Data

A dataset of specimen data relating to vascular plants (those with specialised tissues for the transport of water, encompassing ferns and allied groups, and all seed plants) was downloaded from GBIF (GBIF.org 2018) in Darwin Core (Wieczorek et al. 2012) archive format. This was input into a data-mining process based on the clustering technique DBSCAN in order to detect collector entities, as outlined in chapter 4. Specimen records are eligible for data-mining if they have a numeric component in their *recordnumber* (the sequential number managed by an individual collector and assigned to field collection events), a precise date recorded to the level of day (*eventdate*), and a collector name (*recordedby*). The data-mining process augments the specimen dataset with a numeric identifier for the primary collector of the specimen represented in the metadata record. This allows data to be grouped as the product of the work of a particular collector, irrespective of the lexical variation in the transcription of the collectors names.

6.4.2 Detection of duplicate groups and establishing a confidence measure

A group of specimens are asserted to be generated from a single collection event if they share the same collector identifier (the results of the collector data-mining exercise), eventdate (when the field collection event was carried out) and collector-assigned record number. The record number has any alphabetic prefixes stripped from the value - this normalises values which are sometimes presented with the surname of the collector as a prefix in the recordnumber field.

A confidence measure is applied to candidate duplicate groups by examining the range of variation in fields within the duplicate group. Three assessments are made, a spatial assessment using the countrycode field (duplicate specimen records originating from the same collection event should logically be located in the same country) and two taxonomic assessments using the order and family fields. Biological taxonomy uses a hierarchical system, where species are arranged into families, and families into orders. Although a specimen may be re-determined (have different scientific names applied to it) during its lifetime in a specimen repository, it is less likely to be re-determined across higher taxonomic boundaries. These flags detect variation in these higher-level categories within a duplicate group.

Three Boolean flags were created (one for each assessment field), these were set to True if all members of the candidate duplicate group share the

Algorithm 6.1: labelDuplicateGroups

```
Input  : specimens
Output: labelled_specimens
Let duplicate_groups be specimens grouped by specimen.collector_id,
  specimen.eventdate, specimen.recordnumber
Apply an identifier to each group:
for i ←| duplicate_groups | do
    duplicate_group = duplicate_groups[i]
    Assert a duplicate_group identifier
    for specimen in duplicate_group do
        specimen.duplicate_group_id = i
    end
    Examine range of variation in assessment fields
    for assessment_field in {countrycode,order,family} do
        Create a new Boolean field ${assessment_field}_conservative, which
          is set to True if all members of the duplicate group share a single
          value for this field
        assessment_values = []
        for specimen in duplicate_group do
            assessment_values.append(specimen[assessment_field])
        end
        assessment_values = unique(assessment_values)
        duplicate_group.assessment_conservative ← |assessment_values| == 1
        Copy assessment flag value down to specimen level:
        for specimen in duplicate_group do
            specimen.${assessment_field}_conservative =
              duplicate_group.${assessment_field}_conservative
        end
    end
    Establish overall assessment field:
    conservative = True
    for assessment_field in {countrycode,order,family} do
        conservative = conservative and
          duplicate_group.${assessment_field}_conservative
    end
    for specimen in duplicate_group do
        specimen.conservative = duplicate_group.conservative
        labelled_specimens.append(specimen)
    end
end
return labelled_specimens
```

same value of the assessment field. All possible combinations of these three flags were used to assess the duplicate groups. Only duplicate groups meeting the most conservative assessment criteria (where all of the assessment flags are True, indicating no variation in these fields within the duplicate group) were carried forward for use in subsequent analyses.

This process is summarised in procedure listing 6.1. The input into this algorithm is a tabular data structure where each row represents a specimen, with fields for collector_id, eventdate, recordnumber, countrycode, order and family.

6.4.3 Assessing annotation status per specimen and detecting groups with uneven annotation statuses

Boolean flags were created to indicate if the specimen is georeferenced, if the specimen has an associated image, and if the specimen is recorded as having type status. *Typestatus* values were used as described in (Bebber et al. 2012).

Algorithm 6.2: findPropagableAnnotations

```
Input  : labelled_specimens
Output: assessed_labelled_specimens
Let duplicate_groups be labelled_specimens grouped by duplicate_group_id
for duplicate_group in duplicate_groups do
    Let specimens be the set of specimens included in duplicate_group
    Annotation fields are Boolean flags indicating if the specimen has this
     annotation set
    for annotation_field in {georeference, type_status, image} do
        dg.${annotation_field}_propagable = any(specimen.annotation_field)
         and not all(specimen.annotation_field)
        Copy the annotation_propagable field down to specimen level:
        for specimen in duplicate_group do
            specimen.${annotation_field}_propagable =
             dg.${annotation_field}_propagable
            assessed_labelled_specimens.append(specimen)
        end
    end
end
return assessed_labelled_specimens
```

For each annotation examined, two new Boolean fields were created on the aggregated dataset - these are set to True if *all* specimens in the duplicate group have the annotation set and if *any* specimens in the duplicate group have the annotation set. A group is said to have propagable annotations if it has *any* and *not all* annotations set for the specimens within the group. Two count fields were also created for each annotation, these were set to hold the number of specimens within the group with and without the annotation set. The number of specimens which could receive propagable annotations was determined by totalling the number of specimens within groups with propagable annotations which did not themselves have the annotation set.

This process is summarised in procedure listing 6.2. The input into this algorithm is a tabular data structure where each row represents a specimen, with a field for duplicate_group_id and a set of Boolean fields to indicate the presence of annotations on the specimen (georeference, typestatus, image). This is the assessed, labelled output from the preceding algorithm 6.1.

6.4.4 Repository relationship analysis

The dataset of specimen duplicate groups and the institutions in which they are stored is a very inter-connected dataset - due to relatively few institutions

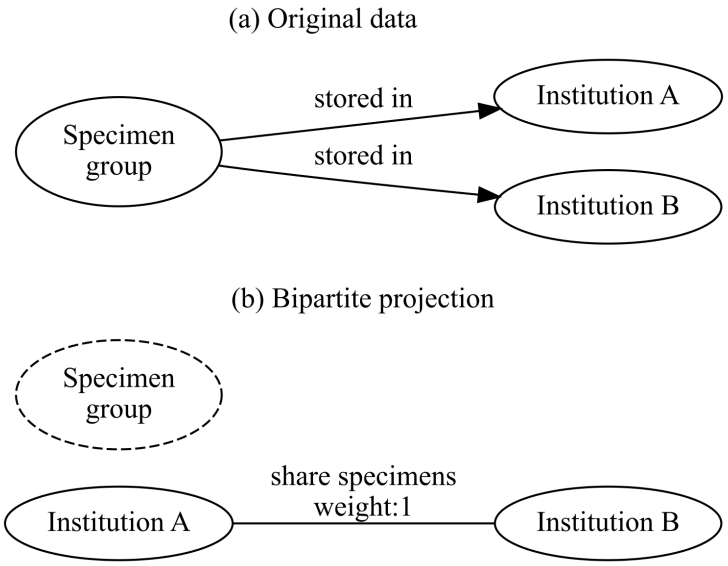


FIGURE 6.2: Bipartite graph projection to infer inter-institutional relationships. (a) Original dataset represented as a directed graph composed of two node types: a large number of *specimen groups*, stored in relatively few *institutions*. A bipartite projection (b) infers weighted relationships between institutions based on shared specimen groups and discards the unused specimen group nodes, resulting in a more tractable weighted (but undirected) graph that can be used for institutional community detection.

being repeated many times as the holders of specimen material. This can be easily represented as a graph data structure, as introduced in section 3.2.3. The graph is composed of two node types - *specimen groups* and *institutions*, and is a *bipartite graph*, as relationships are only permitted between nodes of different types - from specimen groups to the institutions in which they are held. The sharing of specimens in a duplicate group implies a relationship between the two (or more) institutional repositories participating in the group, via a *bipartite graph projection* from an original graph of specimen groups and holding institutions to a *projected graph* consisting only of institutional nodes. This projection process is outlined in figure 6.2. The resulting data structure is a weighted, undirected graph. This inferred network data structure is visualised in Gephi (Bastian et al. 2009), using an OpenOrd (Martin et al. 2011) layout following modularity analysis (Blondel et al. 2008) for community detection, as introduced in section 3.2.3.

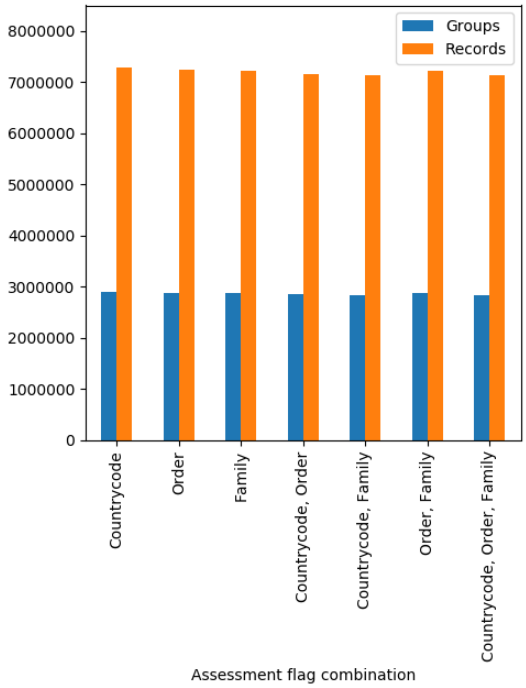


FIGURE 6.3: Duplicate identification assessment: numbers of groups, and numbers of specimen records included in groups

6.5 Results

6.5.1 Data-mining

The initial dataset downloaded from GBIF contained 63.27M records, of these 19.93M records were eligible to be input into the data-mining process to detect the collector. The data-mining process resulted in 19.49M specimen records being labelled with an identifier for the collector.

6.5.2 Duplicate identification and assessment

Of the 19.49M data-mined records, 7.37M records participate in a duplicate relationship, forming 2.92M duplicate groups. All combinations of assessment flags with associated group and record counts are depicted in figure 6.3.

Only the subset of duplicate groups meeting the most conservative assessment criteria were used in subsequent analyses: 7.13M specimens in 2.83M groups. The sizes of the conservatively assessed duplicate groups are shown in figure 6.4.

6.5.3 Propagation of annotations

Members of duplicate sets are located at different institutional repositories and therefore may have been curated differently. Reconciliation of duplicate sets allows the propagation of several classes of annotations - georeferences,

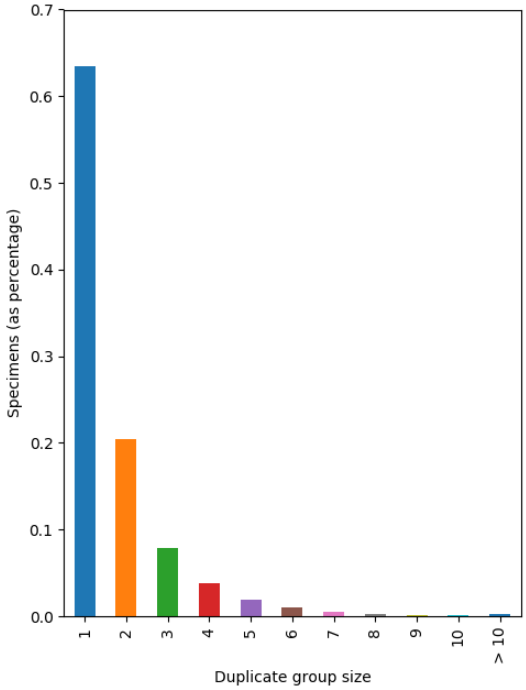


FIGURE 6.4: Sizes of conservatively assessed specimen duplicate groups

type citations, specimen images and determinations - between holders. Of the conservatively assessed duplicate sets:

- 93.5k specimens in 54.61k groups could receive a type citation from a peer in their duplicate group
- 1.13M specimens in 787.73k groups could receive a georeference from a peer in their duplicate group
- 1.11M specimens in 765.72k groups could be linked to an associated specimen image from a peer in their duplicate group
- 2.2M specimens are in 795.73k groups which have multiple scientific names within the group (indicating uneven scientific name determination amongst the members of the specimen duplicate group)

6.5.4 Repository relationship analysis

The relationship graph derived from duplication links at institutional level (see figure 6.5) comprises 245 nodes (institutions) and 6,042 weighted edges (relationships between institutions, based on co-participation in a specimen duplicate group, weighted by the number of co-occurrences). The graph was found to contain eight communities. As the communities appeared to be correlated with the country of the institution, the graph was plotted spatially

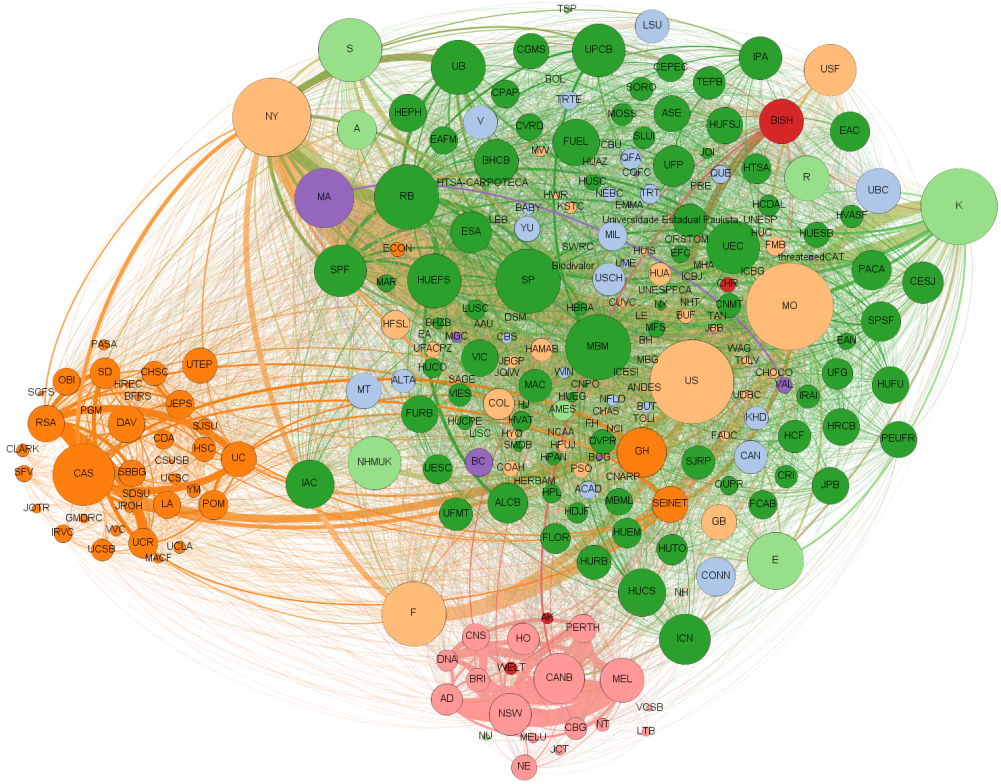


FIGURE 6.5: Institutions connected in an inferred network graph, with communities indicated by Louvain analysis. Layout uses the OpenOrd algorithm, nodes are sized using degree and edges are scaled using weights. Communities as follows: international ; US, regional ; international ; South American, primarily Brazil ; international ; Pacific, New Zealand and Hawaii ; Australia ; Spain

(see figure 6.6). A heatmap was constructed to indicate the correlation between graph community and country of institution (see figure 6.7).

6.6 Discussion

6.6.1 Duplicate identification and assessment

A considerable number of duplicate groups were found in the data-mined dataset, and these appear relatively stable across the different assessment flag combinations (see figure 6.3), permitting the reconciliation of many specimen duplicates between different specimen repositories. The reconciliation of specimen duplicate groups show that many metadata annotations could be propagated between specimen repositories. As these annotations represent both the most expensive parts of the digitisation process (georeferencing) and the most valuable kind of usage citation (type citation), mobilising these between partners would reduce data management costs, improve the utility of the digitised specimen data and improve institutional-level data-usage reporting. It is only possible to supply an estimate range for the cost saving of

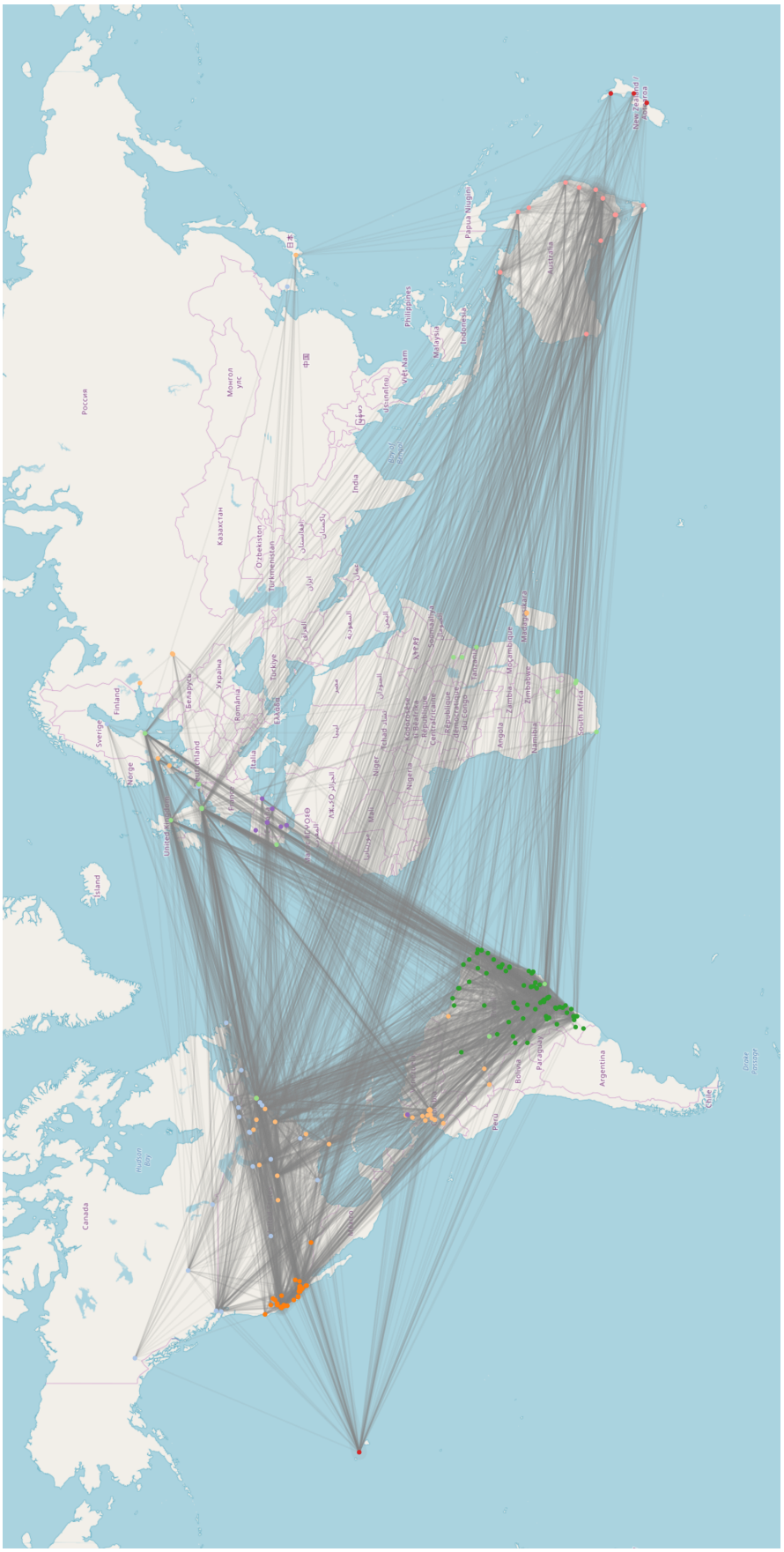


FIGURE 6.6: Spatial layout of institutions connected in an inferred network graph, with communities indicated by Louvain analysis. Node colour as indicated in figure 6.5

FIGURE 6.7: Correlation between country of institution and graph community. Communities as follows ((numbered 1-8, read left to right)): international (1) ; US, regional (2) ; international (3) ; South American, primarily Brazil (4) ; international (5) ; Pacific, New Zealand and Hawaii (6) ; Australia (7) ; Spain (8)

mobilising such a large number of georeferences. Standard procedures tend to batch work by locality, which improves georeferencing speed by focussing on a particular area. A software description paper reports a project georeferencing at a rate of 16.6 (8.3) georeferences per hour and a further separate project achieving a doubling of this rate (Hill et al. 2009). A herbarium type specimen focussed project reported "whole process of georeferencing the ca. 3400 Type specimens took eight months (appx. 100 specimens per week)" (Garcia-Milagros and Funk 2010). It seems that there are significant savings that could be made using the results of this research, given that the number of propagable georeferences is counted at around a million.

6.6.2 Repository relationship analysis

The different repositories represented in the dataset are well connected. Viewed at an institutional level, the low incidence of isolated cliques shows healthy inter-institutional working relationships in botany. There are strong links among regionally focussed herbaria in the United States and Australasia. The interconnections between the Brazilian herbaria and their international counterparts show the volume of work that has been focussed on the world's most mega-diverse country (R. A. Mittermeier and C. G. Mittermeier 1997) and also suggest that the data repatriation projects which aim to mobilise data held out of country (REFLORA 2017) have been

successful. Quantifying the links between specimen repositories enables evidence drawn from specimen duplicate sharing to be used when building project collaborations. Sets of institutions could be selected to maximise overlap or to maximise complementarity. Better sharing of specimen data between institutions facilitates community curation and helps to reduce data management costs.

6.7 Further work

There are several areas in which future work could develop this analysis including further refinement of the analytical approach to cover more data sources, community assessment of interlinked repositories and quality control of annotations by comparison between duplicates. It may be useful to separate future work into two streams: a stream regarding *data management and refinement of data analysis*, and a more conceptual stream regarding *implications of the results*. An example from each area is outlined here: investigation of the reasons why specimens are not currently identified as duplicates ("singleton analysis"), and further work on the research recognition of determination annotations made on specimen objects.

Singleton specimens may be due to uneven digitisation and / or lack of participation in data-mining process, rather than true singletons, further data analysis work is required to investigate this. The heatmap shown in figure 6.8 shows the presence of specimen material for a particular collector. With these per-collector characterisations of the data, it should be possible to calculate for each collector the likely number of specimens gathered at each collecting event. These numbers would give us a potential view on the number of currently un-digitised specimens, and among these, the likely institutional locations of duplicate specimens.

Traditional taxonomic activity can be separated into three phases - collection of specimens, labelling specimen with names and formal publication of results. The first two phases are absent from traditional publication focussed career credit, yet generate long-term research-grade outputs which may be consulted and referenced by others. As these outputs are now mobilised and used much more widely (due to data mobilisation via the internet) there have been calls for these to be included in the career assessment system for taxonomists (McDade et al. 2011). If we recognise that specimens are persistent research objects, which can be uniformly accessed (Güntsch et al. 2017), then the labelling of specimens with scientific names could each be considered to meet the minimum criteria for a *nanopublication* (the smallest unit of research work (Groth et al. 2010) credited to individual researchers).

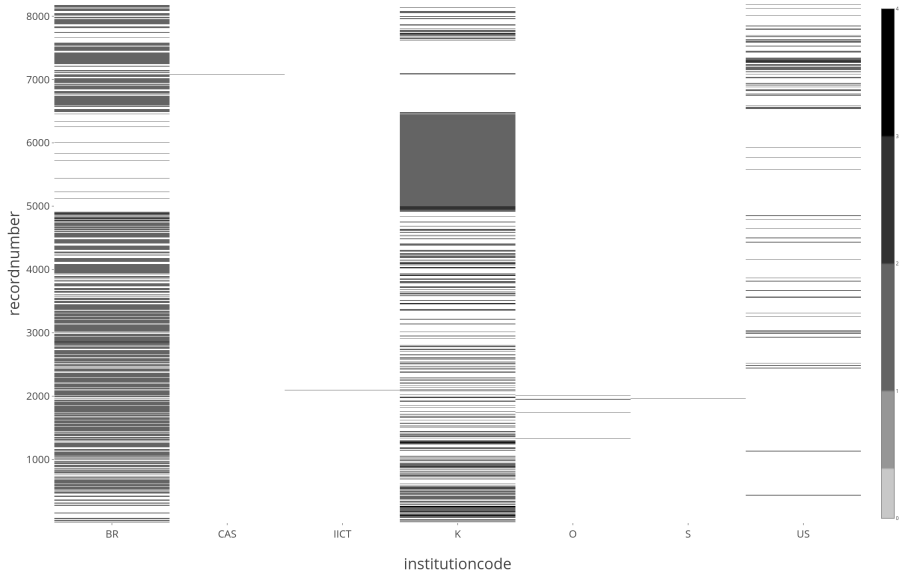


FIGURE 6.8: Heatmap of institutional location of specimens from a single collector (Bidgood). The institutions in which material has been found are listed on the x axis, and the recordnumbers for the collecting events are listed on the y axis. The comprehensively digitised block of collecting events with specimen material in K (numbered c.5000-6500) are likely duplicated but as-yet undigitised elsewhere (e.g. in BR and/or US).

6.8 Conclusion

Specimens are research objects which are managed for long term consultation, facilitate scientific discovery and act as vehicles for the dissemination of results. This chapter demonstrates that specimens form a shared global resource, and that fragmented information management can be overcome by the reconciliation of specimen duplicates across institutional boundaries. Specimen digitisation efforts and work to define standard representations of digitised metadata have built a critical mass of computable information, which can be used as the input into this process. Identification of specimen duplicates allows quantification of potential specimen metadata exchange between institutional specimen repositories. The result of implementing this data exchange would be to develop and strengthen ties between institutional repositories, improve efficiency of data curation (by eliminating repeated work such as specimen georeferencing) and to improve the metadata holdings and reporting figures for institutional repositories. Conceptually, specimens should be recognised as a unit of research work more granular than the scientific paper, but fulfilling the same functions - communication of results and establishment of a long term record. This recognition of the specimen as a research object would eventually allow the annotation of specimens to be regarded as research work and credited to individual researchers. This may start to address some concerns recently

voiced with regard to the many phases of research work conducted by taxonomists which remain absent from publication-focussed career metrics (McDade et al. 2011).

Chapter 7

Conclusions

This chapter re-states the research objectives and context, and evaluates the activities undertaken. Potential criticisms of the work are anticipated and addressed, and potential revisions, generalisations to enable wider application, and suggestions for future related work are given.

7.1 Objectives

The objectives of this work were to:

- Map high-level concepts, identifying key areas for work
- Translate approaches and analyses from smaller-scale editorially managed datasets to larger-scale aggregated heterogeneous datasets
- Develop methods to automate the construction of higher-order data representations
- Enable wider-scale trends analysis

7.1.1 Mapping high-level concepts

A concept map was introduced (figure 1.1), and compared with two existing high-level mappings used in the biodiversity informatics domain: the GBIO framework (figure 2.2) (Hobern et al. 2019), which defines layered focus areas which support higher-level and larger-scale scientific analyses, and the proposed "biodiversity knowledge graph" (figure 2.3) (R. D. Page 2016), which depicts the interactions between recognised data concepts. The concept map was used to visually define the context and scope of each of the analyses in the preliminary and research chapters. A summary of the complete set of concepts covered by the research presented in this thesis is shown in figure 7.1. This shows that the data-mining processes have enabled analyses on collecting activities, agent participation and specimen reconciliation, covering the following concepts:

- *collecting trip*

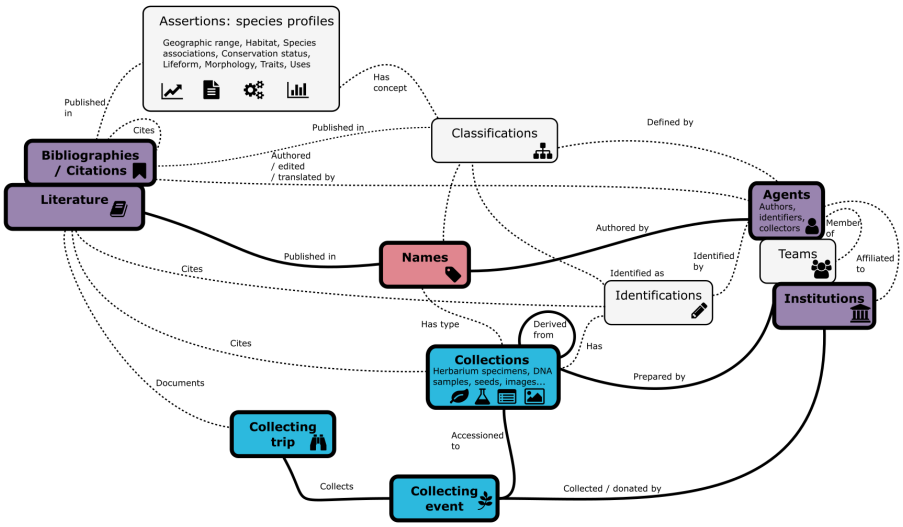


FIGURE 7.1: Visual context: summary of concepts covered in the research presented in this thesis. The highlighted areas cover: *collecting trip*, *collecting event*, specimen *collections*, *institutions* in which the specimens are managed, *agents* responsible for specimen collection and authoring of *scientific names* and the *literature* in which those names are published. The concept map also highlights potential areas for future work - *identifications* and *teams* - given progress made on nearby and referring concepts

- *collecting event* (a sequence of which forms a collecting trip)
- specimen *collections*
- the *institutions* in which those specimen collections are managed
- the *agents* responsible for both specimen collection and authoring of *scientific names*
- the *literature* in which those scientific names are published

7.1.2 Translating approaches and analyses within the biodiversity informatics domain

Three aspects of translation of approach are covered here:

- Translation of approach from analyses on editorially managed data to larger scale, heterogeneous, aggregated datasets
- Translation of approach between systematic domains (from botany to zoology)
- "Reverse" translation of approach - proposing application of computational techniques to aid creation of traditionally editorialised data

The preliminary analysis presented in chapter 3 was based on the publication of new scientific names by an author: units of work contained within a scientific publication. This process was generalised to identify the *object* (scientific name), *creator* (author) and *container* (the publication). This object-creator-container paradigm was applied in the definition of a data-mining process (chapter 4) to a wider scale, more heterogeneous dataset of aggregated specimen data. The data-mining process was applied to specimen metadata to identify object (specimen), creator (collector) and container (collecting trip), allowing the reshaping of the dataset to analyse trends in collector participation and activity over time.

The data-mining process defined in chapter 4 was designed around field practices in botany, but in generalising the process and examining the number of records eligible to participate in the process, it was shown that the process could be revised to track collector activity through physical space and time (where the physical space is represented by geo-positioning coordinates) rather than through numeric space and time. This widens the application of the technique to include zoological material - which is likely to lack the sequential collector-assigned recordnumbers commonly used in botany.

The presentation of research in this thesis has translated approaches from editorialised sources to larger, less-managed sources, defining novel computational techniques that can be applied to the larger-scale data-sources. It is important to recognise that translation of approach need not only work in a single direction, and that the techniques defined here may also be usefully applied to the working practices which generate editorialised data. The collector data-mining approach is based upon the identification of traces of activity through numeric and temporal space in specimen metadata. The discussion of this process (in the conclusions for chapter 4) noted the potential generalisation to identify traces of activity in literature sources, opening the possibility of an iterative translation of approach.

7.1.3 Automating the construction of higher-order data representations

A summary of the higher order data representations used and constructed in the research presented in this thesis is given in table 7.1. Coupled with the visual context provided in figure 7.1, this shows that the work has covered a number of the key entities defined at the start of the research process. The research has additionally defined new higher-order representations not conceived at the start of the process - institutional community - and discussed the use of these in meta-analyses to optimise data curation and plan

digitisation work.

TABLE 7.1: Summary of the higher-order data representations used and/or defined from the work presented in this thesis

Analysis	Higher-order data representation
Preliminary analysis: e-publication (chapter 3)	*Author* (editorially created)
Data-mining from aggregated specimen data (chapter 4)	*Collector*, their *collecting trip*s, and within trips, *collecting state run*s (computationally created)
Agent integration (chapter 5)	*Scientist*: integrating *author* and *collector* (computationally created)
Specimen reconciliation (chapter 6)	*Collecting event* (specimen duplicate group) and *institutional community* (computationally created)

7.1.4 Enabling wider-scale analyses

The analyses facilitated by the definition of these higher-order data representations allow the investigation of *participation* in the species discovery process - the authors participating in the publication of scientific names (chapter 3), and the collectors conducting field work to generate specimens (chapter 4), the *contribution* of individuals towards species discovery (chapter 5), the *reconciliation* of distributed products of a specimen collecting event and the *community detection* of a network of interconnected institutions (chapter 6). These analyses contrast with existing collection analyses, which have used closely defined data and institutional subsets - e.g. type specimens from a select set of institutions (Bebber et al. 2012) or have used summarised data resources covering collector activity rather than the actual specimen data (Penn et al. 2018).

7.2 Evaluation

A fundamental criticism of this work could be the bias towards data which is digitally available and therefore computable. Comparing the number of herbarium specimen records mobilised through GBIF (77.46M records[1]) with the number estimated to exist worldwide from metadata records collated by Index Herbariorum (390.48M specimens) (Thiers continuously updated) shows that the number of specimen records available for computational analysis is still very much a minority. The data are heterogeneous in terms of their completeness as well as in terms of their standardisation. Despite this, it has been possible to use incomplete aggregated datasets to detect collector entities, and to model the activities of collectors by detecting their collecting trips. The areas of research mostly likely to be affected by incomplete data are the specimen duplicate assessment presented in chapter 6 and the inferred

[1]Numbers calculated from GBIF API call executed on 2019-11-05

institutional relationship network presented in the same chapter. It is likely that a number of the records currently assessed as without duplicates in other collections really are duplicated, but the duplicate specimens are not yet digitised and are therefore not computable. The conclusions for chapter 6 propose additional work to characterise the relationships between collectors and numbers of specimens gathered at collecting events, and the relationships between collectors and institutions to make some estimates of the undigitised portion of the data. This area will be investigated in follow-on work which will apply the techniques developed in this research to data derived from literature rather than specimen digitisation.

7.3 Future work

7.3.1 Development of methods and analyses operating on aggregated specimen metadata

This work has developed methods to utilise and detect **agent** entities responsible for key stages of the systematic process (collecting and publication of species descriptions). The intermediate stage - identification - is yet to be addressed. This is an obvious area for future work, given that surrounding stages and elements have been treated in this research (see figure 7.1). Trust in the scientific name applied to a biological specimen and associated metadata is a key factor in downstream use (Goodwin et al. 2015), and the application of names to specimens is an activity which has been noted as providing some form of career credit for the researchers involved (McDade et al. 2011). Further research on the agents involved in systematics activities and methods by which their representations can be data-mined and integrated is the subject of an EU funded project MOBILISE (further details in appendix D.2 and conference papers in appendix sections C.5 and C.4).

Another area absent from the analyses presented in this research, but a candidate for future work is to identify the collecting **team** and the co-working relations between collectors and authors. As above, this work will be simplified by the handling of the surrounding data elements (agents and institutions - see figure 7.1). The revision of the collector data mining process to utilise spatial traces and therefore to apply this to all members of the collecting team will aid the identification of teams and their members. This is an extension of the existing collector detection process, which in utilising primary collector's record-number, was only applied to the primary collector rather than all members in the team.

The **specimen reconciliation** analysis presented in chapter 6 will be further developed with a set of partners to report on the potential for annotation sharing and quality control (core-funded 1 year project), to

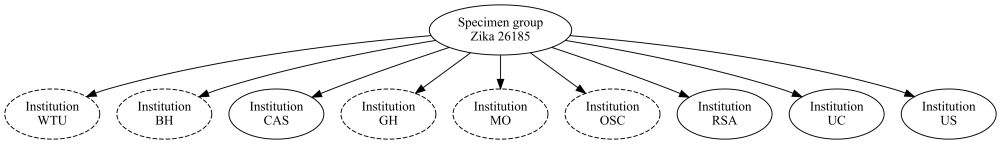


FIGURE 7.2: Rendering of a literature-derived specimen group into a graph structure. Institutions cited as holding specimen material but without digitised records for the specimen (therefore absent from the analysis presented in chapter 6) are shown with a dashed border.

explore the potential for shared data curation between communities of institutions.

7.3.2 Alternative data sources: literature

Alongside specimen digitisation, many commercial publishers and academic projects (e.g. the Biodiversity Heritage Library) are digitising literature sources relevant to systematics. As specimens are cited as evidence in published works (as seen in the examples in chapter 6), literature can be a source of specimen data via text-mining, as minimal metadata (collector, recordnumber, date, locality and institution codes) are available within specimen references.

A follow-on piece of work (scheduled for conference presentation, details in appendix C.6), presents an initial process to exploit the standardised presentation of specimen references in literature to extract specimen metadata.

A sample literature-derived specimen citation is shown below, as per the worked example used in chapter 6:

> **Sedum citrinum** Zika, *sp. nov.* **Type**: UNITED STATES. California: Del Norte County, ridge 1.4 air km north of South Red Mountain, 1050 m, 9 June 2013, *P. F. Zika 26185* **(holotype: WTU!; isotypes: BH!, CAS!, GH!, MO!, OSC!, RSA!, UC!, US!)**. (Zika 2014)

The list of institutional codes to indicate the holders of specimen material (shown in bold face) is used as a feature to detect specimen references; the text extraction of these sets of institution codes would allow the creation of an institutional network inferred from co-relationships as defined in literature (see figure 7.2), rather than from collecting events data-mined from an incomplete dataset of digitised specimens.

The process is developed using a dataset of taxonomic publications categorised into paragraph-level units. This is used as training data to construct a binary text classification system to classify component units of articles (sections or paragraphs) as specimen reference holders, using features derived from the text contents to indicate the presence of specimen reference

statements. Units classified as containing specimen references are processed to extract a minimal representation of the specimen reference - the list of institutional codes in which the specimen material is held. These can be used to construct a network of the relationships between institutions, which can be contrasted with the network built from specimen metadata and presented in chapter 6.

7.3.3 Integration with crowd-sourced approaches

"Crowd-sourcing" is commonly used in the generation of data about biodiversity, with digitisation initiatives seeking to mobilise members of the public (or specialists not directly employed by a project) to help with data transcription and image-labelling. The data generated through these projects can be used as training data for the later application of more automated approaches. Crowd sourced projects which allow collectors and determiners to "claim" their specimens and add them to their research profiles (Shorthouse and R. Page 2019), have to potential to act as training and validation data for the kinds of collector data-mining and agent record-linkage techniques developed in this thesis, expanding the use of supervised machine learning techniques.

7.4 Conclusions

This concluding chapter has summarised the objectives of this research project, evaluated activities against these objectives, anticipated and addressed potential criticisms and identified a number of areas of future work. A common thread through these conclusions is to utilise the research outputs through funded future work to better enable the completion of the specimen digitisation task.

Figure 7.3 summarises the routes by which specimen data are digitised to create structured metadata suitable for data aggregation and analysis. These processes are heterogeneous, varying between different projects and institutions. The work presented in this thesis has been conducted with aggregated specimen metadata, but the products of this research have potential utility in expediting the specimen digitisation process. These include the propagation of specimen metadata as outlined in chapter 6, the generation of resources which can aid human-conducted transcription, and more automated approaches which use these products as training data for data extraction processes - to support the "digitization 2.0" initiative (Hedrick et al. 2019). These activities will be explored in a funded project *SYNTHESYS+* (further details in appendix D.1). Future work on the application of the techniques presented in this thesis at the aggregator level is being investigated with GBIF.

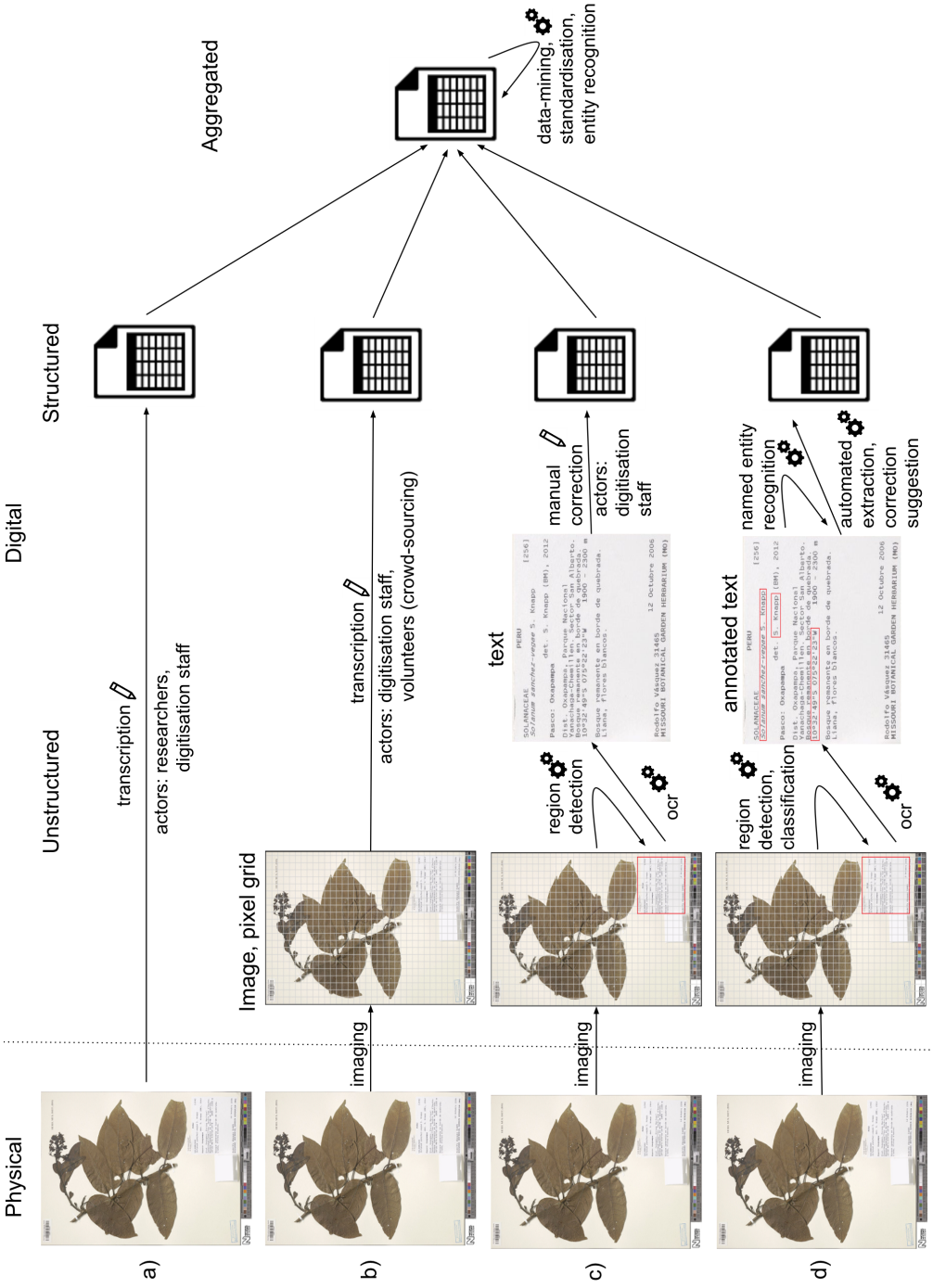


FIGURE 7.3: Digitisation pathways: (a) traditional (b) mass digitisation (c) augmented (d) intelligent

This work has shown that the data derived from botanical specimens can be re-purposed to construct higher-order data representations, which can be used to better understand the species description process. The use of a heterogeneous, aggregated data-set to detect these data representations has uncovered the level of data sharing between institutions. This work has demonstrated the potential for the progress of specimen digitisation and on-going curation as a global collaborative task, recognising the individual effort of collectors and authors in gathering specimens and using them to describe biodiversity.

Glossary

Note: glossary entries include scientific terms from systematics and biodiversity informatics, project acronyms, and technical terms from computer science and information management. Where these are general terms with a specific meaning in a scientific context, the context is indicated in the definition.

accession (systematics) the process of adding new material to a collection.

angiosperm (systematics) plants which produce flowers and have enclosed seeds.

Angiosperm Phylogeny Group (systematics) an effort to create a high level phylogeny for angiosperms, often used in reporting and to organise collections.

annotation (systematics) the addition of data to a specimen sheet.

APG see Angiosperm Phylogeny Group.

arc (computer science) synonym for edge, particularly in a directed graph.

AUC Area Under (the) Curve.

BHL see Biodiversity Heritage Library.

Biodiversity Heritage Library digitisation project to photograph and OCR literature relevant to the study of natural history. `https://www.biodiversitylibrary.org/`.

Biodiversity Information Standards (TDWG) standards body responsible for the definition of data interchange standards in biodiversity informatics. The acronym TDWG relates to the original name of the organisation (the Taxonomic Databases Working Group). `http://www.tdwg.org`.

class (computer science) in machine learning, a category into which samples may be assigned.

class (systematics) a unit in the taxononomic hierarchy.

classification (computer science) a supervised machine learning technique used to assign samples to known classes. When the problem domain has only two classes (which are usually represented as positive and negative), this is known as *binary classification*. In contrast *multi-class classification* describes a problem domain where the number of possible class labels is greater than 2. *Multi-label classification* describes the situation where multiple class labels may be assigned to a single sample.

classification (systematics) a particular taxonomic scheme used to organise biological entities into categories.

clustering (computer science) unsupervised machine learning technique which assigns unlabelled input data samples into categories.

collector (systematics) the person responsible for the field collection of biological material.

Convention on Biological Diversity international convention.

Convention on International Trade in Endangered Species international convention.

CSV Comma-Separated Values, a file format used to represent tabular data.

Darwin Core a data standard comprising a set of metadata terms and definitions used to encode biodiversity data, and used for harvesting into the GBIF system. `https://dwc.tdwg.org/` (Wieczorek et al. 2012).

Darwin Core Archive a single zip file packaging up a set of Darwin Core datafiles. (Remsen et al. 2017).

DBSCAN Density-Based Spatial Clustering of Applications with Noise (Ester et al. 1996).

degree (computer science) in a graph structure, the number of edges incident to a node.

determination (systematics) synonym for identification.

determiner (systematics) the person responsible for the identification of specimen material.

digitisation (biodiversity informatics) the process of creating a digital record from a physical specimen. May include one or more of: structured metadata creation, imaging, georeferencing.

distribution (systematics) the geographical spread of a species.

DOI Digital Object Identifier.

duplicate specimen (systematics) specimens arising from a common field collecting event.

DwC see Darwin Core.

DwCA see Darwin Core Archive.

edge (computer science) the connection between nodes in a graph data structure.

Encyclopaedia of Life aggregation project assembling species profiles `http://www.eol.org`.

endemic (systematics) of a species - found only in a certain region.

EOL see Encyclopaedia of Life.

family (systematics) a unit in the taxononomic hierarchy.

field book (systematics) an inventory of collection locations and habitat descriptions, field notes etc created by a collector and cross referenced to collected material.

filtered push project to mobilise specimen annotations between institutions to interested parties (Macklin et al. 2006) (Wang et al. 2009).

flora (systematics) a listing of the plant species found in a defined area, often published with descriptions and identification keys.

GBIF see Global Biodiversity Information Facility.

GBIO see Global Biodiversity Informatics Outlook.

genus (systematics) a unit in the taxononomic hierarchy.

georeferencing (biodiversity informatics) the process of determining latitude / longitude coordinates from a textual description of a place.

Global Biodiversity Informatics Outlook a framework for understanding the world's biodiversity based on four focal areas: culture, data, evidence and understanding (Hobern et al. 2012). `https://www.biodiversityinformatics.org/en/gbio-framework/overview/`.

Global Biodiversity Information Facility an intergovernmental initiative to aggregate and disseminate biodiversity data, formed as a response to the Organisation for Economic Cooperation and Development (OECD) megascience forum. `http://www.gbif.org`.

Global Taxonomic Initiative part of Convention on Biological Diversity https://www.cbd.int/gti.

GPS Global Positioning System.

graph (computer science) a data structure composed of nodes and edges. May be *directed* (edges have an orientation) or undirected. May also be *weighted* (edges hold a value indicating the weight or strength of the relationship). A *property graph* allows the storage of information (properties) on both the nodes and edges.

H-index a bibliometric calculation summarising the breadth of citations of a set of literature outputs.

herbarium (systematics) a collection of preserved dried plant specimens. Each herbarium is listed in Index Herbariorum and assigned an alphabetic code.

HMM Hidden Markov Model.

holotype (systematics) a class of type specimen, the single specimen intended to be the bearer of a scientific name.

HTTP HyperText Transfer Protocol.

IBC see International Botanical Congress.

ICBN see International Code of Botanical Nomenclature.

ICNafp see International Code of Nomenclature for algae, fungi and plants.

ICZN see International Code of Zoological Nomenclature.

identification (systematics) the labelling of a specimen with a scientific name. An identification is usually accompanied with the name of the person making the identification (the determiner) and the date it was made, it sometimes also includes the institutional affiliation of the person making the identification, and may also include the purpose of the identification (e.g. work on a particular project or towards a published output).

identification key (systematics) a structured tool to aid identification of species.

iDigBio US National Science Foundation funded project to aggregate specimen based information. https://www.idigbio.org/.

IH code see Index Herbariorum.

Index Herbariorum (systematics) resource listing herbaria, with details of holdings, research specialities and staff members. Each institution has an alphabetic code ("IH code") which is used as an abbreviated form of reference to the collection (Thiers continuously updated). http://sweetgum.nybg.org/science/ih/.

Index Kewensis indexing project recording published scientific plant names, now incorporated into International Plant Names Index.

Integrated Publishing Toolkit a GBIF product which helps institutions mobilise their data for harvesting into the GBIF data portal.

International Botanical Congress botanical conference held every 6 years which incorporates a "nomenclature section", at which changes to the International Code of Nomenclature for algae, fungi and plants are proposed, discussed and voted upon.

International Code of Botanical Nomenclature rules governing the naming of organisms traditionally studied in botany, since 2011 renamed as International Code of Nomenclature for algae, fungi and plants.

International Code of Nomenclature for algae, fungi and plants the set of rules and recommendations that govern the scientific naming of all organisms traditionally treated as algae, fungi, or plants, whether fossil or non-fossil. (Before 2011 called the International Code of Botanical Nomenclature). (McNeil et al. 2012) (Turland et al. 2018) https://www.iapt-taxon.org/nomen/main.php.

International Code of Zoological Nomenclature the set of rules governing the scientific naming of organisms treated as animals. https://www.nhm.ac.uk/hosted-sites/iczn/code/.

International Plant Names Index indexing project creating a database of the names and associated basic bibliographical details of seed plants, ferns and lycophytes. http://www.ipni.org.

IPNI see International Plant Names Index.

isotype (systematics) a class of type specimen, a duplicate specimen of the holotype.

kingdom (systematics) a unit in the taxononomic hierarchy.

metadata (systematics) the data recorded about a biological specimen object, including "what, when, where" information about the species represented, and when and where it was collected.

monograph (systematics) a detailed work on a single subject, in systematics, a work focussing on a particular taxon.

mycology the study of fungi.

node (computer science) the fundamental unit of a graph data structure.

nomenclator (systematics) a listing of scientific names, or a project assembling such a list.

nomenclature (systematics) the system of naming for biological organisms.

observation (systematics) the recording of a species in-field, without the long-term preservation of its genetic material.

occurrence (systematics) the recording of a particular species at a particular place and time. May be evidenced with a voucher specimen.

OCR Optical Character Recognition, computational process by which a graphical representation of text is converted into text data.

order (systematics) a unit in the taxononomic hierarchy.

PDF Portable Document Format.

phenology the study of the periodic timing of biological life cycle events (e.g. flowering times, insect emergence, migratory appearance, nesting).

phylogeny (systematics) the study of the evolutionary history of biological entities. Can also be used to refer to a particular representation of evolutionary history ("a phylogeny").

phylum (systematics) a unit in the taxononomic hierarchy.

protologue (systematics) "everything associated with a name at its valid publication, e.g. description, diagnosis, illustrations, references, synonymy, geographical data, citation of specimens, discussion, and comments" (Turland 2019).

recordnumber (systematics) a sequential number used by botanical collectors to cross reference specimen material with field book notes. Also used in specimen references.

ROC Receiver Operating Characteristic.

species (systematics) a unit in the taxononomic hierarchy.

specimen (systematics) a biological sample stored in an accessible collection for long term reference.

specimen reference a reference in literature to a particular specimen (or set of specimens) from a single collecting event. Components of a specimen reference are: collector (primary collector or team), record number, date and institutional locations (indicated by Index Herbariorum code).

supervised learning (computer science) machine learning technique to produce models using labelled training data as input.

systematics the study of the evolutionary relationships of biological organisms.

taxon (systematics) a unit in a taxonomy, an instance of a taxonomic rank.

taxonomic rank (systematics) a level in the taxononomic hierarchy.

taxonomy the branch of science concerned with classification, particularly of biological entities. Can also be used to refer to a particular classification scheme ("a taxonomy").

taxononomic hierarchy (systematics) a hierarchical arrangement of taxonomic ranks, from the broadest (kingdom) through phylum, class, order, family, genus, species.

TDWG see Biodiversity Information Standards (TDWG).

tracheophytes (systematics) synonym for vascular plants.

trait (systematics) a characteristic displayed by a species.

type specimen (systematics) a specimen cited as the bearer of a name.

unsupervised learning (computer science) machine learning technique where models are built without training data, and the structure is learned from the data itself, e.g. clustering.

vascular plants (systematics) plants which develop a vascular system for transport of water. As the vascular system provides physical support, this enables the development of larger body sizes.

vertex (computer science) synonym for node.

voucher (systematics) a specimen used as evidence that a species occurs at a particular place and time.

XML eXtensible Markup Language.

Appendix A

Visualisation tool

This appendix outlines the development and utilisation of an interactive data visualisation tool, developed throughout the research project. A description of this tool was presented as a conference poster and "lightning" talk (the published version is available in appendix C.3).

Originally designed to aid initial data exploration and gather expert input, the visualisation tool was further refined to support process design, quality assurance and refinement by viewing data-mining results at known stages of a pipeline process, and to enable visualisation of data aggregations used to define new features for use in predictive models. Newly defined features can be regarded as additional data, feeding back into data exploration and forming an iterative process. The toolkit has contributed to reproducible research by adding tool support and activity logging at one of the loosest stages of the research process.

A.1 Introduction

The parent research project described in the main part of this thesis can be regarded as inter-disciplinary research, covering both applied computer science and biodiversity informatics. Some of the obvious challenges in a project of this type are the need to work with heterogeneous (messy) data, which is both incomplete and inconsistent, being drawn from many differently managed sources, and the need to elicit expert input at multiple states of the project - to inform initial data exploration, to sanity check initial results, and to refine the process.

Graphical representation of data, particularly datasets featuring combinations of categorical, numeric, geospatial and temporal data provides an intuitive way for a user to understand a summary overview of a large dataset and to interact with particular areas of interest. Many search interfaces allow data selection and exploration using a "dashboard" interface, including the data aggregator (GBIF) used as the source of the data in this project (see figure A.1).

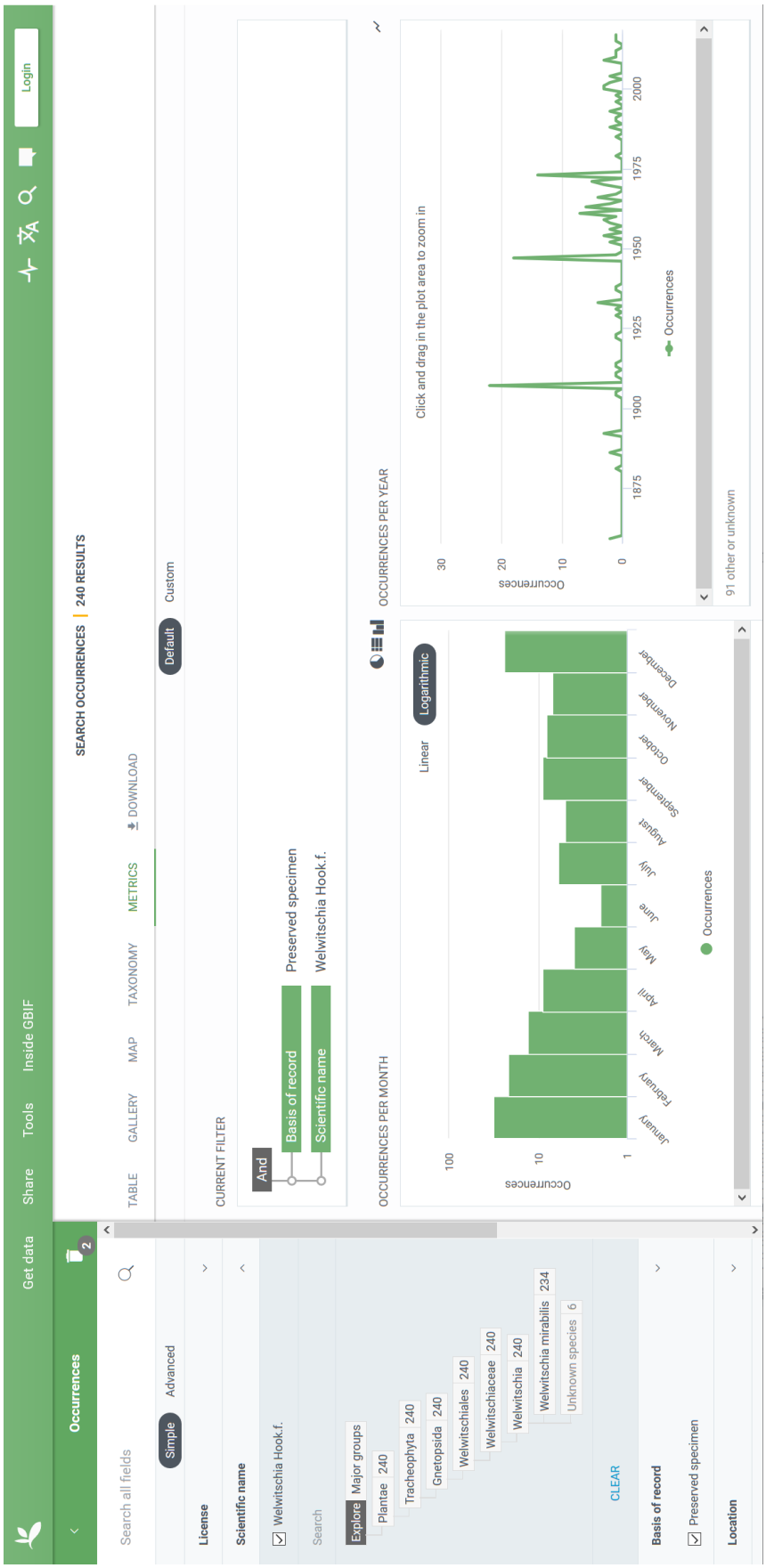


FIGURE A.1: GBIF data exploration dashboard (screenshot). Data selection on right, graphical summary of selected data subset in remainder of screen area.

The parent project exploits some of the field collection processes common amongst botanists to construct a data-mining process to assign field collected specimens to a data-mined "collector" entity. This allows the data to be reshaped to construct higher order data representations - the collector, the collecting trip and particular sequences of intensive collecting activity, which are later used to construct predictive models. As these abstractions are newly created in the data (and relate to a subset of the GBIF data scope, that regarding the field collection of botanical specimens) an existing data visualisation toolkit was not available for the visualisation and exploration of these results, and dedicated visualisation tools were created alongside the core research work conducted for this thesis.

This remainder of this appendix is structured as follows: methods and materials describes the software tools used to construct the toolkit, the data exposed through it and the kinds of visualisations supported. The next section outlines three examples of its use throughout the course of the overall project, with links back to the relevant chapters in the thesis proper. Finally, a short conclusions section summarises the work and offers some ideas for future development.

A.2 Methods and materials

The design aims of the visualisation tool are primarily to maximise access to the tool itself and the data included within it, i.e. to:

- Have a low barrier to entry - no requirement for scripting / programming skills to use the toolkit
- Implement a graphical interface for data subset selection, with live update of visualisation types based on data selections
- Facilitate access to underlying data for seamless navigation

The data packaged into the tool is auto-generated as an output of the data mining process. The toolkit is implemented as a local web application in Dash (`https://plot.ly/`), reading in data-mined results stored in a pandas (McKinney 2010) dataframe.

A.2.1 Visualisation types

A number of visualistion types are available from the home page of the toolkit (figure A.2)

- **Scatter**: shows a view of collecting events plotted as eventdate against record number: the user can select by collector, country of collection,

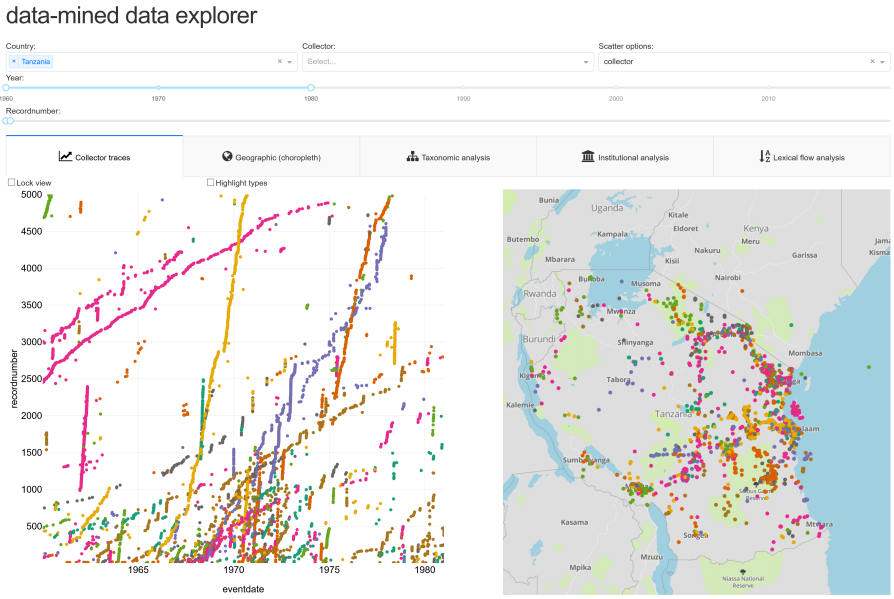


FIGURE A.2: Visualisation tool homepage (screenshot)

temporal and numeric ranges, hover shows record details and click takes the user through to view the record in the aggregator (GBIF)

- **Map**: shows a view of geo-referenced collecting events, selection is synchronised with scatter view described above.
- **Choropleth**: shows a map of collecting event density per political country, for the current selection.
- **Heatmap**: shows a categorical breakdown of the density of collecting events for a single collector. Categories supported are: taxonomic units and institutional holders of specimen material.
- **Sankey**: shows the flow of data through data-mining process steps, as data elements are assigned to different categories, and categories are joined or split.

A.3 Examples of use

A.3.1 Revision of the data-mining process

Interactive visualisation of a trip allows the user to zoom in in increase focus on a particular time-slice of activity from a selected collector. Often, these activity traces show uneven activity - which may represent distinct collecting days (with intensive collecting activity) and travelling days (when collectors are moving between collecting areas and fewer collecting events occur). The data-mining process outlined in chapter 4 was refined following the use of this tool to detect these distinct states using a Markov chain. The visualisation

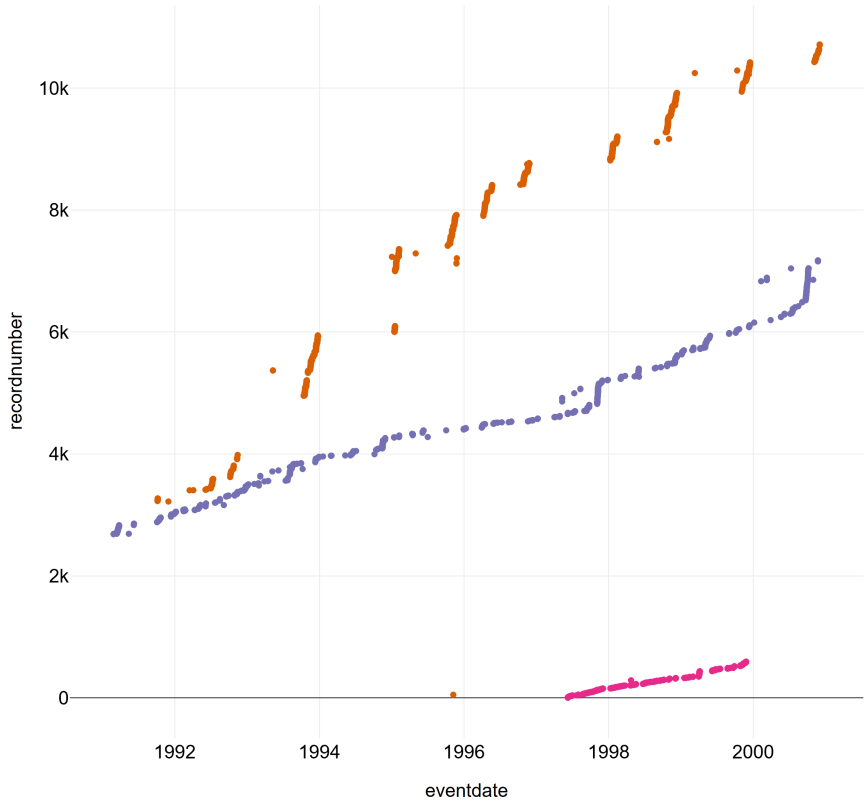


FIGURE A.3: Scatter plot of specimens from multiple separate collectors

tool was then updated to support the visualisation of these more granular aggregations: the linked map plot shows that these are useful distinctions (the example shows that specimens allocated to distinct collecting states are spatially co-located).

A.3.2 Data generation: feature definition

Visualising and comparing several different collector career traces (figure A.3) shows different kinds of activity patterns. These can be used to distinguish periodic collectors (likely to be visiting the collecting area) and persistent collectors (likely to be resident in the country of collection). New features were generated to be used to distinguish these, including the overall slope of the activity trace and the percentage active months in overall career. These features were used in the development of a classifier to detect specimen aggregations pf particular species discovery value (chapter 5).

A.3.3 Research question generation: relations between institutions

The heatmap visualisation (figure A.4) shows where specimens from a particular collector are lodged for long term storage and consultation. Interacting with the data this way prompted a new research question: is it possible to detect a relationship between institutions based on their sharing of

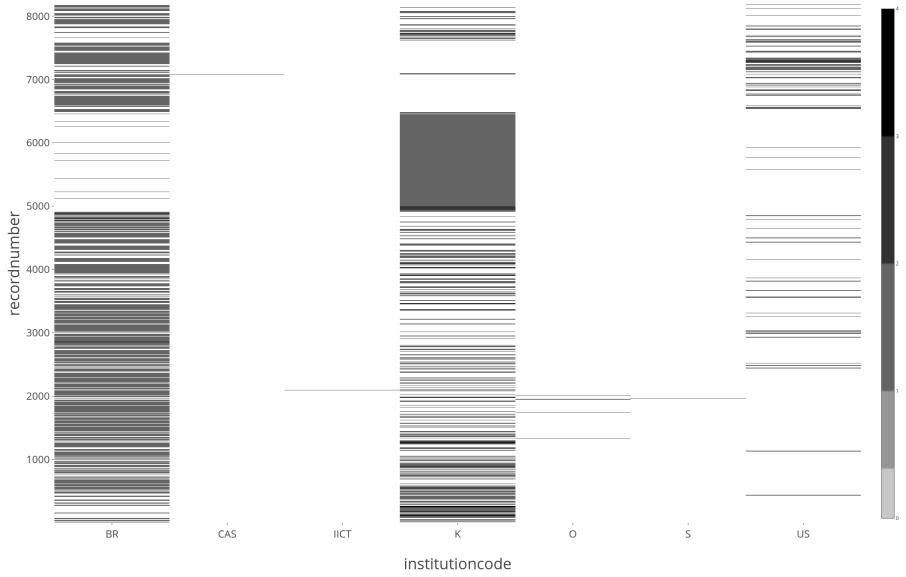


FIGURE A.4: Heatmap of institutional holders for specimens from a single collector

specimen material? As presented in chapter 6, the data were reshaped to a graph structure and analysed to detect communities of institutions. In this case, the graph visualisation toolkit Gephi (Bastian et al. 2009) was used for visualisation purposes (figure A.5), rather than implementing graph visualisation in the toolkit itself.

A.4 Conclusions

This section has introduced an interactive visualisation toolkit designed to aid the exploration of heterogeneous biodiversity data throughout the different phases of this research project. It has been used to:

- Facilitate discussions with experts
- Design and refine a data-mining process and suggest new steps
- Surface new research questions

More generally, this approach has contributed to a reproducible research process by making data exploration a managed step in a pipeline process, generating revision controlled plots and using application logging to provide a record of data exploration activities.

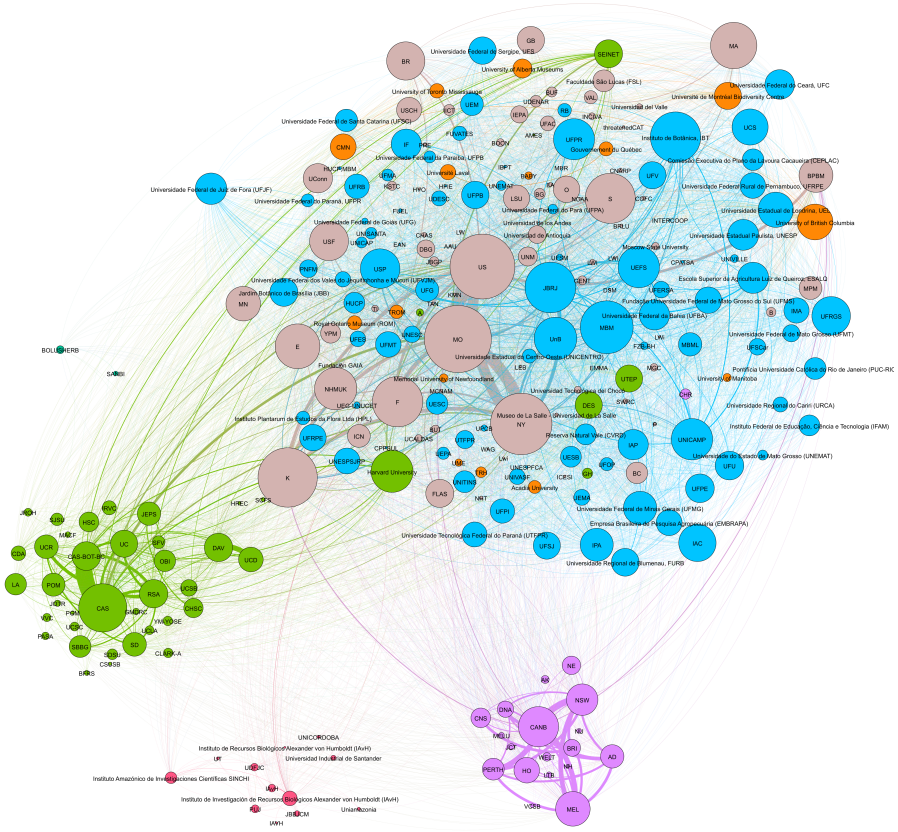


FIGURE A.5: Graph visualisation

Appendix B

Published Articles

This appendix includes published versions of articles arising from the work in this thesis.

B.1 Impact of e-publication changes in the International Code of Nomenclature for algae, fungi and plants (Melbourne Code, 2012) - did we need to "run for our lives"?

This paper results from preliminary work conducted for this thesis, which is presented in chapter 3.

Reprinted from: Nicolson, N., Challis, K., Tucker, A., Knapp, S., 2017. Impact of e-publication changes in the International Code of Nomenclature for algae, fungi and plants (Melbourne Code, 2012) - did we need to “run for our lives”? BMC Evolutionary Biology 17, 116. doi:10.1186/s12862-017-0961-8

Nicolson *et al. BMC Evolutionary Biology* (2017) 17:116
DOI 10.1186/s12862-017-0961-8

BMC Evolutionary Biology

RESEARCH ARTICLE Open Access

Impact of e-publication changes in the International Code of Nomenclature for algae, fungi and plants (Melbourne Code, 2012) - did we need to "run for our lives"?

Nicky Nicolson[1,3], Katherine Challis[2], Allan Tucker[3] and Sandra Knapp[4*]

Abstract

Background: At the Nomenclature Section of the XVIII International Botanical Congress in Melbourne, Australia (IBC), the botanical community voted to allow electronic publication of nomenclatural acts for algae, fungi and plants, and to abolish the rule requiring Latin descriptions or diagnoses for new taxa. Since the 1st January 2012, botanists have been able to publish new names in electronic journals and may use Latin or English as the language of description or diagnosis.

Results: Using data on vascular plants from the International Plant Names Index (IPNI) spanning the time period in which these changes occurred, we analysed trajectories in publication trends and assessed the impact of these new rules for descriptions of new species and nomenclatural acts. The data show that the ability to publish electronically has not "opened the floodgates" to an avalanche of sloppy nomenclature, but concomitantly neither has there been a massive expansion in the number of names published, nor of new authors and titles participating in publication of botanical nomenclature.

Conclusions: The e-publication changes introduced in the Melbourne Code have gained acceptance, and botanists are using these new techniques to describe and publish their work. They have not, however, accelerated the rate of plant species description or participation in biodiversity discovery as was hoped.

Keywords: Publishing, On-line, Botany, Nomenclature, Taxonomy

Background

Publication of results is one of the cornerstones of the scientific endeavour. Differences between scientific and general publishing were first articulated by Henry Oldenburg, who as Secretary of the Royal Society, established the first English-language scientific journal, Philosophical Transactions of the Royal Society [1]. Oldenberg's functions for scientific publication were dissemination, registration, certification and archiving (called by him the "Minutes of Science"); scientific publishing therefore has a role in informing not only in the present, but also for future generations. Scientific (scholarly) publication has seen great change driven in part by increased interconnectivity of research communities, massive increases in funding for research and development since the middle of the 20th century, and key technological advances such as the Internet. These drivers are characterised as having as big an effect as the replacement of parchment by paper, or the advent of mass printing technologies [2]. Moves away from print on paper to electronic-only publishing parallel increasing scientific activity on-line, and the pace of change in this area of scientific publishing is increasing, with more and more journals converting to on-line only publishing (e.g., Evolution, New Phytologist, Biological Journal of the Linnean Society).

"Published work" has a central place in nomenclature (the scientific naming of organisms), and until January 2012, nomenclatural acts published in electronic-only

* Correspondence: s.knapp@nhm.ac.uk
[4]Department of Life Sciences, Natural History Museum, Cromwell Road, London SW7 5BD, UK
Full list of author information is available at the end of the article

form were not considered valid/effective (see [1, 3, 4]) leading many to consider the taxonomic community as distinctly behind the curve relative to the rest of the scientific community. Discussions about publication went on in both the zoological and botanical (those working on algae, fungi and plants) communities, but largely separately, since the two rulebooks for naming (Codes of nomenclature) are governed very differently (see [5] for a history of the Codes), although many of the central issues were the same for both. Here we treat only e-publication as it pertains to algae, fungi and plants, whose nomenclature rules are contained in the current *International Code of Nomenclature for algae, fungi and plants* [6], hereafter referred to as the ICN or Melbourne Code. Decisions about changes to the rules of naming for this community are made at Nomenclature Sections of International Botanical Congresses (IBC) held every six years [7, 8].

Discussions about electronic publication in the botanical community began in the 1990s, formal proposals at the 1999 XVI IBC in St Louis [9] and at the XVII IBC in Vienna [10] to allow e-publication were defeated, but suggestions about e-publication were included in the Vienna Code [11, 12]. Issues arising were largely those of archiving, accessibility and tracking dates of publication; this last is critical, because the principle of priority that is one of the pillars of nomenclature depends upon accurate knowledge of date of publication (see [8] for an explanation of the principle of priority). A Special Committee was established at the Vienna Congress to examine the issues, with the mandate to prepare proposals for the next IBC in Melbourne in 2011 [13]. Over the six years between the XVII (Vienna) and XVIII (Melbourne) Congresses, publication rules were challenged by Knapp [14], who published new species in PLoS ONE - an on-line only journal - and complied with letter of the Code by depositing ten offprints in botanical libraries [15].

> 30A.2. To aid availability through time and place, authors publishing nomenclatural novelties should give preference to periodicals that regularly publish taxonomic articles, or else printed copies of a publication (even if also distributed electronically) should be deposited in at least ten, but preferably more, botanical or other generally accessible libraries throughout the world including a name-indexing centre appropriate to the taxonomic group. [11].

This posed a significant cataloguing and preservation challenge for libraries, who felt they might be facing a deluge of single- or few-page paper copies of papers describing new species [15] (also see Doug Holland conference presentation: "Libraries and the Code: The changing role of botanical libraries in the age of electronic publication.", Biodiversity Information Standards (TDWG) 2011). Proposals put forward by the Special Committee on Electronic Publication [16] to allow e-publication under the then "botanical code" were accepted overwhelmingly at the Nomenclature Section of the XVIII IBC in Melbourne Australia in July 2011 [17, 18]. At the same time, proposals to change the rules that required a description or diagnosis of new taxa to be in Latin were also accepted [18, 19]. Changes to the rules of naming usually come into force two years after the IBC, but such was the excitement of many in the community for change that these two major changes were voted to come into force in January 2012, a year earlier than "normal" [20]. About six months later, the zoological Commissioners voted to accept e-publication of new names and nomenclatural acts for animals [4], and backdated their new rule to January 2012 to harmonise dates. One major difference in the implementation of e-publication in the two communities is that in zoology, e-publication must be accompanied by registration in ZooBank (www.zoobank.org; for description of ZooBank see [21], while for algae, fungi and plants, e-publication is only another publication type and is not necessarily linked to registration. Fungal names, however, must be registered to be validly published [6].

The advent of e-publication for nomenclatural acts for algae, fungi and plants was both welcomed and feared (see Table 2). Tracking the realization of these effects is difficult for many groups of organisms, but with vascular plants, we have a unique opportunity to conduct an analysis using data from the International Plant Names Index (IPNI, www.ipni.org) which records new names and combinations (generic reassignments, see [8]) for these taxa. IPNI began as Index Kewensis, which was started with a £250 legacy from Charles Darwin in his will for the "establishment of an index of all plants" [22, 23]. It was conceived in a time when it was feasible for a scientist to own all the relevant literature for their field, but it was even then necessary to have a bibliographic index to avoid repeated reference to scattered primary sources. The Index captured the name, authorship and basic bibliographic details of published plant names. In 1983 the data were digitised to an electronic database format, and in the late 1990s Index Kewensis was amalgamated with the Gray Card Index (GCI) maintained by the Harvard University Herbaria and the Australian Plant Names Index (APNI) to form the International Plant Names Index (IPNI, www.ipni.org see [22]). This dataset is accessible online and is continuously updated by a dedicated editorial team as new names are published; approximately 8000 new name records are added each year. The dataset is a valuable resource for trends analysis regarding the time, location and method of publication of new plant names.

In this paper, we analyse publication trajectories for nomenclature governed by the ICN [6] using data from IPNI to examine whether the hopes-increased participation,

increased rate of description-or fears-avalanche of sloppy nomenclature, proliferation of new on-line journals - have been realised. It is not our intention to review the debates on e-publication in taxonomy here, nor are we comparing the effects of the changes in the rules between zoology and botany (algae, fungi and plants). Problems with the new rules have been highlighted by some [24, 25], and within the community working with algae, fungi and plants, new changes to improve the rules surrounding e-publication continue to be proposed [26]. These will be discussed at the Nomenclature Section of the XIX IBC in Shenzhen, China in July 2017 (Shenzen XIX IBC).

Methods

Where the data are from and how they were recorded

The IPNI database contains basic bibliographic information about the place of first publication of vascular plant names (ferns and fern allies, conifers, cycads and flowering plants). Nomenclatural acts representing new names, new combinations, replacement names and names at new ranks are recorded, with the date of effective publication. Note that lectotypifications are also nomenclatural acts which may be published electronically, but these are not included in this analysis. See [6] and [8] for definitions of nomenclatural acts. This dataset does not include nomenclatural acts in algae or fungi, also governed by the same rules as vascular plants.

The authorship of the nomenclatural act is standardised using the principles laid out in Authors of Plant Names [27] (also referenced under recommendation 46A of the ICN [6]). Publication titles are also standardised by linking to an authoritative list. The set of data recorded for each nomenclatural act has been expanded since the changes in the Melbourne Code came into effect (1st January 2012), to include:

- publication channel: to indicate if the work was published on paper or as an e-publication. This is set to e-publication if the article containing the nomenclatural act is either published online before print or is published online only. The default value of the flag indicates paper publication.
- language: indicates that the description or diagnosis is written in English. The default value of the flag indicates use of Latin language.
- Digital Object Identifiers (DOIs) - these can be resolved to access metadata about the publication and to navigate to the reference online (if available).

Members of the editorial team apply the rules of the ICN and exercise nomenclatural judgement about the nomenclatural acts recorded. Annotations to indicate if an act is illegitimate, not effectively published, or not validly published are added to that nomenclatural act record.

The records are fully versioned, with date of application of each edit recorded.

Selection of data subset for analysis

All data recorded with publication years between 2009 and 2014 (inclusive) were analysed. The years 2009 - 2011 represent the three years before the changes in the ICN agreed at the Melbourne Congress, which came into effect on 1st January 2012; 2012–2014 represent the three years after. Although we conducted our analysis in 2017, the most recent data included in the dataset are two years old - this was to ensure that more obscure titles have had a chance to be seen by the IPNI editorial team. The lag time for some types of publications (e.g., small print-run journals and some books) can be up to a year or more. Three years after the implementation of the ICN change date gives us a valuable range of samples, because at least some of the work published in 2012 would have been already in the publication system and thus done using the previous rules; thus authors would have been unable to fully take advantage of the changes in the ICN which came into effect on 1st January of that year.

Emergence trends for authors and publications were created, in order to see if more people were participating in the publication of new vascular plant names, and if the range of places available in which to publish have expanded. To get a better view of underlying trends, a longer timescale was chosen for this part of the analysis - the full decade between 2005 and 2014 (inclusive).

We recognised that taxon-specific communities of botanists may exist, such as those working in plant families with considerable horticultural interest. To assess the degree to which these communities were using e-publication in different ways, we drilled down into the flowering plant data to collate information on the rate of take-up of e-publication in particular families. For this analysis we compared three families with considerable horticultural and collector interest – Orchidaceae (orchids), Cactaceae (cacti) and Bromeliaceae (airplants and pineapples) – with three families that are of less horticultural interest – Fabaceae (beans), Solanaceae (nightshades) and Cyperaceae (sedges). We performed the same analyses on these smaller datasets as were done for the whole dataset (see above).

Preparation

The nomenclatural acts recorded in the IPNI database were classified into three broad groupings:

tax. nov. (names of new taxa) - *tax. nov*
comb. nov. (new combinations) - *comb. nov*, *stat. nov*, *comb. et stat. nov.*
nom. nov. (replacement names and names at new rank) - *nom. nov.*, *nom. et stat. nov.*

SQL queries were executed against the underlying IPNI database on 2017-04-28, these were scripted in the Python programming language.

Analyses

Volume of nomenclatural acts: numbers of nomenclatural acts were grouped by publication year and citation type (to distinguish names of new taxa, new combinations, and replacement names and names at new rank.

Use of publication channel for all nomenclatural acts: numbers of all nomenclatural acts were grouped by year and then by publication channel. This analysis was repeated on a per-family basis for a selected number of families, some with horticulural interest.

Use of any ICN changes (e-publication channel, English language diagnosis) for acts representing new names: numbers of nomenclatural acts representing new names (*tax. nov.*) were grouped by year, and then by language and publication channel combined.

Authors - number active: the unique number of authors specified as members of the publishing author team in nomenclatural acts between 2005 and 2014 were counted, broken down by year.

Authors - number emergent: for all the authors active in the selected period (2005-2014), their date of emergence was calculated - this is the date when they were first recorded as a member of the publishing team of a nomenclatural act. This dataset was grouped by year of emergence to give a count for each year.

Publications - number active: as per the analysis for authors described above, this is the unique number of serial publications recorded as containing nomenclatural acts published between 2005 and 2014 were counted, broken down by year. A serial publication is defined as a multi-volume work.

Publications - number emergent: (as per the analysis for emergent authors described above) - for all serial publications active in the selected period, their date of emergence was calculated - this is when they were first recorded as containing a nomenclatural act. This dataset was grouped by year of emergence to give a count for each year of the study.

Results

Volume of nomenclatural acts

The volume of nomenclatural acts – excluding lectotypifications - has remained relatively constant (Table 1; data shown graphically in Additional file 1: Figure S1). In fact, the number of new taxa described per year (ca. 3000) has remained relatively constant since its recovery from a dip due to the Second World War (unpublished data) more than 50 years ago.

Table 1 Numbers of nomenclatural acts recorded in IPNI

publication_year	tax. nov.	comb. nov.	nom. nov.
2009	3022	2548	195
2010	2759	2533	166
2011	2754	3155	198
2012	2960	3611	195
2013	2806	2677	133
2014	2804	2580	384

Use of publication channel, and description or diagnosis language

The use of e-publication has increased steadily from its introduction in 2012; the most recent year of the study (2014) shows that almost half (48.3%) of all acts recorded were using e-publication (Fig. 1a), data in Additional file 1: Table S1). The use of e-publication was consistent when data were analysed on a per-family basis (Additional file 1: Figure S2, data in Additional file 1: Table S1a).

When looking at only the publication of new names for taxa, the previous status quo (a description or diagnosis formed in Latin, contained within a work published on paper) is steadily diminishing, with almost three quarters (74.7%) of the new taxonomic descriptions in the final year of the study utilising at least one of the major ICN changes introduced (Fig. 1b), data in Additional file 1: Table S2). It appears that those using e-publication also more often use a diagnosis or description in English rather than Latin (Fig. 1b).

Emergence of new authors and serial titles

The data show no sudden difference in the emergence or participation of either authors or serials after the starting date for e-publication in 2012 (indicated by the vertical line in the plot) (data in Additional file 1: Table S3). The apparent dramatic dip in the last year of the sample is likely due to the lag in discovery of nomenclatural acts published in less accessible media (e.g., small print-run local journals or books).

Discussion

Underlying many of the hopes regarding e-publication is a recognition of a potential opportunity to overcome the so-called "taxonomic impediment" (as defined by the Convention on Biological Diversity, https://www.cbd.int/gti/problem.shtml), and to increase the communities working in taxonomy through adoption of e-dimensions to their work [28]. This means that if e-publication were a significant part of these impediments, we would expect

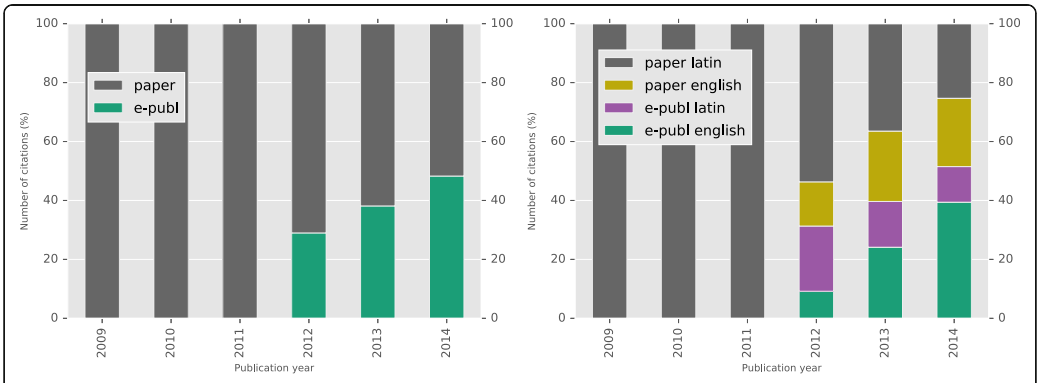


Fig. 1 Publication channel use (all nomenclatural acts) and take-up of Code changes (plotted for tax. nov. acts only)

to see more species being described, by more people, more quickly.

We structure our discussion around the principal hopes and fears regarding e-publication that were expressed during the discussions surrounding the acceptance of the changes to the ICN introduced at the XVIII IBC in Melbourne (Table 2) and which have continued to be discussed elsewhere.

Taking each of these hopes in turn:

- **Rapidity**: e-publication has not had an effect on the speed of publication of new names, and thus the rapidity of biodiversity description. The same numbers of plant species are being described every year as were before the change in the ICN (Table 1). This is likely to be the result of a number of factors, including the speed of peer-review, and the increasing numbers of specimens available for examination before decisions about the novelty of taxa can be taken. It is also abundantly evident that taxonomists now do many more things than describe and publish new taxa [29].
- **Accessibility**: e-publication as permitted in the ICN does not necessarily imply accessibility via Open Access publication. An amendment to the ICN proposed, but defeated [18] in Melbourne was to require OA publishing for all nomenclatural acts, and considerable discussion is going on in the zoological community suggesting this should be a requirement. The cost of OA publishing, however, is seen by many as restricting participation by those for the developing world, despite initiatives set up to alleviate this [1]. It is clear that accessibility needs to be associated with physical or virtual access to the work rather than any costs which may be associated with access. Accessibility is an issue for *all* types of publications, electronic and print-on-paper.
- **Inclusivity**: Our data do not show any upward trends in the numbers of authors actively publishing nomenclatural acts, nor in the number of people involved in the authorship of botanical nomenclature. Neither have we seen a decrease in either of these measures. This short term trend is only a snapshot of the longer term trend seen (for a smaller plant related dataset) by previous authors [29]. Anecdotally more authors appear to be associated with plant names, but further analysis of these trends is required. Biographical data on the authors of nomenclatural acts is not routinely collected by IPNI, and new efforts will be needed to ascertain if the community is truly changing.

Table 2 Hopes and fears regarding e-publication, expressed in the discussions at the XVIII IBC, Melbourne 2011

Hopes		Fears	
Rapidity	speed up publication process; biodiversity description becomes faster	Avalanche of sloppy nomenclature, leading to bad taxonomy	many new journals, little quality control
Accessibility	increase connectivity worldwide	Accessibility	lack of connectivity in the developing world; potential disenfranchisement
Inclusivity	more people in involved in description of biodiversity	Date of publication	difficulties in applying the principle of priority
Modernity	part of normal publication; improve the visibility and opinion of taxonomy	Archiving	lack of permanency; ephemeral nature of the electronic environment

- **Modernity**: This is difficult to assess with the data we have assembled. The results of our analyses show no perceptible change in numbers of publications pre and post the permitted use of e-publication, suggesting that the changes are seen as part of the normal publication process. Also discussed under modernity was the wish to improve the visibility and opinion of taxonomy - electronic publication and the use of English rather than Latin for the diagnosis or descriptions, is on the rise (see Fig. 1), and the influence of the ICN on nomenclatural practice in general can be seen in the choice of the ICZN to back-date e-publication to match the ICN starting date (1 Jan 2012) [4].

If the hopes have not been fully realised, what of the fears? The fears expressed about the acceptance of e-publication are underpinned by concerns about the potential for fragmentation of the community - either geographically or through time - and a consequent lessening in taxonomic quality. Our analysis of families with considerable horticultural interest versus those without such associated communities showed no difference in the rates of use of e-publication (see Additional file 1), so we suggest that fears regarding a fragmentation of the community based upon specialisation currently seem unrealised.

- **Sloppy nomenclature**: We have not seen an avalanche of nomenclatural activity creating "bad taxonomy" since the acceptance of e-publication. The numbers of journals continuing to be active in the process of publishing botanical nomenclature has remained more or less constant (Fig. 2) and there has not been a dramatic upsurge in the establishment of new journals. Acts of "sloppy nomenclature", such as publication that is not effective or not valid under the ICN, have also not increased since the advent of e-publication (data not shown), but longer term trends are needed.
- **Accessibility**: Our data cannot address the fear that e-publications will potentially be less accessible to the wider botanical community than print-on-paper publications. Anecdotally, however, it seems that the move in the publishing world from printed copies to electronic-only publication of journals has not limited access to the scientific literature. The ubiquity of internet connectivity seems only to be on the rise. We are currently assembling data to examine this aspect of publishing botanical nomenclature. Issues regarding accessibility to the literature containing nomenclatural acts will be better addressed in a separate analysis, which is more focussed on the literature itself (rather than the abstracted subset available here).
- **Date of publication**: The principle of priority is dependent upon the retrieval of an effective date of publication for any nomenclatural act, be it a new species, new combination, name at new rank or lectotypification. Over the course of the 3 years of data we assembled, there have been a handful of cases requiring investigation, these are not common and in fact are no different than any other nomenclatural problem needing investigation to resolve. It is, however, an issue that many journals still do not place the date of effective publication in the required PDF of the publication, but instead place it elsewhere, for example in the table of contents for the journal. It is imperative that botanists work with publishers to ensure that

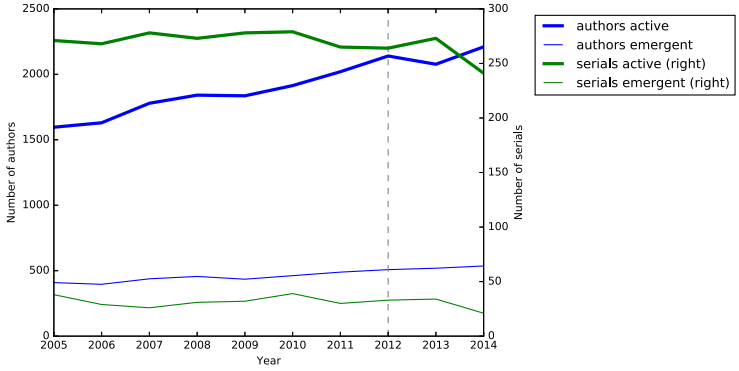


Fig. 2 Number of authors & serials active & emergent/year (2005–2014)

e-publications best serve future generations of botanists – this will be an on-going conversation [20] and we have really only just started.

- **Archiving**: Access to past literature is fundamental for systematics, as it is for all of science. At the current point in time - just five years after the first e-published works - we are too early to fully assess issues regarding archival storage and long term accessibility. Because archiving of works is not part of the requirements for effective publication under the ICN (for either print or e-publications), resolution of this concern does not lie in the rules of the ICN, but rather, in continued dialogue with publishers of works that contain botanical nomenclature.

Conclusions

Our analysis shows that in the time frame we have analysed, three years after the implementation date for e-publication, nomenclature as applied to vascular plants continues to be in a steady state - both in terms of the number and quality of nomenclatural acts recorded, and the participation of those doing the science that results in these acts. We can therefore conclude that one of the more modest hopes - that e-publication is seen as part of the normal publication process - has been realised. In fact, we did not need to run for our lives [15]: the issues imagined have not flooded us with problems different to those perennially associated with nomenclature.

The result that the acceptance of e-publication has not elevated the rate of species description nor increased the numbers of people involved in naming new taxa means that as a community, botanists must consider other ways to speed up taxonomy. Some issues that have been raised include the large numbers of specimens now available for examination before a decision can be reached about the novelty of a taxon, the advent of a perception that molecular as well as morphological data are necessary for making a taxonomic decision, and the rigour of the peer-review process that accompanies modern publication. It still takes as long to make a decision about the identity of a specimen as it always has done, and if a botanist has 3000 specimens to look at it necessarily will take longer. Human resource issues are likely to be crucial for increasing the rate of taxonomy; our efforts perhaps should be focusing on this rather than on technological quick fixes.

It is clear that much discussion remains to be had with the publishers of nomenclature about some of the issues that have arisen, such as display of the date of effective publication, access and archiving. The results of this analysis of one part of the names governed by the *International Code of Nomenclature for algae, fungi and plants* shows that treating e-publication as an instance of publication, rather than something special to be regulated differently has been a good decision that still has the potential to help the botanical community both publish and access work describing life on Earth.

Additional file

Additional file 1: Supplementary information contains plotted figures for data presented as tables in the text and original data tables for figures in the text. In addition, all data and plots for the per-family analysis are presented here. **Figure S1.** – Volume of nomenclatural acts by type. **Table S1.** – Data for use of publication channel (all nomenclature acts). **Figure S2.** - Use of publication channel for all nomenclatural acts - per-family breakdown. **Table S1a.** – Data for use of publication channel for all nomenclatural acts - per-family breakdown. **Table S2.** – Data for use of any Melbourne Code changes (e-publication channel, English language diagnosis) (tax. nov. acts only). **Table S3.** – Data for authors and publications – numbers active and emergent. (DOCX 76 kb)

Acknowledgements
The authors acknowledge the work of the editorial team past and present who manage and contribute to the IPNI database: Christine Barker, Irina Belyaeva, Rosemary Davies, Kanchi Gandhi, Rafael Govaerts, Helen Hartley and Heather Lindon.

Funding
All authors undertook the work as part of their core tasks at their institutions; this study was not externally funded.

Availability of data and materials
The datasets supporting the conclusions of this article are included within the article and its additional files. Please see separate document (Additional file 1) https://doi.org/10.6084/m9.figshare.c.3780956.

Authors' contributions
NN and SK conceived the study. KC gathered the data. NN extracted and analysed the data. NN and SK wrote the manuscript, with intellectual input and assistance from AT. All authors read and approved the manuscript.

Competing interests
The authors declare that they have no competing interests.

Consent for publication
Not applicable.

Ethics approval and consent to participate
Not applicable.

Publisher's Note

Author details
[1]Biodiversity Informatics & Spatial Analysis, Royal Botanic Gardens, Kew, Richmond Surrey TW9 3AA, UK. [2]IPNI, Royal Botanic Gardens, Kew, Richmond Surrey TW9 3AA, UK. [3]Department of Computer Science, Brunel University London, Uxbridge UB8 3PH, UK. [4]Department of Life Sciences, Natural History Museum, Cromwell Road, London SW7 5BD, UK.

Received: 24 March 2017 Accepted: 9 May 2017
Published online: 25 May 2017

References

1. Knapp S, Wright D. E-Publish or Perish. In: Polaszek A, editor. Systema naturae 250-the linnaean ark. London: Taylor & Francis; 2010. p. 83–93.
2. Guédon JC. In Oldenburg's long shadow: Librarians, research scientists, publishers, and the control ofscientific publishing. Washington: Association

of Research Libr; 2001.http://www.arl.org/storage/documents/publications/in-oldenburgs-long-shadow.pdf.
3. International Commission on Zoological Nomenclature. Proposed amendment of articles 8, 9, 10, 21 and 78 of the International Code of Zoological Nomenclature to expand and refine methods of publication. Bull Zool Nomencl. 2008;65:265–75. doi:10.21805/bzn.v65i4.a9.
4. International Commission on Zoological Nomenclature. Amendment of Articles 8, 9, 10, 21 and 78 of the International Code of Zoological Nomenclature to expand and refine methods of publication. ZooKeys. 2012;219:1–10. doi:10.3897/zookeys.219.3944.
5. Knapp S, Lamas G, Lughadha EN, Novarino G. Stability or stasis in the names of organisms: the evolving codes of nomenclature. Philos Trans R Soc B. 2004;359:611–22.
6. McNeill J, Barrie F, Buck W, Demoulin V, Greuter W, Hawksworth D, et al. International code of nomenclature for algae, fungi, and plants (Melbourne Code). Königstein: Koeltz Scientific Books; 2012.
7. Brummitt RK. The democratic processes of botanical nomenclature. In: Leadlay E, Jury SL, editors. Taxonomy and plant conservation: the cornerstone of the conservation and the sustainable use of plants. Cambridge: Cambridge University Press; 2006. p. 101–29.
8. Turland N. The code decoded a user's guide to the International Code of Nomenclature for algae, fungi, and plants. Königstein: Koeltz Scientific Books; 2013.
9. Zander RH, Wilson KL. (10–13) four proposals to amend the code, and report of the Special Committee on Electronic Publishing and Databasing. Taxon. 1998;47:175–7. doi:10.2307/1224041.
10. Zander RH. (180–181) report of the Special Committee on Electronic Publishing with two proposals to amend the Code. Taxon. 2004;53:592–4.
11. McNeill J, Burdet F, Demoulin V, Hawksworth D, Marhold K, Nicolson D, et al. International Code of Botanical Nomenclature:(Vienna Code). 2006.
12. Knapp S, Wilson K, Watson M. Electronic publication. Taxon. 2006;55:2–3.
13. Chapman AD, Turland NJ, Watson MF. Report of the Special Committee on Electronic Publication. Taxon. 2010;59:1853–62.
14. Knapp S. Four new vining species of solanum (Dulcamaroid Clade) from montane habitats in Tropical America. PLoS ONE. 2010;5:e10502. doi:10.1371/journal.pone.0010502.
15. Knapp S, Paton A, Challis K, Nicolson N. "Run for your lives! End of the World!" – Electronic publication of new plant names. Taxon. 2010;59:1009–10.
16. Special Committee on Electronic Publication. (203–213) proposals to permit electronic publications to be effectively published under specified conditions. Taxon. 2010;59:1907–8.
17. Cressey D. Botanists shred paperwork in taxonomy reforms. Nat News. 2011. doi:10.1038/news.2011.428.
18. Flann C, Turland N, Monro AM. Report on botanical nomenclature—Melbourne 2011. XVIII International Botanical Congress, Melbourne: Nomenclature Section, 18–22 July 2011. PhytoKeys. 2014;41:1–289. doi:10.3897/phytokeys.41.8398.
19. Figueiredo E, Moore G, Smith GF. Latin diagnosis: time to let go. Taxon. 2010;59:617–20.
20. Knapp S, McNeill J, Turland NJ. Changes to publication requirements made at the XVIII International Botanical Congress in Melbourne - what does e-publication mean for you? BMC Evol Biol. 2011;11:250. doi:10.1186/1471-2148-11-250.
21. Pyle RL, Michel E. ZooBank: developing a nomenclatural tool for unifying 250 years of biological information. Zootaxa. 1950;2008:39–50.
22. Croft J, Cross N, Hinchcliffe S, Lughadha EN, Stevens PF, West JG, et al. Plant names for the 21st century: the International Plant Names Index, a distributed data source of general accessibility. Taxon. 1999;48:317–24. doi:10.2307/1224436.
23. Lughadha EN. Towards a working list of all known plant species. Philos Trans R Soc Lond B Biol Sci. 2004;359:681–7. doi:10.1098/rstb.2003.1446.
24. Shipunov A. We need at least two baskets for our eggs: PDF alone is not enough for e-publication. Taxon. 2014;63:134–5.
25. Dubois A, Crochet P-A, Dickinson EC, Nemésio A, Aescht E, Bauer AM, et al. Nomenclatural and taxonomic problems related to the electronic publication of new nomina and nomenclatural acts in zoology, with brief comments on optical discs and on the situation in botany. Zootaxa. 2013;3735:1–4. doi:10.11646/zootaxa.3735.1.1.
26. Turland NJ, Wiersema JH. Synopsis of proposals on nomenclature - Shenzhen 2017: a review of the proposals concerning the International Code of Nomenclature for algae, fungi, and plants submitted to the XIX International Botanical Congress. Taxon. 2017;66:217–74. http://www.ingentaconnect.com/contentone/iapt/tax/2017/00000066/00000001/art00037.
27. Brummitt RK, Powell CE. Authors of plant names: a list of authors of scientific names of plants, with recommended standard forms of their names, including abbreviations. London: Royal Botanic Gardens, Kew; 1992.
28. Scoble MJ. Networks and their role in e-taxonomy. In: The New Taxonomy. CRC Press; 2008. p. 19–31. doi:10.1201/9781420008562.ch2.
29. Joppa LN, Roberts DL, Pimm SL. The population ecology and social behaviour of taxonomists. Trends Ecol Evol. 2011;26:551–3. doi:10.1016/j.tree.2011.07.010.

B.2 Identifying novel features from specimen data for the prediction of valuable collection trips

This paper is an early output from work further developed in chapters 4 and 5.

Identifying novel features from specimen data for the prediction of valuable collection trips

Nicky Nicolson[1,2] and Allan Tucker[2]

[1] Biodiversity Informatics and Spatial Analysis, Royal Botanic Gardens, Kew, UK, n.nicolson@kew.org
[2] Department of Computer Science, Brunel University London, UK

Abstract. Primary biodiversity data provide "what, where, and when" data points: the assertion that a species occurred at a particular point in space and time. These are most valuable when associated with specimens stored in natural history museums and herbaria, which evidence the assertions with reference to a physical specimen. The research presented uses novel data-mining techniques to uncover two hidden dimensions in specimen data - *who* collected the specimens and *how* they were collected. A combination of unsupervised and supervised learning techniques are used, which establish two new entities: *collector* and *collection trip*. Features are defined against these higher order representations of the data, which support the use of the data to answer novel questions such as *which collection trips discover the most new species?*. We explore the features by building classifiers to predict species discovery, and compare these with a baseline model grouped using collector team transcriptions derived from the raw specimen data. Preliminary results are promising and whilst the particular focus of this research was botanical specimens, the technique is equally applicable to datasets of field-collected specimens from other scientific domains. . . .

Keywords: Data-mining, Clustering, Classification, Species discovery

1 Introduction

Biological specimens collected over hundreds of years and held in natural history museums and herbaria are a rich reference source with which to understand the natural world, and to analyse its changes over time. Estimates of the total number of specimens vary between 2.5-3 billion specimens globally [1]. Only a small percentage have associated digital data. Aggregation initiatives such as the Global Biodiversity Informatics Facility (GBIF) harvest and mobilise digital specimen data: at the time of writing (May 2017) the GBIF data portal includes information on 129,006,858 specimens. In order to aid the mobilization of the data, there has been an effort to develop standards regarding the representation of the data [2], and references to it [3]. These standards are important as due to the scale of the overall task, data have been digitised in a distributed fashion, at different rates and to different levels of completeness.

In addition to the structured data held on the specimens themselves, field collected specimens are often accompanied by a wealth of information about the collection site, habitat and associated species, logged in collectors notebooks, which are also being digitised via literature digitisation initiatives.
Although plants are a comparatively well known group, and are well represented with digitised specimen data, species discovery is not yet complete, and approximately two thousand new species are described per year [4]. Not all species discovery is via field work: a significant proportion of species discovery is conducted from pre-existing specimens already lodged in institutional collections [5]. Estimates of the total number of plant species recognise the importance of species discovery from pre-existing collections and the use of collections data to plan species discovery in the field.
The application of intelligent data analysis techniques on the specimen data can help meet two key aims: *data mobilisation* by better utilising and curating the existing data, and finding efficiencies that will help the digitisation process, and *data understanding* by uncovering patterns that will help plan future scientific effort as research is conducted with specimens or in the field.
The novel data-mining techniques demonstrated here detect new entities (*collector* and *collection trip*) from the duplicated, incomplete and variably transcribed specimen datasets. These are used to draw together heterogonous data, which has been recorded in different places, to different standards, in order to build classification models to support and develop our understanding of a complex system - species discovery.
The remainder of the paper is structured as follows: a background section further introduces the nature of the specimen data available by defining terms and outlining the specimen collection process, methods describes a data-mining process to detect collector and collection trip entities from raw specimen data, defines a novel set of features using these new entities to group the raw specimen data, and describes the creation of classifiers using these and baseline data. Preliminary results of the data-mining and classification steps are shown, and ideas for further work are discussed.

2 Background & definition of terms

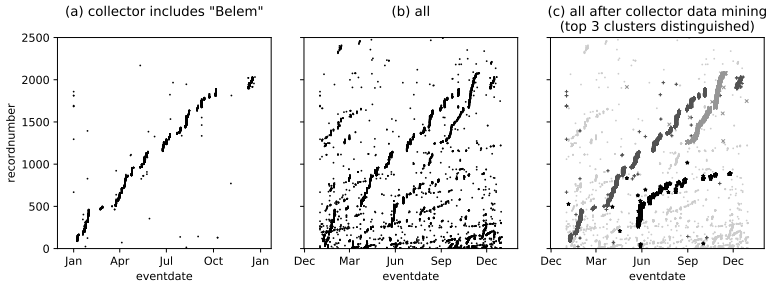
A *specimen* is a physical sample of biological material collected in the field. In botany, a collected sample may consist of multiple specimens, named *duplicates*. The *collecting team* is the team of collectors responsible for gathering and documenting the specimen, this may include multiple collectors, referred to by personal name. The *primary collector* is the first listed member of the *collecting team*, and controls the *recordnumber* - a number given to the specimen in the field, usually sequential and unique to the primary collector. Recordnumbers are locally managed, rather than centrally assigned. When duplicate specimens are collected, they are given the same recordnumber [6]. A *collection trip* is a circumscribed period of specimen collecting activity conducted by a particular primary collector, focussed on a particular place and time. An *itinerary* is a list of the

collecting localities visited by a primary collector in a collection trip, which may be documented in a *collectors fieldbook*, cross referenced to specimens via the recordnumber.
An *institution* is the holder of specimens for long term storage and reference consultation, usually natural history museums or herbaria (botanically focussed institutions). Institutions may distribute duplicate specimens to external institutions, to form a globally distributed reference collection. *Digitisation* is the process of creating electronic records from the data held on the physical specimen, which may include *imaging* - the creation of a digital image of the specimen, and / or *georeferencing* - the process of determining a latitude / longitude pair from a textual description of the collecting locality. This is necessary due to the historic nature of the specimen collection effort, which pre-dates technologies such as hand-held global positioning systems. Duplicates are recognised as a source of data to speed the digitisation process [7]. The *collector name transcription* is the transcription of the collector names made when specimen data is read for digitisation. A single collector may have multiple varying collector name transcriptions, depending on the standards used in the different institutions, transcription errors and spelling mistakes. As the collector name transcription is necessary to identify duplicates [7], variability in this data element impedes efficient use of the global specimen dataset. *Aggregation* is the collation of digitised specimen records from many institutions into a single data repository, represented using a structured *data standard*.
A *type specimen* is the reference use of a specimen as the basis of a new species description, published in the academic literature. The reference to a specimen is made using the collector name and recordnumber [8]. The use of a specimen as a type specimen is indicated in its digital record. A *name author* is the author of a new species description, a person who may also act as a collector. A *career* is the complete body of work performed by one collector / name author. The subject focus of the career may be examined to determine if the person is a *specialist* (focussed on a particular taxonomic subset) or *generalist* (working across many different areas of taxonomy). Some generalists may be regional specialists, focussing on the plants of a particular geographical region.

Primary biodiversity data derived from specimens have many applications in research [1] including species description and discovery. The collector - who makes decisions in preparatory planning and in the field about what to collect - is obviously a major contributor to species discovery [9], and the collection trip has been recognised as a way to understand the accumulation of knowledge regarding the species found in a particular geographic area [10].
Differentiating collection trips based on the characteristics of the collector has also been proposed [10], including the differentiation between specialist and generalist collectors. Despite the scope for more advanced analyses of specimen data when differentiated and grouped by collector and / or collection trip, these entities are not formally managed - only the collecting team is a component of the main data standard used to share specimen data, which is supplied as a text transcription [2] [11]. Studies involving the grouping of specimen data by

collector and or collection trip have had to use manual specimen record allocation [9] and / or expert knowledge [10], which limits scope. This means that the sequential nature of collectors recordnumbers has been minimally exploited to date - but it has been used to create collectors itineraries, by cross-referencing by hand between specimen data and collectors field-books [12] as an aid to georeferencing.

Fig. 1. Example specimen data from 1965, recordnumber less than 2500



An example use of sequential recordnumber for a single collector is a test for a positive correlation - as a particular collector moves forward through time, their own personal sequential recordnumber increases (see figure 1(a)). Exploration of the data in this way can be useful to identify outliers (resulting from data transcription errors), but applications are limited due to the difficulty in initially identifying the set of specimens relating to a single primary collector, due to the variation in collector name transcriptions. Plotting a fuller corpus of specimen data (see figure 1(b) - a sample of points from specimens collected in a single year), shows some visually distinguishable elongated "clusters", each of which correspond to the set of specimens collected by a particular primary collector and labelled with their own sequential recordnumber, which ascends over time.

In this research, we propose the exploitation of the sequential recordnumber as a feature for clustering to detect the primary collector, thereby overcoming the variability encountered when using the un-standardised transcription of personal names. We employ a novel combination of data-mining techniques to detect these clusters, in order to identify higher order abstractions (collector and collection trip) from an incomplete raw specimen dataset. These abstractions are recognised in the domain, but are absent from digital datasets. We exploit the sequential nature of the collectors recordnumber to cluster specimens as they were gathered over time, resulting in a grouping by primary collector. We then use the collector grouping to detect the collection trips made by that primary collector. These abstractions are used to group the data and to define features at

the grouped level, these features are used to train classifiers to identify high-value collection trips which are relevant to species discovery.

3 Methods

The process described here was developed to allow visualisation of intermediate results at each stage, in order to allow an analyst to influence the design of the process. Visualisations were created as interactive scatter plots (as shown in figure 1), allowing the analyst to focus on particular areas of the data, and to examine the underlying specimen record.

The main specimen dataset was downloaded from the Global Biodiversity Informatics Facility, encompassing data generated from botanical specimens [11]. This large (59 million record) dataset was used for exploratory data analysis, and a subset representing the specimens collected since 1700 from a single political country (Brazil) was selected for data-mining. Brazil is recognised as a mega-diverse country [13], and Brazilian specimen data has been digitised and repatriated via the REFLORA project [14], meaning that a considerable amount of data is digitally available, from many different institutions. The subset of data used for data-mining contains 3493107 specimen records, which were collected between 1705 and 2016, and held in 132 different institutions. A biographical dataset was used as a data lookup, to check if collectors detected in the data-mining steps are also known to have authored new species. This is managed as part of the International Plant Names Index (IPNI), and contains the personal names and lifespan dates for those who have published new names since the start date for botanical nomenclature (1753) [4].

3.1 Data-mining

Preparation: data are read from the data store and prepared for data-mining by making a `numeric feature-set` from `recordnumber` and `eventdate` (expressed as days since 1st January 1970). The details of the collector team transcription are quantified by extracting the primary collector name, standardising the order of recording of the name elements and deriving minimal textual features from the name to form a `lexical feature-set`. The first initial, and the first uppercase character, first lowercase character and last lowercase character from the first word (usually the surname) are extracted and converted to indicator variables where the value `1000` represents presence and `0` represents absence. A field for type status (`is_type`) is created and populated following the criteria used in [9].

The actual data mining process is composed of 4 steps, steps 1-3 identify collectors, step 4 examines the set of specimen data allocated to a particular collector to detect collection trips.

Step 1 (cluster) uses DBSCAN [15]. This clustering algorithm is used as via exploratory data analysis the data are observed to form elongated rather than spherical clusters, due to the use of the sequential `recordnumber` and `eventdate` features (along with the `lexical feature-set`). DBSCAN is configured to use

a value of `300` for `epsilon` and `2` for `min_samples`. A low value of `min_samples` is used as the clustering results are computationally post-processed to lexically examine the collector names included within a cluster. Analyst examination of the data immediately after DBSCAN clustering shows that the primary collector names are so variably recorded that clusters contain multiple logical collector names. A pessimistic approach is taken, clusters are divided into multiple separate clusters if the lexical variation of the primary collector names included is too great (e.g. due to differing initials).

Expert analysis of the dataset after step one identified a common problem with many clusters, that the huge variation in the transcriptions of the primary collector names introduces variation into the lexical feature-set, and results in the assignment of logical collectors into separate clusters. When the data are examined using visualisation (an interactive scatter plot of `eventdate` against `recordnumber`, with the colour of the points determined by the `cluster_id`, these clusters show up as *interpolations*: an elongated stream of points flips between two or more different clusters, but the specimen data underlying is seen to have the same primary collector (transcribed in very different ways).

Step 2 (classify) uses a decision tree to detect sets of distinct clusters which are similar in terms of the numeric feature-set, but which differ in terms of the lexical feature-set (described as *interpolated* above). The classifier is trained on the numeric feature-set to predict the cluster identifier. Commonly confused classes are identified using the classifier, and these are considered candidates for joining after computational assessment for lexical similarity with respect to their primary collector names. Those with very similar names (as may result from differing transcriptions e.g. abbreviation to initials) are joined. As cluster manipulation will affect the extent of cluster interpolation, this is an iterative process, and is run for `10` iterations or until there are no more candidates for joining, whichever occurs first.

Step 3 (join) joins clusters to result in a grouping representing the career work of a single collector, so that all specimens collected by the same collector will be held in the same cluster. This is implemented in two stages: (i) the clusters output from step 3 are joined if their most frequently occurring first collector name is shared and all name variants in the cluster agree lexically and (ii) clusters are matched against an external bibliographic database of taxonomic name authors, those matching to the same bibliographic database record are joined. A unique identifier value for each collector is created (`collector_id`).

Step 4 (detect collection trips) subdivides the dataset by `collector_id`. The specimen data for each `collector_id` is passed into a DBSCAN clustering using minimal features - `eventdate` and `recordnumber`. A lower value of epsilon is used (`90`, in comparison to `300` used in the collector data-mining in step 2). The minimal value for `min_samples` (`2`) is retained in recognition of the gaps in the incomplete specimen dataset - a cluster of two specimen data points may indicate a trip which collected many specimens, only two of which are currently digitised. The clusters identified by each iteration of the DBSCAN process rep-

resent the collection trips for a single collector, a unique trip identifier is created and applied to the specimen dataset.

3.2 Data abstraction: creation of new features

The results of the data mining are used to group the specimen data and to define metrics using these groupings. The identification of a collector allows the detection and elimination of specimen duplicates. Duplicates are defined as those specimens which share `recordnumber` and `collector_id`, all but the first occurrence of a particular `recordnumber` / `collector_id` pairing are flagged as duplicates and excluded from subsequent analyses.

Grouping the specimen data using the data-mined entity types (collector and collection trip) allows the definition of a new set of features. These are categorised as *temporal* (the start year of the grouping), the *scale* of the grouping (duration, the total number of specimens included and the range of recordnumbers allocated), the *rate* of accumulation (the slope of a line of best fit through the `eventdate` and `recordnumber` values), and the `correlation_score` of these points). The *character* of the grouping is defined in two ways - by creating a `specialist` flag, set if the grouping is more 60% composed of specimens from a single taxonomic family, and by creating a `nomenclaturalist` flag which is set if the collector is known to have also acted as a name author. The *experience* of the collector at a point in time is assessed by creating features for the total number of previous specimens collected and the total number of previous collection trips made. Finally, a feature is defined that will later be used as the class variable in classifiers: this encodes the *species discovery value* of the grouping, and is simply a flag indicating if the grouping contains material that was later used as a type specimen, representing a contribution towards species discovery. Data files containing these features are constructed, these are used as training data in the next step.

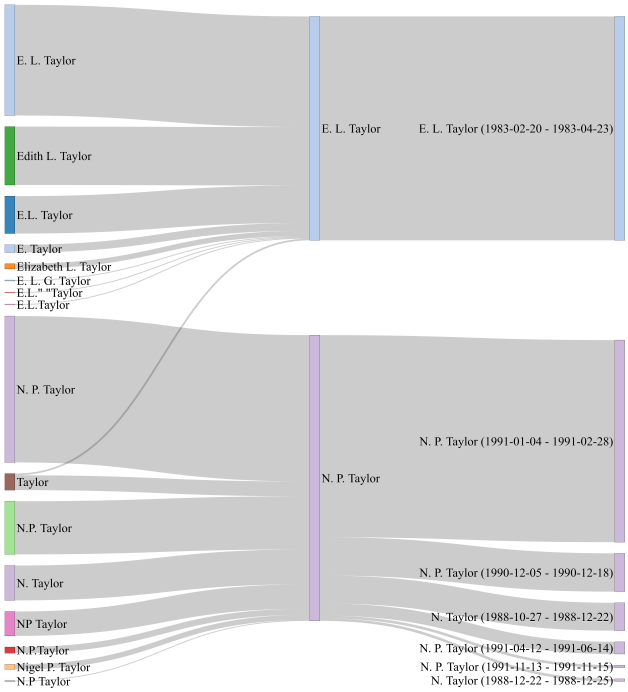
3.3 Development of a classification model using the results of data-mining & data abstraction

The feature-sets generated by the data-mining process are used to train classifiers to predict the species discovery value of the grouping. These are compared to a baseline grouping, derived without the data-mining process. The baseline is simply a grouping of the specimen data by the different values of the transcribed primary collector name, which was the source of the lexical feature-set used in the data-mining steps. Both the baseline and the data-mined datasets were down-sampled to balance the binary class variable, as the samples for the positive class were far less frequent. A decision tree classifier was trained on the down-sampled data, using 10-fold stratified cross-validation. Feature selection was also conducted to examine which of the features defined were the most indicative.

4 Results

Data mining results: DBSCAN identified 42096 clusters (step 1a); lexically post-processed to 51192 clusters (step 1b); resolved via decision-tree classifier to 44768 clusters (step 2); joined to 19706 clusters representing collector entities (step 3). 79012 different collecting trips were identified (step 4). The raw specimen data underlying the entities recognised via data mining comprises 131582 unique collector team transcriptions and 41511 unique primary collector name transcriptions. 1127 (5.7%) of collectors and 3412 (4.3%) of trips collected specimens later labelled as type specimens.

Fig. 2. Specimen grouping by (l-r) baseline, collector & collection trip (collection trips shown with start & end dates)



The results of the collector data-mining process on the illustrative sample used in the background section are shown in figure 1(c). An alternative visualisation of the data, using textual rather than numeric attributes, demonstrates the grouping provided by the data-mining process, and the level of variation in

the input data. A Sankey diagram represents flow. It is used here (see figure 2) to demonstrate how specimen records "flow" between the groupings established at each stage of the data mining process. The diagram compares three groupings - on the left the data are grouped using the baseline method (the text transcription of the primary collector name extracted from the collector team transcription). The central column shows the grouping as done by the data-mined collector entities, and the rightmost column divides the specimen data still further, into collection trips (shown with start and end dates). The width of the connections between the groups in each of the columns is proportional to the number of specimens included. The diagram shows that the variation in the dataset is reduced as a result of the data-mining, and that specimens can be more meaningfully grouped by collector and / or collection trip.
The selected subset shows both the strengths and the weaknesses of the current data-mining technique - a strength is that the ambiguous name *Taylor* is split into two different collectors based on the context provided by the date and recordnumber values; a weakness is the over-enthusiastic grouping of what seem to be two distinct collectors (*Edith L Taylor* and *Elizabeth L Taylor*) in the top-most collector. The trip detection results should be considered preliminary - due to an incomplete dataset and immature trip data-mining process, many very small trip groupings are defined.

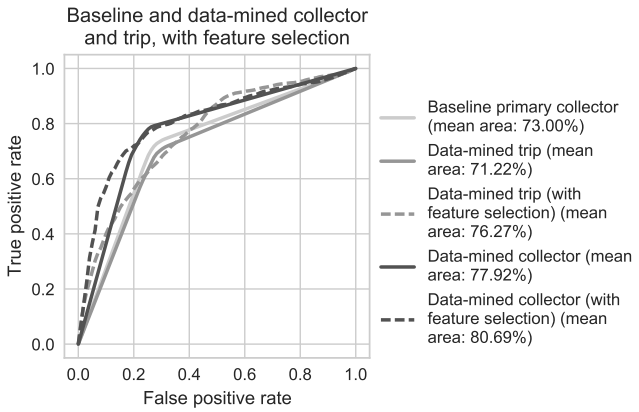
Classification & feature selection results: these were assessed by calculating the mean area under the receiver operator curve from the 10-fold cross-validated runs. Classification results (see figure 3) from the baseline and trip datasets are similar (73.00% and 71.22%), collector shows an improvement over these (77.92%), and the execution of feature selection on the trip and collector datasets improves each over the comprehensive feature-set (76.27% and 80.69% respectively).

Collector appears to perform better than trip; the similarity of the performance of the trip groupings to the baseline is likely to be due to the numbers of small trips detected in the data-mining process. The datasets derived from data-mining were used to conduct feature selection using an exhaustive search strategy, scoring using the area under the receiver operator curve (ROC AUC) metric. The features selected via this process were the temporal feature (start_year) and the two features encoding character (specialist and nomenclaturalist) for both the collector and trip groupings.

5 Discussion

When analysing the results of the data-mining via a simple classification task, collector has shown to perform well, and trip has similar results to the baseline. Re-examination of the methods shows that collector data-mining has a set of post-processing steps which validate and modify the entities, trip data-mining by contrast is rather immature and is perhaps more badly affected by the incomplete input data: when trying to subdivide data into trips, a time gap can be either due to a legitimate trip boundary or an artefact due to incomplete input data.

Fig. 3. Receiver-operator curves for classifiers trained on baseline & data-mined datasets



Further work with the trip data-mining process to define a similar set of post-processing steps to be applied after DBSCAN clustering would likely improve this part of the method.

The immediate context for this research is data generated from field-collected scientific specimens. The particular focus of this research was botanical specimens, however the technique is equally applicable to datasets of field-collected specimens from other scientific domains. It is possible to define simple eligibility criteria for this kind of analysis: the specimen dataset must contain a string representation of the primary collector (or collecting team), a date of collection and a field-assigned recordnumber. Applying these eligibility criteria to the datasets available via GBIF shows that datasets comprising specimens from *icthyology*, *ornithology* and *mycology* meet the criteria for this kind of analysis.

The data-mining process created in this research can be generalized to the use of a *product* (specimen) dataset, to identify the *agent* responsible for its generation (collector), and to place the product within a *sequence of work* (a collection trip). This is possible as the product is identified by an *agent-managed sequence* (recordnumber) which ascends over time. We scanned the biodiversity (and related) domains for other examples of data generated via a similar process. Many of the datasets generated in the digital age have recognised the need for shared persistent identifiers across distributed datasets (e.g. the use of DOIs in publishing) and by implementing these have sidestepped the need for this kind of analysis. The examples selected represent data generation via digitisation of historic information, that which pre-dates easily accessible shared identifiers. Species names for plants are referenced using micro-citations, page level bibliographic references. There is an effort to standardise the authorship for these to

enable trends analysis e.g. to detect changes in gender balance of the authors of plant names [16]. Page level microcitations can be seen as another representation of the *product / agent / sequence of work* data generation process: the product is a page-level microcitation, generated by an author, fitting into a bibliographic container (article or book) as a sequence of work. As page number is sequential and pages located in close proximity are likely to be authored by the same person, this dataset is a candidate for data-mining using this technique.
Feature selection on the grouped datasets result in the inclusion of the two character features - which indicate if a grouping is specialist (taxonomically focussed) and if the collector was also a nomenclaturalist (participating in the publication of new species) - these results support previous work which propose the specialist / generalist distinction as relevant [10]. Future work will implement the selective inclusion of features using the categorisations defined earlier, for example to define classification models applicable across temporal scales or to see how relevant different feature categories remain over time. There is scope for further data integration to expand the set of features used here by including data from bibliographic sources. An expanded feature-set should allow further advances towards understanding the process of species discovery and the people involved in it.

6 Conclusions

This paper proposes the application of data-mining techniques to specimen data in order to create higher-order data abstractions. These abstractions were used to define new features, which were tested by building classifiers to predict species discovery value. The input data are acknowledged as being incomplete - both in terms of the number of records available and the population of individual fields in a particular record. Specimen data are expensive to fully digitise, one of the aims of this research was to understand what could be done with a minimal dataset, with no dependence on expensive augmentation processes such as georeferencing.
The positive preliminary results shown here have impacts in two core areas - data mobilization and data understanding. In data mobilization we are able to suggest practical modifications to increase efficiency of the specimen digitisation process. Recognition of the data-mined collector and trip entities allows better integration of data from different sources, one element of the data-mining process - classification to predict collector from easily digitised features of specimen (`eventdate`, `recordnumber`) - has potential application as a tool to aid the transcription of specimen data by digitisation staff. We have advanced data understanding by demonstrating the potential for reshaping specimen data to define novel features. These were used to populate models which can detect subsets of particular value in species discovery.
This work has demonstrated that although incomplete and variably recorded, the aggregated specimen data now form a critical mass which support the de-

velopment and application of alternative approaches towards data mobilization and data understanding.

References

1. Chapman, D., A.: Uses of Primary Species-Occurrence Data. Global Biodiversity Information Facility, Copenhagen (2005)
2. Wieczorek, J., Bloom, D., Guralnick, R., Blum, S., Döring, M., Giovanni, R., Robertson, T., Vieglais, D.: Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. PLoS ONE **7**(1) (January 2012) e29715
3. Güntsch, A., Hyam, R., Hagedorn, G., Chagnoux, S., Röpert, D., Casino, A., Droege, G., Glöckler, F., Gödderz, K., Groom, Q., Hoffmann, J., Holleman, A., Kempa, M., Koivula, H., Marhold, K., Nicolson, N., Smith, V.S., Triebel, D.: Actionable, long-term stable and semantic web compatible identifiers for access to biological collection objects. Database **2017**(1) (January 2017)
4. ipni.org: International Plant Names Index. http://www.ipni.org
5. Bebber, D.P., Carine, M.A., Wood, J.R.I., Wortley, A.H., Harris, D.J., Prance, G.T., Davidse, G., Paige, J., Pennington, T.D., Robson, N.K.B., Scotland, R.W.: Herbaria are a major frontier for species discovery. Proceedings of the National Academy of Science **107** (December 2010) 22169–22171
6. Bridson, D.M...: The Herbarium Handbook. 3rd. ed edn. Kew : Royal Botanic Gardens, 1998 (1998)
7. Tulig, M., Tarnowsky, N., Bevans, M., Kirchgessner, Anthony, Thiers, B.M.: Increasing the efficiency of digitization workflows for herbarium specimens. ZooKeys (209) (July 2012) 103–113
8. Turland, N.: The code decoded a user's guide to the International code of nomenclature for algae, fungi, and plants. Koeltz Scientific Books, Königstein (2013)
9. Bebber, D.P., Carine, M.A., Davidse, G., Harris, D.J., Haston, E.M., Penn, M.G., Cafferty, S., Wood, J.R.I., Scotland, R.W.: Big hitting collectors make massive and disproportionate contribution to the discovery of plant species. Proceedings of the Royal Society of London B: Biological Sciences **279**(1736) (June 2012) 2269–2274
10. Utteridge, T., de Kok, R.: Collecting Strategies for Large and Taxonomically Challenging Taxa. In: Reconstructing the Tree of Life: Taxonomy and Systematics of Species Rich Taxa. CRC Press (2006) 297–304
11. GBIF.org: (4th October 2016) GBIF occurrence download (taxon: Tracheophyta, basis of record: specimen). http://doi.org/10.15468/dl.68z1mf
12. Smith, L., Smith, R.: Itinerary of William John Burchell in Brazil, 1825-1830. Phytologia **14**(8) (1967) 492–505
13. Mittermeier, R.A.: Megadiversity: Earth's biologically wealthiest nations. Agrupacion Sierra Madre (1997)
14. REFLORA: REFLORA programme. http://reflora.jbrj.gov.br
15. Ester, M., Kriegel, H.P., Sander, J., Xu, X., others: A density-based algorithm for discovering clusters in large spatial databases with noise. In: Kdd. Volume 96. (1996) 226–231
16. Lindon, H.L., Gardiner, L.M., Brady, A., Vorontsova, M.S.: Fewer than three percent of land plant species named by women: Author gender over 260 years. Taxon **64**(2) (May 2015) 209–215

B.3 Specimens as research objects: reconciliation across distributed repositories to enable metadata propagation

This paper results from work presented in chapter 6.

doi:10.1109/eScience.2018.00028

Specimens as research objects: reconciliation across distributed repositories to enable metadata propagation.

1st Nicky Nicolson
Royal Botanic Gardens, Kew and *Brunel University, London*
London, UK
n.nicolson@kew.org

2nd Alan Paton
Royal Botanic Gardens, Kew
London, UK
a.paton@kew.org

3rd Sarah Phillips
Royal Botanic Gardens, Kew
London, UK
sarah.phillips@kew.org

4th Allan Tucker
Brunel University, London
London, UK
allan.tucker@brunel.ac.uk

***Abstract*—Botanical specimens are shared as long-term consultable research objects in a global network of specimen repositories. Multiple specimens are generated from a shared field collection event; generated specimens are then managed individually in separate repositories and independently augmented with research and management metadata which could be propagated to their duplicate peers. Establishing a data-derived network for metadata propagation will enable the reconciliation of closely related specimens which are currently dispersed, unconnected and managed independently. Following a data mining exercise applied to an aggregated dataset of 19,827,998 specimen records from 292 separate specimen repositories, 36% or 7,102,710 specimens are assessed to participate in duplication relationships, allowing the propagation of metadata among the participants in these relationships, totalling: 93,044 type citations, 1,121,865 georeferences, 1,097,168 images and 2,191,179 scientific name determinations. The results enable the creation of networks to identify which repositories could work in collaboration. Some classes of annotation (particularly those regarding scientific name determinations) represent units of scientific work: appropriate management of this data would allow the accumulation of scholarly credit to individual researchers: potential further work in this area is discussed.**

***Index Terms*—research objects, data citation, record linkage, annotation**

I. Introduction

Botanical specimens are core research objects in the science of taxonomy (the naming of biological organisms), stored for long term consultation in institutional repositories and referenced in academic works. Worldwide there are 3,001 herbaria (botanical specimen repositories), containing 387,007,790 specimens - representing collections gathered over hundreds of years [1]. Due to their physical characteristics (flattened, dried plant material is typically mounted on a large sheet of paper, stored inside a paper folder) and their management as a long term, consultable record, specimens act as vehicles for the communication of results and theories, as researchers annotate the paper sheet underlying the specimen. Annotations placed on specimen sheets are public and available for use by other researchers, this public yet potentially unpublished status is discussed in [2].

Taxonomic researchers populate institutional repositories by conducting collection events (usually field-based) which generate multiple specimens. Recommended botanical practice is for a single collection event to generate five to six specimens, which will be deliberately distributed to separate institutional repositories. Physical distribution of specimens has three main goals: to maximise access - researchers working on their local flora should be able to consult the relevant specimens in their national herbarium, to provide resilient storage - duplicate specimens insure against disastrous loss of a single repository, and to ensure efficient use of storage space within repositories [3]. Duplicate specimens are also used in genetic analyses: if the samples were collected from separate individuals, the duplicate set can be used to assess genetic diversity across the sampled population. Scientific theories regarding the recognition of species and their interrelationships are developed by researchers as they work with the specimens, which are traditionally accessed either by loan or by visits to institutions; more recently specimen digitisation initiatives have enabled online access to specimen metadata records and high quality images, this simplifies search and retrieval of specimens and associated metadata, and allows some level of specimen examination to be conducted remotely. Independent creation and management of metadata for specimen duplicates can be inefficient (metadata creation is repeated unnecessarily), and inadvertently misleading (metadata diverges between different members of a specimen duplicate group).

One particular class of research annotation is the application of a scientific name to the specimen: this may be an existing name, or the researcher may recognise that the specimen represents a new species. Species description in plants is ongoing with circa two thousand new plant species described each year [4]. When a new species is described, one specimen is chosen as a physical representation of the otherwise abstract scientific name. Specimens which formally represent a scientific name are called type specimens; the selection of these is called type citation. When a specimen is cited as a type, all peers ("duplicates") which are generated from the same collection event - but which may be stored and managed remotely, in

separate repositories - are also considered to have type status. New scientific names are created via a formal publication process governed by the International Code of Nomenclature for algae, fungi and plants [5]. The majority of new species are discovered from historic specimens already lodged within specimen repositories, being formally described years after collection [6]. The use of duplicate specimens as vehicles for the communication of results is illustrated by the historic use of "exsiccatae". These are uniform specimen sets with information displayed on printed labels distributed to multiple herbaria, and until 1953 were considered a valid publication mechanism for new scientific names [7] [5].

Taxonomists consider type specimens to be the most valuable specimens in a repository, and management reporting often includes both the total number of specimens held and the number of type specimens. The first major digitisation effort in botany (JSTOR Global Plants Initiative) focussed on the digitisation of type specimens across more than 300 institutions in over 70 countries [8]. In addition to reporting on the total numbers of specimens and types housed in an institutional repository [9], managers are also interested in the numbers of new type citations published each year as a metric of on-going research use of their specimens [10]. Some natural history institutions have experimented with bibliometrics to quantify use of their specimens in a publication context [11].

In addition to their core use in the science of taxonomy, specimens provide physical "what, where, when" evidence and are used for a wide range of scientific applications such as species distribution modelling [12]. Specimen exchange networks have also been used for historical social network analysis [13]. These applications are generally dependent on aggregations of specimen metadata mapped to a common data standard and sourced from many different institutional repositories.

Problem statement Despite the widespread recognition that botanical specimens form a global collection, there is currently no flow of data from the point of creation (via the field collection event) to the generated specimens wherever they may be located for long term storage. Despite advances in the mobilisation and standardized representation of specimen metadata across the different specimen repositories, duplicate specimens have so far gone undetected, with metadata records for duplicates appearing unlinked in aggregated datasets. The main data elements needed to assess specimens as potentially arising from a shared collection event - collector name, along with the collector's recordnumber and eventdate - are not formally managed. These missing links mean that valuable research annotations and type citations are not easily shared between repositories, and impacts all downstream users of specimen data: taxonomic researchers working with individual specimens are unable to benefit from knowledge added elsewhere, leading to misinterpretation due to inaccurate and/or out of date naming, and users working with large aggregations of specimen data can find that specimen number estimates are overstated, as their datasets contain hidden duplicates.

The research described in this paper applies machine learning to a set of aggregated specimen metadata to identify and reconcile the collectors responsible for the creation of specimens, enabling the detection and linkage of specimen duplicates generated from the field work of the identified collectors. In contrast to existing work on annotation propagation - which has focussed on potential changes in working practices and tools and techniques to enable and incentivize this [14] [15] - this work applies these techniques to a dataset of existing digitally available specimen data in order to calculate the numbers of existing metadata elements and annotations which may be propagated between separate institutional repositories.

The remainder of this paper is structured as follows: a background section further introduces the problem domain with an explanation of the specimen life cycle and the kinds of annotations applied at each stage, and worked examples of distributed specimen sets whose members are independently managed at different institutions. Materials and methods describes the application of a machine learning process to a dataset of specimen data from the Global Biodiversity Information Facility to identify specimen duplicates. Criteria for the identification and assessment of duplicate sets are proposed. The resulting specimen duplicate analysis is used to answer the following questions:

1. How many distributed, independently managed specimens can be reconciled across separate institutional repositories and linked as generated products of a common collection event?
2. How many metadata elements and research annotations can be propagated between institutional specimen repositories?
3. Can specimen duplicate linkages be used to infer network relationships between institutional repositories, which institutions are most frequently linked and do subcommunities or cliques exist in the inferred network?

Preliminary results are presented and ideas for expansion and future work are proposed.

II. Background

This section outlines the stages in the specimen life cycle, and indicates relevant projects at each stage.

Collection and storage: these activities represent standard practice across the specimen repositories

- **Collection**: material is gathered from the field and details of the collection locality (associated species, geology, habitat etc) are recorded in the collectors field notebook. The collectors recordnumber provides the cross-reference between the data recorded in the field notebook and the physical material collected, this is usually a sequential number managed individually by the collector.
- **Accessioning**: material is received by a specimen repository and prepared for long term storage, including mounting on a sheet of paper (for dried specimens).

Digitisation: due to the number of specimens held in the global collection, digitisation is incomplete, and is progressing through a variety of cross-cutting institutional, regional,

international and thematic projects. The JSTOR Global Plants Initiative selected a particular class of specimens for digitisation (type specimens) across 300 institutions [8], other projects have been set up to digitise all specimens gathered from a particular country to enable data repatriation, as in the Brazilian REFLORA programme [16] and to digitise specimens held within a particular country as in the US National Science Foundation funded Advancing Digitisation of Biocollections programme [17]. These latter projects show a trend of government funding for digitisation, recognising that these are part of the national scientific infrastructure [9] [18].

- **Databasing**: details of the specimen (metadata) are added to an institutional data repository.
- **Aggregation**: databased records can be mapped to a data standard (e.g. Darwin Core [19]) and shared with aggregation projects. The Global Biodiversity Information Facility is an intergovernmental organisation that aggregates specimen-derived species occurrence records (alongside records from observations) to facilitate scientific research, iDigBio is a US based aggregator which focusses only on specimen derived data.
- **Georeferencing**: the metadata record in the institutional repository can have latitude and longitude added (this may be a costly step for historic records where the original collection locality is only a textual description of the place). Economies of scale are possible if records can be ordered so that similar places are georeferenced together [20] [21].
- **Imaging**: the specimen is imaged and a reference to the image is added to the metadata. If the specimen metadata is shared with an aggregator the digital image may also be mobilised.

Depending on their range of holdings, some institutions are involved in multiple digitisation projects, others not at all. With technical advances in digitisation and the setup of high-throughput imaging facilities, some of these steps may be performed out of sequence - if the digitisation project is of a sufficient scale, it may be cost effective to rapidly image the specimens first and perform the metadata capture later, from a high quality digital image [22] [23] [24].

Use as a research object: these steps outline the use of the specimen as a taxonomic research object. The use of specimens as a data source for computational applications such as species modelling is covered in the digitisation steps above, digitisation steps also facilitate discovery and access of specimens for taxonomic research. Annotation mobilisation work has focussed on tooling for the collection and propagation of newly generated annotations, including the projects AnnoSys [14] and Filtered Push [15]. There has also been an effort to standardise the citation of specimens so that different repositories use a common HTTP URI based naming convention by which their digital metadata records can be accessed [25]. By convention, the citation of specimen records irrespective of digitisation status is made by stating the collector name, number and date, along with the herbarium code [1] in which the physical specimen may be found. These kinds of references can be found throughout the botanical literature, and examples are shown in the worked examples in the next section.

- **Determination**: the specimen is labelled with a scientific name, the date and the name of the researcher who made the determination are also added.
- **Citation**: the specimen is cited in a published academic work (e.g. to evidence the presence of a species in a geographic region).
- **Type citation**: the specimen is referenced as a type specimen in a published academic work to create a new species name.

The long term creation of a global network of specimen repositories, the more recent efforts to enable virtual access to specimens and their metadata, and the practice of sharing research annotations all fit well with the FAIR principles for scientific data management [26]: ensuring that the metadata and specimens on which scientific analyses are based are Findable, Accessible, Interoperable and Retrievable.

A. Worked examples

This section is intended to illustrate the problem statement - that specimen duplicates are (1) widely present in distributed specimen repositories, (2) unidentified in data aggregations built by combining specimen datasets and (3) that specimen metadata attached to derived specimens generated from a single source can diverge due to separate and independent data curation practices. Two examples have been selected, representing the two extremes of species description citing botanical specimens: species discovery in-field formalised by rapid publication just one year after collection, and species discovery in-repository with formalised description decades after field collection. A considerable proportion of new species are described from material already collected and stored in specimen repositories [6]. The second example shows a species description occurring 46 years after the field collection of the plant material on which is it based.

For each example we will assemble a dataset of potential specimens, which is constructed as the superset of the specimens referenced in the literature (which may or may not be digitised) and the relevant specimen records found in digital form in a data aggregator. We then examine the metadata attached to the specimens, showing where this has diverged due to independent management. These are shown in table I.

1) Example 1: Rapid publication of species discovered in-field: See table I, example 1. (Table data source: gbif.org)

The publication data (displayed below) shows that there are at least 9 specimen duplicates, stored in different institutional repositories, indicated by the capitalised alphabetic herbarium codes (WTU, BH etc [1]). The exclamation mark (!) after a code is a convention to indicate that the author has actually seen the specimen. In this case the author is also the collector of the specimen, so all are listed as having been seen.

> ***Sedum citrinum*** Zika, *sp. nov.* **Type**:—UNITED STATES. California: Del Norte County, ridge 1.4 air

TABLE I
WORKED EXAMPLES

recordedBy	recordNumber	eventDate	scientificName	institutionCode	referenced in publication	digitised	typestatus	georeferenced	imaged
P. F. Zika	26185	2013-06-09	Sedum citrinum Zika	BH	✓	-	-	-	-
Zika, Peter F.	26185	2013-06-09	Sedum citrinum Zika	CAS	✓	✓	✓	✓	-
Peter F. Zika	26185	2013-06-09	Sedum citrinum Zika	CAS-BOT-BC	-	✓	-	-	-
P. F. Zika	26185	2013-06-09	Sedum citrinum Zika	CHSC	-	✓	-	✓	-
P. F. Zika	26185	2013-06-09	Sedum citrinum Zika	GH	✓	-	-	-	-
Zika, P.F.	26185	2013-06-09	Sedum citrinum Zika	K	-	✓	✓	-	✓
P. F. Zika	26185	2013-06-09	Sedum citrinum Zika	MO	✓	-	-	-	-
P. F. Zika	26185	2013-06-09	Sedum citrinum Zika	NY	-	✓	✓	✓	✓
P. F. Zika	26185	2013-06-09	Sedum citrinum Zika	OSC	✓	-	-	-	-
Peter F. Zika	26185	2013-06-09	Sedum citrinum Zika	RSA	✓	✓	-	✓	-
Peter F. Zika	26185	2013-06-09	Sedum citrinum Zika	UC	✓	✓	-	✓	-
P. F. Zika	26185	2013-06-09	Sedum citrinum Zika	US	✓	✓	✓	-	✓
P. F. Zika	26185	2013-06-09	Sedum citrinum Zika	WTU	✓	-	-	-	-
P. C. Hutchison & J. K. Wright	5738	1964-06-19	Solanum sanchez-vegae S.Knapp	F	✓	✓	✓	-	✓
P. C. Hutchison & J. K. Wright	5738	1964-06-19	Solanum aligerum Schltdl.	F	-	✓	-	-	-
Hutchison, P.C.	5738	1964-06-19	Solanum sanchez-vegae S.Knapp	K	✓	✓	✓	✓	✓
Paul C. Hutchison—J. Kenneth Wright	Hutchison 5738	1964-06-19	Solanum cutervanum Zahlbr.	MO	-	✓	-	-	-
P. C. Hutchison	5738	1964-06-19	Solanum sanchez-vegae S.Knapp	NY	-	✓	✓	✓	✓
P. C. Hutchison	5738	1964-06-19	Solanum sanchez-vegae S.Knapp	NY	-	✓	✓	✓	✓
P.C. Hutchison & J.K. Wright	5738	1964-06-19	Solanum sanchez-vegae S.Knapp	P	✓	-	-	-	-
P. C. Hutchison & J. K. Wright	5738	1964-06-19	Solanum sanchez-vegae S.Knapp	US	✓	✓	✓	-	✓
P.C. Hutchison & J.K. Wright	5738	1964-06-19	Solanum sanchez-vegae S.Knapp	USM	✓	-	-	-	-

km north of South Red Mountain, 1050 m, 9 June 2013, *P. F. Zika 26185* (holotype: WTU!; isotypes: BH!, CAS!, GH!, MO!, OSC!, RSA!, UC!, US!). [27]

There are 8 digitally available records for this set of specimens, drawn from 8 separate institutional specimen repositories. These are independently managed and not interlinked. Despite being generated from the same collection event, the specimen metadata show variation due to isolated management in separate repositories: 5 of the 8 are georeferenced, 4 of the 8 specify a type status and 3 of the 8 have an associated image. We can therefore calculate that the group contains propagable annotations for georeferences, typestatus and image (i.e. that for each annotation class, the group contains records with and without the annotation set, meaning that the annotation could be propagated from the specimens with the annotation to their peers without it). Of the digitised specimens in the group: 3 could receive a georeference, 4 could receive a type status annotation and 5 could be linked to an associated image. The creation of a specimen group could also make the initial creation of the specimen records for the currently undigitised members more efficient, by using existing data as a starting point rather than independently re-creating it.

2) Example 2: Species discovery in-repository: See table I, example 2. (Table data source: gbif.org)

The publication data (displayed below) shows that there are at least 6 specimen duplicates, stored in 5 different institutional repositories. The author has supplied a numeric identifier for some of the specimens (shown in square brackets), to help the reader locate the relevant records in specimen repository and / or its associated metadata catalogue(s).

> ***Solanum sanchez-vegae*** S.Knapp, *sp. nov.* [urn:lsid:ipni.org:names:77103635-1] **Type**: Peru. Amazonas: Prov. Chachapoyas, W side of Cerros Calla-Calla, 45 km above Balsas, mid-way on road to Leimebamba, 3100 m, 19 Jun 1964, *P.C. Hutchison & J.K. Wright 5738* (holotype, USM; isotypes, F [F-163831], K [K000545365], P [P00549320], US [US-246605], USM). [28]

There are 7 digitally available records for this set of specimens, from 5 separate institutional specimen repositories. These are independently managed and not interlinked. As per the first example, despite being generated from the same collection event, the specimen metadata show variation due to isolated management in separate repositories, with all annotation categories holding inconsistent information: 3 of the 7 are georeferenced, 5 of the 7 specify a type status, 5 of the 7 have an associated image and 2 of the 7 have an outdated scientific name. We can therefore calculate that of the 7 digitised specimens in the group: 4 could receive a georeference, 2 could receive a type status annotation and 2 could be linked to an associated image.

These two different examples both show that the separate specimen records held in different specimen repositories hold divergent metadata, and that there is the potential for metadata propagation between members of a specimen group. Specimen groups can be identified by grouping on the collector, their field-assigned record number and the eventdate, but this is non-trivial due to the variation in the recording style of the collecting team (shown in the recordedBy column), as duplicate records have been independently digitised to different data standards in different institutions and projects.

III. MATERIALS AND METHODS

A. Data

A dataset of specimen data relating to vascular plants (those with specialised tissues for the transport of water, encompassing ferns and allied groups, and all seed plants) was downloaded from GBIF [29] in Darwin Core [19] archive format. This was input into a data mining process based on

the clustering technique DBSCAN in order to detect collector entities [30]. Specimen records are eligible for data mining if they have a numeric component in their *recordnumber* (the sequential number managed by an individual collector and assigned to field collection events), a precise date recorded to the level of day (*eventdate*), and a collector name (*recordedby*). The data mining process augments the specimen dataset with a numeric identifier for the primary collector of the specimen represented in the metadata record. This allows data to be grouped as the product of the work of a particular collector, irrespective of the lexical variation in the transcription of the collectors names.

B. Detection of duplicate groups

A group of specimens are asserted to be generated from a single collection event if they share the same collector identifier (the results of the collector data mining exercise), eventdate (when the field collection event was carried out) and collector-assigned record number. The record number has any alphabetic prefixes stripped from the value - this normalises values which are sometimes presented with the surname of the collector in the recordnumber field (see the worked example in I).

See procedure listing detectDuplicateGroups. The input into this algorithm is a tabular data structure where each row represents a specimen, with fields for collector_id, eventdate and recordnumber.

C. Establishing a confidence measure

A confidence measure is applied to candidate duplicate groups by examining the range of variation in fields within the duplicate group. Three assessments are made, a spatial assessment using the countrycode field (duplicate specimen records originating from the same collection event should logically be located in the same country) and two taxonomic assessments using the order and family fields. Biological taxonomy uses a hierarchical system, where species are arranged into families, and families into orders. Although a specimen may be re-determined (have different scientific names applied to it) during its lifetime in a specimen repository, it is less likely to be re-determined across higher taxonomic boundaries. These flags detect variation in these higher-level categories within a duplicate group.

Three Boolean flags were created (one for each assessment field), these were set to True if all members of the candidate duplicate group share the same value of the assessment field. All possible combinations of these three flags were used to assess the duplicate groups. Only duplicate groups meeting the most conservative assessment criteria (where all of the assessment flags are True, indicating no variation in these fields within the duplicate group) were carried forward for use in subsequent analyses.

See procedure listing assessDuplicateGroups. The input into this algorithm is a tabular data structure where each row represents a specimen, with fields for duplicate_group_id, countrycode, order and family. This is the labelled output from the preceding algorithm detectDuplicateGroups.

D. Assessing annotation status per specimen and detecting groups with uneven annotation statuses

Boolean flags were created to indicate if the specimen is georeferenced, if the specimen has an associated image, and if the specimen is recorded as having type status. *Typestatus* values were used as described in [31].

For each annotation examined, two new Boolean fields were created on the aggregated dataset - these are set to True if *all* specimens in the duplicate group have the annotation set and if *any* specimens in the duplicate group have the annotation set. A group is said to have propagable annotations if it has any and not all annotations set for the specimens within the group. Two count fields were also created for each annotation, these were set to hold the number of specimens within the group with and without the annotation set. The number of specimens which could receive propagable annotations was determined by totalling the number of specimens within groups with propagable annotations which did not themselves have the annotation set.

See procedure listing findPropagableAnnotations. The input into this algorithm is a tabular data structure where each row represents a specimen, with a field for duplicate_group_id and a set of Boolean fields to indicate the presence of annotations on the specimen (georeference, typestatus, image). This is the assessed, labelled output from the preceding algorithm assessDuplicateGroups.

E. Repository relationship analysis

The sharing of specimens in a duplicate group implies a relationship between the two (or more) institutional repositories participating in the group. In this analysis, the data are reshaped to build a graph data structure where nodes are institutional repositories and links are created between a pair of nodes if the corresponding repositories share specimens in a duplicate group. The links are weighted by the number of groups shared. The resulting data structure is a weighted, undirected graph. This inferred network data structure is visualised in Gephi [32], using an OpenOrd [33] layout following modularity analysis [34] for community detection.

IV. Results

A. Data mining

The initial dataset downloaded from GBIF contained 63,492,620 records, of these 19,827,998 records were eligible to be input into the data mining process to detect the collector. The data mining process resulted in 19,489,798 specimen records being labelled with an identifier for the collector.

B. Duplicate identification and assessment

Of the 19,489,798 data mined records, 7,347,705 records participate in a duplicate relationship, forming 2,914,181 duplicate groups. All combinations of assessment flags with associated group and record counts are depicted in figure 1.

Procedure detectDuplicateGroups(Specimens)

Input: Specimens
Output: LabelledSpecimens
let S be Specimens, the set of specimens to be grouped
let $DuplicateGroups$ be S grouped by s.collector_id, s.eventdate, s.recordnumber
Apply an identifier to each group
for $i \leftarrow 1$ **to** $|DG|$ **do**
  $dg \leftarrow DG[i]$
  for s *in* dg **do**
    $s.duplicate\_group\_id \leftarrow i$
    LabelledSpecimens.append(s)
  end
end

Procedure assessDuplicateGroups(LabelledSpecimens)

Input: LabelledSpecimens
Output: AssessedLabelledSpecimens
let $DuplicateGroups$ be LabelledSpecimens grouped by duplicate_group_id
for dg *in* $DuplicateGroups$ **do**
  for $assessment\_field$ *in* $\{countrycode, order, family\}$ **do**
    Create a new boolean field [assessement]_conservative, which is set to *True*
    if all members of the duplicate group share a single value for this field
    $assessment\_values \leftarrow []$
    for s *in* dg **do**
      assessment_values.append(s[assessment_field])
    end
    $dg[assessment\_conservative] \leftarrow |assessment\_values| == 1$
    Copy the assessment flag down to specimen level
    for s *in* dg **do**
      $s.assessment\_conservative \leftarrow dg.assessment\_conservative$
      AssessedLabelledSpecimens.append(s)
    end
  end
end

Procedure findPropagableAnnotations(assessedLabelledSpecimens)

Input: AssessedLabelledSpecimens
Output: AssessedLabelledCountedSpecimens
let $DuplicateGroups$ be AssessedLabelledSpecimens grouped by duplicate_group_id
for dg *in* $DuplicateGroups$ **do**
  let s be the set of specimens included in *dg*
  Annotation fields are Boolean flags indicating if the specimen has this annotation set
  for $annotation\_field$ *in* $\{georef, typestatus, image\}$ **do**
    $dg[annotation\_propagable] \leftarrow any(s.annotation\_field)$ *and not* $all(s.annotation\_field)$
    Copy the propagable flag down to specimen level
    for s *in* dg **do**
      $s.annotation\_propagable \leftarrow dg.annotation\_propagable$
      AssessedLabelledCountedSpecimens.append(s)
    end
  end
end

Fig. 1. Assessment flag combination counts

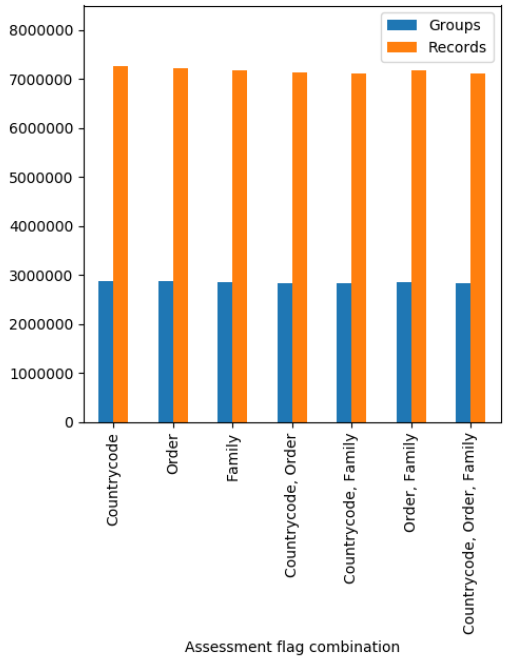
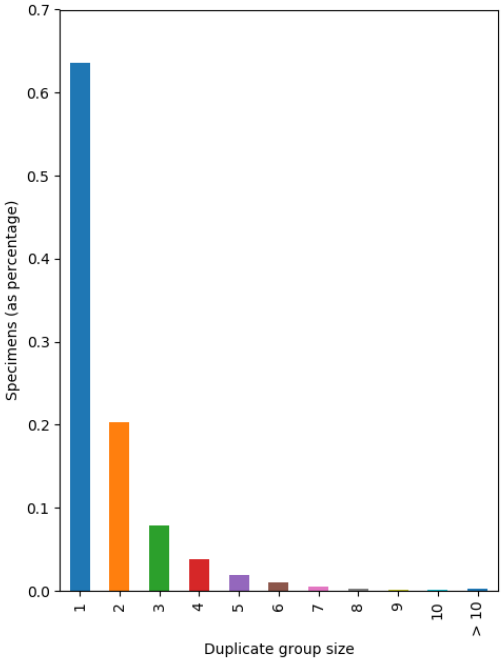


Fig. 2. Duplicate group sizes



Only the subset of duplicate groups meeting the most conservative assessment criteria were used in subsequent analyses: 7,102,710 specimens in 2,823,651 groups. The sizes of the conservatively assessed duplicate groups are shown in figure 2.

C. Propagation of annotations

Members of duplicate sets are located at different institutional repositories and therefore may have been curated differently. Reconciliation of duplicate sets allows the propagation of several classes of annotations - georeferences, type citations, specimen images and determinations - between holders. Of the conservatively assessed duplicate sets:

- 93,044 specimens in 54,435 groups could receive a type citation from a peer in their duplicate group
- 1,121,865 specimens in 782,655 groups could receive a georeference from a peer in their duplicate group
- 1,097,168 specimens in 758,416 groups could be linked to an associated specimen image from a peer in their duplicate group
- 2,191,179 specimens are in 792,274 groups which have multiple scientific names within the group (indicating uneven scientific name determination amongst the members of the specimen duplicate group)

D. Repository relationship analysis

The relationship graph derived from duplication links at institutional level (see figure 3) comprises 260 nodes (institutions) and 6,588 weighted edges (relationships between institutions, based on co-participation in a specimen duplicate group, weighted by the number of co-occurrences). The graph was found to contain seven communities: Brazilian herbaria, Australasian herbaria, the regional herbaria in the United States, Colombian herbaria, Canadian herbaria, South African herbaria, and the internationally focussed herbaria found in North America and Europe.

V. Discussion

A. Duplicate identification and assessment

A considerable number of duplicate groups were found in the datamined dataset, and these appear relatively stable across the different assessment flag combinations (see 1), permitting the reconciliation of many specimen duplicates between different specimen repositories. The reconciliation of specimen duplicate groups show that many metadata annotations could be propagated between specimen repositories. As these annotations represent both the most expensive parts of the digitisation process (georeferencing) and the most valuable kind of usage citation (type citation), mobilising these between partners would reduce data management costs, improve the

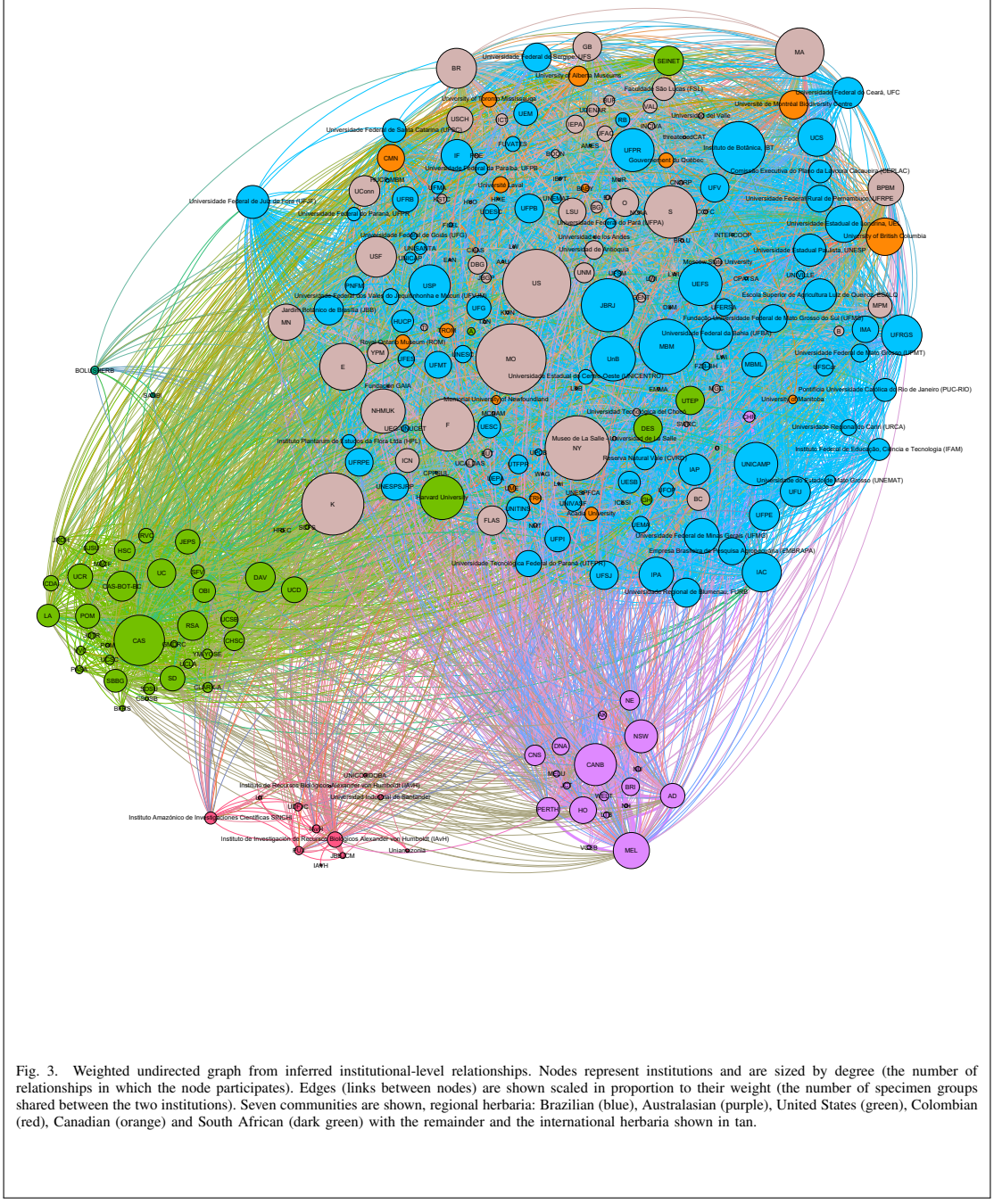


Fig. 3. Weighted undirected graph from inferred institutional-level relationships. Nodes represent institutions and are sized by degree (the number of relationships in which the node participates). Edges (links between nodes) are shown scaled in proportion to their weight (the number of specimen groups shared between the two institutions). Seven communities are shown, regional herbaria: Brazilian (blue), Australasian (purple), United States (green), Colombian (red), Canadian (orange) and South African (dark green) with the remainder and the international herbaria shown in tan.

utility of the digitised specimen data and improve institution level data usage reporting. It is only possible to supply an estimate range for the cost saving of mobilising such a large number of georeferences. Standard procedures tend to batch work by locality, which improves georeferencing speed by focussing on a particular area. A software description paper reports a project georeferencing at a rate of 16.6 (8.3) georeferences per hour and a further separate project achieving a doubling of this rate [20]. A herbarium type specimen focussed project reported "whole process of georeferencing the ca. 3400 Type specimens took eight months (appx. 100 specimens per week)" [21]. It seems that there are significant savings that could be made using the results of this research, given that the number of propagable georeferences is counted at around a million.

B. Repository relationship analysis

The different repositories represented in the dataset are well connected. Viewed at an institutional level, the low incidence of isolated cliques shows healthy inter-institutional working relationships in botany. There are strong links among regionally focussed herbaria in the United States and Australasia. The interconnections between the Brazilian herbaria and their international counterparts show the volume of work that has been focussed on the world's most mega-diverse country [35] and also suggest that the data repatriation projects which aim to mobilise data held out of country [16] have been successful. Quantifying the links between specimen repositories enables evidence drawn from specimen duplicate sharing to be used when building project collaborations. Sets of institutions could be selected to maximise overlap or to maximise complementarity. Better sharing of specimen data between institutions facilitates community curation and helps to reduce data management costs.

VI. Further work

There are several areas in which future work could develop this analysis including further refinement of the analytical approach to cover more data sources, community assessment of interlinked repositories and quality control of annotations by comparison between duplicates. It may be useful to separate future work into two streams: a stream regarding data management and refinement of the data pipeline, and a more conceptual stream regarding implications of the results. An example from each area is outlined here: investigation of the reasons why specimens are not currently identified as duplicates - singleton analysis - and further work on the research recognition of determination annotations made on specimen objects.

Singleton specimens may be due to uneven digitisation and / or lack of participation in data mining process, rather than true singletons, further data analysis work is required to investigate this. It should be possible to use the results from the data mining process to calculate for each collector the likely number of specimens gathered at each collecting event. These numbers would give us a potential view on the number of currently un-digitised specimens, and among these, the likely location of duplicates (in which institutional repositories will they be found).

Traditional taxonomic activity can be separated into three phases - collection of specimens, labelling specimen with names and formal publication of results. The first two phases are absent from traditional publication focussed career credit, yet generate long-term research-grade outputs which may be consulted and referenced by others. As these outputs are now mobilised and used much more widely (due to data mobilisation via the internet) there have been calls for these to be included in the career assessment system for taxonomists [36]. If we recognise that specimens are persistent research objects, which can be uniformly accessed [25], then the labelling of specimens with scientific names could each be considered to meet the minimum criteria for a nanopublication - the smallest unit of research work [37] and credited to individual researchers.

VII. Conclusion

Specimens are research objects which are managed for long term consultation, facilitate scientific discovery and act as vehicles for the dissemination of results. This paper demonstrates that specimens form a shared global resource, and that fragmented information management can be overcome by the reconciliation of specimen duplicates across institutional boundaries. Specimen digitisation efforts and work to define standard representations of digitised metadata have built a critical mass of computable information, which can be used as the input into this process. Identification of specimen duplicates allows quantification of potential specimen metadata exchange between institutional specimen repositories. The result of implementing this data exchange would be to develop and strengthen ties between institutional repositories, improve efficiency of data curation (by eliminating repeated work such as specimen georeferencing) and to improve the metadata holdings and reporting figures for institutional repositories. Conceptually, specimens should be recognised as a unit of research work more granular than the scientific paper, but fulfilling the same functions - communication of results and establishment of a long term record. This recognition of the specimen as a research object would eventually allow the annotation of specimens to be regarded as research work and credited to individual researchers. This may start to address some concerns recently voiced with regard to the many phases of research work conducted by taxonomists which remain absent from publication-focussed career metrics [36].

References

[1] B. Thiers, "The Worlds Herbaria 2017: A Summary Report Based on Data from Index Herbariorum," New York Botanical Garden, Tech. Rep., 2018.
[2] B. J. Conn, "Information Standards in Botanical Databases–the Limits to Data Interchange," *Telopea*, vol. 10, p. 1, 2003.
[3] D. M. .-. Bridson, *The Herbarium Handbook*, 3rd ed. Kew : Royal Botanic Gardens, 1998, 1998.
[4] ipni.org, "International Plant Names Index."

[5] J. McNeil, F. Barrie, W. Buck, V. Demoulin, W. Greuter, D. Hawksworth, P. Herendeen, S. Knapp, K. Marhold, and J. Prado, *International Code of Nomenclature for Algae, Fungi, and Plants (Melbourne Code)*, ser. Regnum vegetabile, 2012, no. 154.

[6] D. P. Bebber, M. A. Carine, J. R. I. Wood, A. H. Wortley, D. J. Harris, G. T. Prance, G. Davidse, J. Paige, T. D. Pennington, N. K. B. Robson, and R. W. Scotland, "Herbaria Are a Major Frontier for Species Discovery," *Proceedings of the National Academy of Science*, vol. 107, pp. 22 169–22 171, Dec. 2010.

[7] D. Triebel, P. Scholz, T. Weibulat, and M. Weiss, "An Online Thesaurus for Standard Bibliographic Data on Exsiccatae in Botany and Mycology," New Orleans, US, 2011.

[8] ITHAKA, "JSTOR Global Plants," 2015, 300+ herbaria 2.5 million objects in 56 collections.

[9] G. L. Bras, M. Pignal, M. L. Jeanson, S. Muller, C. Aupic, B. Carré, G. Flament, M. Gaudeul, C. Gonçalves, V. R. Invernón, F. Jabbour, E. Lerat, P. P. Lowry, B. Offroy, E. P. Pimparé, O. Poncy, G. Rouhan, and T. Haevermans, "The French Muséum National dhistoire Naturelle Vascular Plant Herbarium Collection Dataset," *Scientific Data*, vol. 4, p. 170016, Feb. 2017.

[10] I. Friis, "How to Trace Publications Based on Collections in Specific Herbaria and/or Museums?" Jan. 2012.

[11] K. Winker and J. J. Withrow, "Natural History: Small Collections Make a Big Impact," *Nature*, vol. 493, no. 7433, pp. 480–480, Jan. 2013.

[12] A. Chapman, D., *Uses of Primary Species-Occurrence Data*. Copenhagen: Global Biodiversity Information Facility, 2005.

[13] Q. J. Groom, C. O, and T. Humphrey, "Herbarium Specimens Reveal the Exchange Network of British and Irish Botanists, 1856–1932," *New Journal of Botany*, vol. 4, no. 2, pp. 95–103, 2014.

[14] L. Suhrbier, W.-H. Kusber, O. Tschöpe, A. Güntsch, and W. G. Berendsohn, "AnnoSys—Implementation of a Generic Annotation System for Schema-Based Data Using the Example of Biodiversity Collection Data," *Database*, vol. 2017, Jan. 2017.

[15] J. Macklin, R. Rabeler, and P. Morris, "Developing a Framework for Exchange of Botanical Specimen Data to Reduce Duplicate Effort and Improve Quality Using a Filtered Push," *Botany 2006. California State University–Chico. July28-August 2*, 2006.

[16] REFLORA, "REFLORA Programme."

[17] "Advancing Digitization of Biodiversity Collections — NSF - National Science Foundation."

[18] L. M. Page, B. J. MacFadden, J. A. Fortes, P. S. Soltis, and G. Riccardi, "Digitization of Biodiversity Collections Reveals Biggest Data on Biodiversity," *BioScience*, p. biv104, Aug. 2015.

[19] J. Wieczorek, D. Bloom, R. Guralnick, S. Blum, M. Döring, R. Giovanni, T. Robertson, and D. Vieglais, "Darwin Core: An Evolving Community-Developed Biodiversity Data Standard," *PLoS ONE*, vol. 7, no. 1, p. e29715, Jan. 2012.

[20] A. W. Hill, R. Guralnick, P. Flemons, R. Beaman, J. Wieczorek, A. Ranipeta, V. Chavan, and D. Remsen, "Location, Location, Location: Utilizing Pipelines and Services to More Effectively Georeference the World's Biodiversity Data," *BMC Bioinformatics*, vol. 10, no. 14, pp. 1–9, 2009.

[21] E. Garcia-Milagros and V. A. Funk, "Data: Improving the Use of Information from Museum Specimens: Using Google Earth to Georeference Guiana Shield Specimens in the US National Herbarium," *Frontiers of biogeography*, vol. 2, no. 3, 2010.

[22] J. P. van Oever and M. Gofferjé, "From Pilot to Production: Large Scale Digitisation Project at Naturalis Biodiversity Center," *ZooKeys*, no. 209, pp. 87–92, Jul. 2012.

[23] M. Heerlien, J. Van Leusen, S. Schnörr, S. De Jong-Kole, N. Raes, and K. Van Hulsen, "The Natural History Production Line: An Industrial Approach to the Digitization of Scientific Collections," *J. Comput. Cult. Herit.*, vol. 8, no. 1, pp. 3:1–3:11, Feb. 2015.

[24] P. W. Sweeney, B. Starly, P. J. Morris, Y. Xu, A. Jones, S. Radhakrishnan, C. J. Grassa, and C. C. Davis, "Large-Scale Digitization of Herbarium Specimens: Development and Usage of an Automated, High-Throughput Conveyor System," Mar. 2018.

[25] A. Güntsch, R. Hyam, G. Hagedorn, S. Chagnoux, D. Röpert, A. Casino, G. Droege, F. Glöckler, K. Gödderz, Q. Groom, J. Hoffmann, A. Holleman, M. Kempa, H. Koivula, K. Marhold, N. Nicolson, V. S. Smith, and D. Triebel, "Actionable, Long-Term Stable and Semantic Web Compatible Identifiers for Access to Biological Collection Objects," *Database*, vol. 2017, no. 1, Jan. 2017.

[26] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. d. S. Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. G. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. C. Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, and B. Mons, "The FAIR Guiding Principles for Scientific Data Management and Stewardship," Mar. 2016.

[27] P. Zika, "A New Species of Stonecrop (Sedum Section Gormania , Crassulaceae) from Northern California," *Phytotaxa*, vol. 159, no. 2, pp. 111–121, Feb. 2014.

[28] S. Knapp, "Four New Vining Species of Solanum (Dulcamaroid Clade) from Montane Habitats in Tropical America," *PLoS ONE*, vol. 5, no. 5, p. e10502, May 2010.

[29] GBIF.org, "(11th July 2018) GBIF Occurrence Download (Taxon: Tracheophyta, Basis of Record: Preserved Specimen)," Jul. 2018.

[30] N. Nicolson and A. Tucker, "Identifying Novel Features from Specimen Data for the Prediction of Valuable Collection Trips," ser. Lecture Notes in Computer Science. Springer, Cham, Oct. 2017, pp. 235–246.

[31] D. P. Bebber, M. A. Carine, G. Davidse, D. J. Harris, E. M. Haston, M. G. Penn, S. Cafferty, J. R. I. Wood, and R. W. Scotland, "Big Hitting Collectors Make Massive and Disproportionate Contribution to the Discovery of Plant Species," *Proceedings of the Royal Society of London B: Biological Sciences*, vol. 279, no. 1736, pp. 2269–2274, Jun. 2012.

[32] M. Bastian, S. Heymann, and M. Jacomy, "Gephi: An Open Source Software for Exploring and Manipulating Networks," 2009.

[33] S. Martin, W. M. Brown, R. Klavans, and K. W. Boyack, "OpenOrd: An Open-Source Toolbox for Large Graph Layout," in *Visualization and Data Analysis 2011*, vol. 7868. International Society for Optics and Photonics, 2011, p. 786806.

[34] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast Unfolding of Communities in Large Networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, no. 10, p. P10008, 2008.

[35] R. A. Mittermeier, *Megadiversity: Earth's Biologically Wealthiest Nations*. Agrupacion Sierra Madre, 1997.

[36] L. A. McDade, D. R. Maddison, R. Guralnick, H. A. Piwowar, M. L. Jameson, K. M. Helgen, P. S. Herendeen, A. Hill, and M. L. Vis, "Biology Needs a Modern Assessment System for Professional Productivity," *BioScience*, vol. 61, no. 8, pp. 619–625, Aug. 2011.

[37] P. Groth, A. Gibson, and J. Velterop, "The Anatomy of a Nanopublication," *Information Services & Use*, vol. 30, no. 1-2, pp. 51–56, Jan. 2010.

Appendix C

Conference presentations

C.1 Clustering botanical collections data with a minimised set of features drawn from aggregated specimen data

Nicky Nicolson and Allan Tucker , "Clustering botanical collections data with a minimised set of features drawn from aggregated specimen data," in TDWG 2016 ANNUAL CONFERENCE (Symposium on Big Data Analysis Methods and Techniques as Applied to Biocollections), 2016

This is an early version of work presented in chapter 4.

Abstract:

Current state of play Numerous digitisation and data aggregation efforts are mobilising botanical specimen data. Although digitisation is not yet complete, it is likely that we now have a critical mass of data available from which we can determine patterns.

Problem We know that many duplicate specimens exist, shared between separate botanical collections: these are digitised and transcribed in different herbaria and are yet to be comprehensively linked. Parallel digitisation efforts mean that the transcription of label data also happens in parallel, this results in some critical data fields (such as collector name) being much too variable to be easily used to resolve duplicates. Although not explicitly managed, we have the concept of a collecting trip (a sequence of collections from a particular individual or team). This research aims to uncover this implicit trip data from the aggregated whole. Once we have identified a collecting trip, we should be able to more easily resolve duplicates by cross-linking on the trip identifier, along with the record number and date - i.e. avoiding the transcription variations that we often see in the collector field.

Method and input data This talk will show the output of a clustering analysis run in Python using the machine learning library scikit-learn. The data analysed were drawn from aggregated botanical specimen data accessed via the GBIF portal. Input to the analysis was optimised to use numeric features wherever possible (collection date and record number) along with minimal textual features extracted from the collector team.

Results The outputs of this clustering analysis will be used in a research context - to identify different kinds of collector trip â“ but also have immediate practical applications in data management: to identify duplicate specimens between herbaria, and to identify outliers and label transcription errors. Examples of each of these kinds of outliers will be shown. Numbers of geo-references which can be shared between institutions will also be included. Other applications of this clustering technique within problem domains relevant to biodiversity informatics (e.g. bibliographic reference management) will also be discussed.

C.2 Building your own big data analysis infrastructure for biodiversity science

Matthew Collins, **Nicky Nicolson**, Jorrit Poelen, Alexander Thompson, Jennifer Hammock, Anne Thessen, "Building Your Own Big Data Analysis Infrastructure for Biodiversity Science," Biodiversity Information Science and Standards, vol. 1, e20161, 2017. doi: 10.3897/tdwgproceedings.1.20161

Abstract:

The size of biodiversity data sets, and the size of people's questions around them, are outgrowing the capabilities of desktop applications, single computers, and single developers. Numerous articles in the corporate sector (Delgado 2016) have been written on how much time professionals spend manipulating and formatting large data sets compared to the time they spend on the important work of doing analysis and modeling. To efficiently move large research questions forward, the biodiversity domain needs to transition towards shared infrastructure with the goal of providing a mise en place for researchers to do research with large data.

The GUODA (Global Unified Open Data Access) collaboration was formed to explore tools and use cases for this type of collaborative work on entire biodiversity data sets. Three key parts of that exploration have been: the software and hardware infrastructure needed to be able to work with hundreds of millions of records and terabytes of data quickly, removing the impediment of data formatting and preparation, and workflows centered around GitHub for interacting with peers in an open and collaborative manner.

We will describe our experiences building an infrastructure based on Apache Mesos, Apache Spark, HDFS, Jupyter Notebooks, Jenkins, and Github. We will also enumerate what resources are needed to do things like join millions of records, visualize patterns in whole data sets like iDigBio and the Biodiversity Heritage Library, build graph structures of billions of nodes, analyze terabytes of images, and use natural language processing to explore gigabytes of text. In addition to the hardware and software, we will describe

the kinds of skills needed by staff to design, build, and use this sort of infrastructure and highlight some experiences we have with training students.

Our infrastructure is one of many that are possible. We hope that by showing the amount and type of work we have done to the wider community, other organizations can understand what they would need to speed up their research programs by developing their own collaborative computation and development environments.

C.3 Interactive visualisation of field-collected botanical specimen metadata: supporting data mining process development

Nicky Nicolson and Allan Tucker, "Interactive visualisation of field-collected botanical specimen metadata: Supporting data mining process development," presented at the International Symposium on Intelligent Data Analysis, Den Bosch, Netherlands, 2018. doi: 10.6084/m9.figshare.7321166.v1.

This includes work presented in appendix A.

Abstract:

We outline the development and utilisation of an interactive data visualisation tool, developed throughout a data-intensive research project. Originally designed to aid initial data exploration and gather expert input, the toolkit was further refined to support process design, quality assurance and refinement by viewing data mining results at known stages of a pipeline process, and to enable visualisation of data aggregations used to define new features for use in predictive models. Newly defined features can be regarded as additional data, feeding back into data exploration and forming an iterative process.

The toolkit has contributed to reproducible research by adding tool support and activity logging at one of the loosest stages of the research process.

C.4 Integrating collector and author roles in specimen and publication datasets

Nicky Nicolson, Alan Paton, Sarah Phillips and Allan Tucker, "Integrating Collector and Author Roles Across Specimen and Publication Datasets," Biodiversity Information Science and Standards (Symposium: More than Names : Identifying and Crediting People in Biodiversity Data), 2019. doi: 10.3897/biss.3.35866.

This includes work presented in chapter 5.

Abstract:

This work builds on the outputs of a collector data-mining exercise applied to GBIF mobilised herbarium specimen metadata, which uses unsupervised learning (clustering) to identify collectors from minimal metadata associated with field collected specimens (the DarwinCore terms *recordedby*, *eventdate* and *recordnumber*). Here, we outline methods to integrate these data-mined collector entities (large scale dataset, aggregated from multiple sources, created programmatically) with a dataset of author entities from the International Plant Names Index (smaller scale, single source dataset, created via editorial management). The integration process asserts a generic "scientist" entity with activities in different stages of the species description process: collecting and name publication. We present techniques to investigate specialisations including content - taxa of study - and activity stages: examining if individuals focus on collecting and / or name publication. Finally, we discuss generalisations of this initially herbarium-focussed data-mining and record linkage process to enable applications in a wider context, particularly in zoological datasets.

C.5 Progress in authority management of people names for collections

Quentin J. Groom, Chloé Besombes, Josh Brown, Simon Chagnoux, Teodor Georgiev, Nicole Kearney, Arnald Marcer, **Nicky Nicolson**, Roderic Page, Sarah Phillips, Heimo Rainer, Greg Riccardi, Dominik Röpert, David Peter Shorthouse, Pavel Stoev and Elspeth Margaret Haston, "Progress in Authority Management of People Names for Collections," Biodiversity Information Science and Standards (Symposium: More than Names : Identifying and Crediting People in Biodiversity Data), 2019. doi: 10.3897/biss.3.35074

This work is an early output from the MOBILISE project and is an invited submission to the symposium "More than Names : Identifying and Crediting People in Biodiversity Data" in the forthcoming biodiversity next conference. It includes work presented in chapters 4 and 5.

Abstract:

The concept of building a network of relationships between entities, a knowledge graph, is one of the most effective methods to understand the relations between data. By organizing data, we facilitate the discovery of complex patterns not otherwise evident in the raw data.

Each datum at the nodes of a knowledge graph needs a persistent identifier (PID) to reference it unambiguously. In the biodiversity knowledge graph, people are key elements (R. D. Page 2016). They collect and identify

specimens, they publish, observe, work with each other and they name organisms.

Yet biodiversity informatics has been slow to adopt PIDs for people and people are currently represented in collection management systems as text strings in various formats. These text strings often do not separate individuals within a collecting team and little biographical information is collected to disambiguate collectors.

In March 2019 we organised an international workshop to find solutions to the problem of PIDs for people in collections with the aim of identifying people unambiguously across the world's natural history collections in all of their various roles. Stakeholders were represented from 11 countries, representing libraries, collections, publishers, developers and name registers.

We want to identify people for many reasons. Cross-validation of information about a specimen with biographical information on the specimen can be used to clean data. Mapping specimens from individual collectors across multiple herbaria can geolocate specimens accurately. By linking literature to specimens through their authors and collectors we can create collaboration networks leading to a much better understanding of the scientific contribution of collectors and their institutions. For taxonomists, it will be easier to identify nomenclatural type and syntype material, essential for reliable typification. Overall, it will mean that geographically dispersed specimens can be treated much more like a single distributed infrastructure of specimens as is envisaged in the European Distributed Systems of Scientific Collections Infrastructure (DiSSCo).

There are several person identifier systems in use. For example, the Virtual International Authority File (VIAF) is a widely used system for published authors. The International Standard Name Identifier (ISNI), has broader scope and incorporates VIAF. The ORCID identifier system provides self-registration of living researchers. Also, Wikidata has identifiers of people, which have the advantage of being easy to add to and correct. There are also national systems, such as the French and German authority files, and considerable sharing of identifiers, particularly on Wikidata. This creates an integrated network of identifiers that could act as a brokerage system. Attendees agreed that no one identifier system should be recommended, however, some are more appropriate for particular circumstances.

Some difficulties have still to be resolved to use those identifier schemes for biodiversity : 1) duplicate entries in the same identifier system; 2) handling collector teams and preserving the order of collectors; 3) how we integrate identifiers with standards such as Darwin Core, ABCD and in the Global Biodiversity Information Facility; and 4) many living and dead collectors are only known from their specimens and so they may not pass notability standards required by many authority systems. The participants of

the workshop are now working on a number of fronts to make progress on the adoption of PIDs for people in collections. This includes extending pilots that have already been trialled, working with identifier systems to make them more suitable for specimen collectors and talking to service providers to encourage them to use ORCID IDs to identify their users. It was concluded that resolving the problem of person identifiers for collections is largely not a lack of a solution, but a need to implement solutions that already exist.

C.6 Examining herbarium specimen citation: developing a literature based institutional impact measure

Nicky Nicolson, Alan Paton, Sarah Phillips and Allan Tucker, "Examining herbarium specimen citation: Developing a literature based institutional impact measure," Biodiversity Information Science and Standards, 2019. doi: 10.3897/biss.3.37198

This includes work discussed in the conclusions to this thesis (chapter 7).

Abstract:

Herbarium specimens are critical components of the research process - providing "what, where, when" evidence for species distributions and through type designation, providing the basis for un-ambiguous, standardised nomenclature facilitating the interpretation of scientific names. Specimen references are embedded within research article texts, by convention usually presented in a relatively formalised fashion. As this is a domain-specific practice, general publishers tend not to provide tools for detecting and tracking specimen references to enable bibliometric-style calculations and navigation to the referenced specimen, as is common practice in literature reference management. This means that it is difficult to measure impact, which affects both the individuals responsible for the collection and determination of herbarium specimens (McDade et al. 2011), and the institutions responsible for their long-term management.

Specimen digitisation - creating searchable data repositories of metadata and/or images - has enabled many new and larger scale uses for herbarium specimens and their associated data, and stimulated interest in quantifying usage and measuring institutional impact. To date, these impact measures have been conducted by examining usage statistics for specimen portals, or by text searching for specimen identifier patterns.

This research uses text mining and document classification techniques to detect article sections likely to contain specimen references, which are then extracted, classified and counted. A dataset of taxonomic publications categorised into paragraph-level units is used to train a text classifier to

predict the presence of specimen references within component units of articles (sections or paragraphs). The input to the classifier is a set of features derived from the text contents of paragraphs, which detect content such as latitude/longitude, dates and bracketed lists of herbarium codes. Article units classified as containing specimen references are processed to extract a minimal representation of the specimen reference, including the abbreviated codes for the institutional holder(s) of the specimen material. This allows total and per-institution counts to be calculated, which can be compared to datasets of Global Biodiversity Information Facility data citations, to institutional-level type citations in nomenclatural acts recorded by the International Plant Names Index and to usage statistics recorded by institutional data repositories. As well as counting specimen references, distinct specimen reference styles are detected and quantified, including the use of numeric and persistent identifiers (Güntsch et al. 2017) which can be used to access a standardised metadata record for the specimen.

We will present an assessment of the classification and detection process and initial results, and discuss future work to develop this approach to work with different kinds of literature inputs. These techniques have the potential to allow institutions to make better use of existing information to help assess the use and impact of their specimen and data holdings.

Appendix D

Grants

This appendix lists competitively awarded grants that will support further research on the topics presented in this thesis.

D.1 SYNTHESYS+

D.1.1 Funder and timescale

SYNTHESYS+ is "a pan-European collections infrastructure project and the fourth iteration of the SYNTHESYS programme, funded by the European Commission. SYNTHESYS+ will commence on 1 February 2019 and run until 31 January 2023" (*SYNTHESYS - an Integrated European Infrastructure for Researchers in the Natural Sciences* 2019).

D.1.2 Aims and objectives

> SYNTHESYS aims to produce an accessible, integrated European resource for research users in the natural sciences. SYNTHESYS will create a shared, high quality approach to the management, preservation, and access to leading European natural history collections.
>
> A core element in SYNTHESYS is to provide funded researcher visits (*Access*) to the 390,000,000 specimens housed by SYNTHESYS institutions. In particular, the 4,049,800 type specimens.
>
> Alongside the Access, a *Joint Research Activity* (JRA) aims to improve the quality of and increase access to digital collections and data within natural history institutions by developing virtual collections.
>
> *Network Activities* (NA) will provide enhanced quality and quantity of online collections information to virtual Users and will implement best practice benchmarks in collections care to raise standards and improve accessibility to collections for all physical

> Users. (*SYNTHESYS - an Integrated European Infrastructure for Researchers in the Natural Sciences* 2019)

One of the Joint Research Activities defined in SYNTHESYS+ is to develop a "*Specimen Data Refinery*":

> This research will integrate machine learning, Artificial Intelligence, and human approaches to extract, enhance, and annotate data from digital images and records at scale. Many collections-holding institutions still need to digitise the bulk of their collections. Digitisation takes time and resources. One of the major challenges in digitising massive collections is finding ways of ensuring high-quality collections data can be processed at pace.
>
> We will use new technological approaches, such as computer vision, data mining and machine learning, to rapidly enhance minimal natural history specimen records using images (e.g. of labels, specimens or registers) and unstructured text at scale. These approaches will be largely automated and may support record enhancement by experts as well as members of the public (crowdsourcing).

From: `https://www.synthesys.info/joint-research-activities.html`

D.1.3 Contributions from this research

The data-mining process outlined in chapter 4 will be utilised in the development of the *specimen data refinery* for use in the mass digitisation of specimens.

The data-driven generation of a institutional network as presented in chapter 6 will be used in the development of standards to encode collections descriptions, a network activity jointly associated with the Biodiversity Information Standards organisation.

D.1.4 Progress to date

A kick-off meeting for the *Specimen Data Refinery* was held in April 2019 (virtual meeting), and work is intended to start in the final quarter of 2019.

The *collections descriptions* network activity is meeting in September 2019 (London, UK).

D.2 MOBILISE

D.2.1 Funder and timescale

MOBILISE is a European Cooperation in Science and Technology network (COST action) on "Mobilising Data, Experts and Policies in Scientific Collections". It is funded for a period of four years, and is intended to contribute towards the Research Infrastructure DiSSCo - Distributed System of Scientific Collections.

D.2.2 Aims and objectives

The main aim of the MOBILISE COST action is to

> build an inclusive, bottom-up and responsive network to address the urgent challenges around the mobilisation and linking of natural science collections reference information

Which will be realised via a number of objectives:

1. *Assessing*: Assessment and comparison of existing standards and protocols on digitisation, mobilisation of biodiversity / collection data, data management and publication.
2. *Bridging*: Linking complementary expertise of information scientists, biodiversity researchers and geoscientists leading to new concepts, technical innovations and products.
3. *Compiling*: Develop recommendations and best practices linking regional and global community standards and guidelines
4. Planning: Increase sustainability of bio- and geodiversity data providing infrastructures and define a common research agenda for long-term preservation and re-use of biodiversity data.
5. *Facilitating*: Facilitate implementation of common standards and of newly developed techniques by training and education.
6. *Disseminating*: Raise awareness and open bio- and geo-diversity information systems to interdisciplinary research and to the society in general.

Further information is available at www.mobilise-action.eu and www.cost.eu/actions/CA17106.

D.2.3 Contributions from this research

Entities and interrelationships resultant from data-mining in chapter 4 will be used to inform an assessment of data standards.

Agent analyses presented in chapter 5 will be further developed through the MOBILISE group convened to research "Authority Management of People Names".

Annotation flow and inter-institutional relationships resulting from chapter 6 will be further developed in a working group on "New concepts and standards for data management".

D.2.4 Progress to date

The group working on "Authority Management of People Names" met in March 2019 (Sofia, Bulgaria). Research from chapters 4 and 6 was presented at this meeting.

A conference paper reporting progress to date (details in section C.5) has been accepted in the forthcoming biodiversity next conference, which will also feature a pre-conference MOBILISE working meeting.

Bibliography

Access to Biological Collections Data task group (2007). *Access to Biological Collection Data (ABCD) Schema, Version 2.06*. Biodiversity Information Standards (TDWG). URL: http://www.tdwg.org/standards/115.

Advancing Digitization of Biodiversity Collections | NSF - National Science Foundation (2018). URL: https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=503559.

Akella, L. M., C. N. Norton, and H. Miller (2012). "NetiNeti: Discovery of Scientific Names from Text Using Machine Learning Methods". In: *BMC Bioinformatics* 13, p. 211. ISSN: 1471-2105. DOI: 10.1186/1471-2105-13-211. URL: http://dx.doi.org/10.1186/1471-2105-13-211.

Barber, D. (2012). *Bayesian Reasoning and Machine Learning*. Cambridge University Press. 739 pp. ISBN: 978-0-521-51814-7.

Bastian, M., S. Heymann, and M. Jacomy (2009). "Gephi: An Open Source Software for Exploring and Manipulating Networks". In: *Proceedings of the Third International Conference on Weblogs and Social Media*. San Jose, California: AAAI Press, p. 396. ISBN: 978-1-57735-421-5. URL: https://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154.

Bebber, D. P., M. A. Carine, G. Davidse, D. J. Harris, E. M. Haston, M. G. Penn, S. Cafferty, J. R. I. Wood, and R. W. Scotland (2012). "Big Hitting Collectors Make Massive and Disproportionate Contribution to the Discovery of Plant Species". In: *Proceedings of the Royal Society of London B: Biological Sciences* 279.1736, pp. 2269–2274. ISSN: 0962-8452, 1471-2954. DOI: 10.1098/rspb.2011.2439. URL: http://rspb.royalsocietypublishing.org/content/279/1736/2269.

Bebber, D. P., M. A. Carine, J. R. I. Wood, A. H. Wortley, D. J. Harris, G. T. Prance, G. Davidse, J. Paige, T. D. Pennington, N. K. B. Robson, and R. W. Scotland (2010). "Herbaria Are a Major Frontier for Species Discovery". In: *Proceedings of the National Academy of Science* 107, pp. 22169–22171. ISSN: 0027-8424. DOI: 10.1073/pnas.1011841108.

Bebber, D. P., J. R. I. Wood, C. Barker, and R. W. Scotland (2013). "Author Inflation Masks Global Capacity for Species Discovery in Flowering Plants". In: *The New Phytologist*. ISSN: 1469-8137. DOI: 10.1111/nph.12522.

Berthold, M. R., C. Borgelt, F. Höppner, and F. Klawonn (2010). *Guide to Intelligent Data Analysis*. Texts in Computer Science. London: Springer

London. 410 pp. ISBN: 978-1-84882-259-7 978-1-84882-260-3. URL: http://link.springer.com/10.1007/978-1-84882-260-3.

Blondel, V. D., J.-L. Guillaume, R. Lambiotte, and E. Lefebvre (2008). "Fast Unfolding of Communities in Large Networks". In: *Journal of Statistical Mechanics: Theory and Experiment* 2008.10, P10008. ISSN: 1742-5468. DOI: 10.1088/1742-5468/2008/10/P10008. URL: http://stacks.iop.org/1742-5468/2008/i=10/a=P10008.

Bras, G. L., M. Pignal, M. L. Jeanson, S. Muller, C. Aupic, B. Carré, G. Flament, M. Gaudeul, C. Gonçalves, V. R. Invernón, F. Jabbour, E. Lerat, P. P. Lowry, B. Offroy, E. P. Pimparé, O. Poncy, G. Rouhan, and T. Haevermans (2017). "The French Muséum National d'histoire Naturelle Vascular Plant Herbarium Collection Dataset". In: *Scientific Data* 4, p. 170016. ISSN: 2052-4463. DOI: 10.1038/sdata.2017.16. URL: http://www.nature.com/articles/sdata201716.

Bridson, D. M. (1998). *The Herbarium Handbook*. 3rd. ed. Royal Botanic Gardens, Kew. 334 pp. ISBN: 978-1-900347-43-3.

Brummitt, R. K. (2006). "The Democratic Processes of Botanical Nomenclature". In: *Taxonomy and Plant Conservation: The Cornerstone of the Conservation and the Sustainable Use of Plants*. Ed. by E. Leadlay and S. Jury. Cambridge: Cambridge University Press, pp. 101–129.

Brummitt, R. K. and C. E. Powell (1992). *Authors of Plant Names: A List of Authors of Scientific Names of Plants, with Recommended Standard Forms of Their Names, Including Abbreviations*. [London]: Royal Botanic Gardens, Kew. ISBN: 0-947643-44-3 978-0-947643-44-7.

Buttigieg, P. L. (2015). "Shaping the Semantic Layer by Mining Digitised Data: An Encounter between iDigBio's Plant Records and the Environment Ontology (ENVO)" (Bremen, Germany and East Lansing, MI, USA). URL: https://www.idigbio.org/content/webinar-shaping-semantic-layer-mining-digitised-data-encounter-between-idigbios-plant.

Chapman, A. D. (2005). *Uses of Primary Species-Occurrence Data*. Copenhagen: Global Biodiversity Information Facility. 100 pp.

Clark, T., S. Martin, and T. Liefeld (2004). "Globally Distributed Object Identification for Biological Knowledgebases". In: *Briefings in Bioinformatics* 5.1, pp. 59–70. ISSN: 1467-5463. DOI: 10.1093/bib/5.1.59. URL: https://academic.oup.com/bib/article/5/1/59/430459.

Conn, B. J. (2003). "Information Standards in Botanical Databases–the Limits to Data Interchange". In: *Telopea* 10, p. 1.

Convention on Biological Diversity (2007). *What Is the Problem? (The Taxonomic Impediment)*. URL: https://www.cbd.int/gti/problem.shtml.

Croft, J., N. Cross, S. Hinchcliffe, E. N. Lughadha, P. F. Stevens, J. G. West, and G. Whitbread (1999). "Plant Names for the 21st Century: The International Plant Names Index, a Distributed Data Source of General Accessibility".

In: *TAXON* 48.2, pp. 317–324. ISSN: 1996-8175. DOI: 10.2307/1224436. URL: https://onlinelibrary.wiley.com/doi/abs/10.2307/1224436.

Cryer, P., R. Hyam, C. Miller, N. Nicolson, É. Ó. Tuama, R. Page, J. Rees, G. Riccardi, K. Richards, and R. White (2009). *Adoption of Persistent Identifiers for Biodiversity Informatics*. Copenhagen: Global Biodiversity Information Facility (GBIF) Secretariat. 23 pp. URL: http://www.gbif.org/resource/80662.

Cui, H. (2012). "CharaParser for Fine-Grained Semantic Annotation of Organism Morphological Descriptions". In: *Journal of the American Society for Information Science and Technology* 63.4, pp. 738–754. ISSN: 1532-2890. DOI: 10.1002/asi.22618. URL: http://onlinelibrary.wiley.com/doi/10.1002/asi.22618/abstract.

Dallwitz, M. (2006). *Description Language for Taxonomy (DELTA), Version 2006-11-24*. Biodiversity Information Standards (TDWG). URL: http://www.tdwg.org/standards/107.

Darwin Core Task Group (2009). *Darwin Core*. Biodiversity Information Standards (TDWG). URL: http://www.tdwg.org/standards/450.

Ester, M., H.-P. Kriegel, J. Sander, X. Xu, et al. (1996). "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise." In: *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*. Vol. 96. Portland, Oregon: AAAI Press, pp. 226–231.

Final Report of the OECD Megascience Forum Working Group on Biological Informatics (1999). Organisation for Economic Co-operation and Development, p. 74. URL: http://www.oecd.org/sti/inno/2105199.pdf.

Flann, C., N. Turland, and A. M. Monro (2014). "Report on Botanical Nomenclature—Melbourne 2011. XVIII International Botanical Congress, Melbourne: Nomenclature Section, 18–22 July 2011". In: *PhytoKeys* 41, pp. 1–289. ISSN: 1314-2003, 1314-2011. DOI: 10.3897/phytokeys.41.8398. URL: http://phytokeys.pensoft.net/browse_journal_issue_documents.php?issue_id=595.

Friis, I. (2012). *How to Trace Publications Based on Collections in Specific Herbaria and/or Museums?* E-mail. URL: http://mailman.nhm.ku.edu/pipermail/taxacom/2012-January/121650.html.

Garcia-Milagros, E. and V. A. Funk (2010). "Improving the Use of Information from Museum Specimens: Using Google Earth to Georeference Guiana Shield Specimens in the US National Herbarium". In: *Frontiers of Biogeography* 2.3, pp. 71–77. ISSN: 1948-6596. DOI: 10.21425/F5FBG12348.

GBIF.org (2018). *(11th July 2018) GBIF occurrence download (Taxon: Tracheophyta, Basis of record: specimen)*. http://doi.org/10.15468/dl.wjjrdk. DOI: 10.15468/dl.wjjrdk.

Goodwin, Z. A., D. J. Harris, D. Filer, J. R. I. Wood, and R. W. Scotland (2015). "Widespread Mistaken Identity in Tropical Plant Collections". In: *Current Biology* 25.22, R1066–R1067. ISSN: 0960-9822. DOI: 10.1016/j.cub.2015.10.002. URL: http://www.sciencedirect.com/science/article/pii/S0960982215012282.

Groom, Q. J., C. O'Reilly, and T. Humphrey (2014). "Herbarium Specimens Reveal the Exchange Network of British and Irish Botanists, 1856–1932". In: *New Journal of Botany* 4.2, pp. 95–103. URL: http://www.tandfonline.com/doi/abs/10.1179/2042349714Y.0000000041.

Groth, P., A. Gibson, and J. Velterop (2010). "The Anatomy of a Nanopublication". In: *Information Services & Use* 30.1-2, pp. 51–56. ISSN: 0167-5265. DOI: 10.3233/ISU-2010-0613. URL: https://content.iospress.com/articles/information-services-and-use/isu613.

Guédon, J.-C. (2001). *In Oldenburg's Long Shadow: Librarians, Research Scientists, Publishers, and the Control of Scientific Publishing*. Association of Research Libraries. 70 pp. ISBN: 0-918006-81-3.

Güntsch, A., R. Hyam, G. Hagedorn, S. Chagnoux, D. Röpert, A. Casino, G. Droege, F. Glöckler, K. Gödderz, Q. Groom, J. Hoffmann, A. Holleman, M. Kempa, H. Koivula, K. Marhold, N. Nicolson, V. S. Smith, and D. Triebel (2017). "Actionable, Long-Term Stable and Semantic Web Compatible Identifiers for Access to Biological Collection Objects". In: *Database: The Journal of Biological Databases and Curation* 2017. ISSN: 1758-0463. DOI: 10.1093/database/bax003.

Guralnick, R., T. Conlin, J. Deck, B. J. Stucky, and N. Cellinese (2014). "The Trouble with Triplets in Biodiversity Informatics: A Data-Driven Case against Current Identifier Practices". In: *PLoS ONE* 9.12. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0114069. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4254916/.

Hagberg, A. A., D. A. Schult, and P. J. Swart (2008). "Exploring Network Structure, Dynamics, and Function Using NetworkX". In: *Proceedings of the 7th Python in Science Conference* (Pasadena, CA USA). Ed. by G. Varoquaux, T. Vaught, and J. Millman, pp. 11–15.

Hagedorn, G., K. Thiele, R. Morris, and P. B. Heidorn (2005). *Structured Descriptive Data (SDD) W3c-Xml-Schema, Version 1.0*. Biodiversity Information Standards (TDWG). URL: http://www.tdwg.org/standards/116.

Hand, D. J. (1997). "Intelligent Data Analysis: Issues and Opportunities". In: *International Symposium on Intelligent Data Analysis*. Springer, pp. 1–14.

Hardisty, A., D. Roberts, and the biodiversity informatics community (2013). "A Decadal View of Biodiversity Informatics: Challenges and Priorities". In: *BMC Ecology* 13.1, p. 16. ISSN: 1472-6785. DOI:

10.1186/1472-6785-13-16. URL: http://www.biomedcentral.com/1472-6785/13/16/abstract.

Hedrick, B., M. Heberling, E. Meineke, K. Turner, C. Grassa, D. Park, J. Kennedy, J. Clarke, J. Cook, D. Blackburn, S. Edwards, and C. Davis (2019). *Digitization and the Future of Natural History Collections*. e27859v1. PeerJ Inc. DOI: 10.7287/peerj.preprints.27859v1. URL: https://peerj.com/preprints/27859.

Heerlien, M., J. Van Leusen, S. Schnörr, S. De Jong-Kole, N. Raes, and K. Van Hulsen (2015). "The Natural History Production Line: An Industrial Approach to the Digitization of Scientific Collections". In: *Journal on Computing and Cultural Heritage* 8.1, 3:1–3:11. ISSN: 1556-4673. DOI: 10.1145/2644822. URL: http://doi.acm.org/10.1145/2644822.

Hill, A. W., R. Guralnick, P. Flemons, R. Beaman, J. Wieczorek, A. Ranipeta, V. Chavan, and D. Remsen (2009). "Location, Location, Location: Utilizing Pipelines and Services to More Effectively Georeference the World's Biodiversity Data". In: *BMC Bioinformatics* 10.14, pp. 1–9. ISSN: 1471-2105. DOI: 10.1186/1471-2105-10-S14-S3. URL: http://dx.doi.org/10.1186/1471-2105-10-S14-S3.

Ho, T. K. (1995). "Random Decision Forests". In: *Proceedings of 3rd International Conference on Document Analysis and Recognition*. Vol. 1. IEEE, pp. 278–282.

Hobern, D., A. Apostolico, E. Arnaud, J. C. Bello, D. Canhos, G. Dubois, D. Field, E. Alonso García, A. Hardisty, J. Harrison, B. Heidorn, L. Krishtalka, E. Mata, R. Page, C. Parr, J. Price, and S. Willoughby (2012). "Global Biodiversity Informatics Outlook: Delivering Biodiversity Knowledge in the Information Age". In: *Global Biodiversity Information Facility*. DOI: 10.15468/6jxa-yb44. URL: https://www.gbif.org/document/80859.

Hobern, D., B. Baptiste, K. Copas, R. Guralnick, A. Hahn, E. van Huis, E.-S. Kim, M. McGeoch, I. Naicker, L. Navarro, D. Noesgaard, M. Price, A. Rodrigues, D. Schigel, C. A. Sheffield, and J. Wieczorek (2019). "Connecting Data and Expertise: A New Alliance for Biodiversity Knowledge". In: *Biodiversity Data Journal* 7, e33679. ISSN: 1314-2828. DOI: 10.3897/BDJ.7.e33679. URL: https://bdj.pensoft.net/article/33679/.

Hogeweg, P. (2011). "The Roots of Bioinformatics in Theoretical Biology". In: *PLoS Computational Biology* 7.3. ISSN: 1553-734X. DOI: 10.1371/journal.pcbi.1002021. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3068925/.

Hunter, J. D. (2007). "Matplotlib: A 2D Graphics Environment". In: *Computing in science & engineering* 9.3, pp. 90–95.

Hyam, R., R. E. Drinkwater, and D. J. Harris (2012). "Stable Citations for Herbarium Specimens on the Internet: An Illustration from a Taxonomic

Revision of Duboscia (Malvaceae)". In: *Phytotaxa* 73, pp. 17–30. DOI: 10.11646/phytotaxa.73.1.4. URL: https://biotaxa.org/Phytotaxa/article/view/phytotaxa.73.1.4.

International Plant Names Index (n.d.). URL: www.ipni.org.

International Union for Conservation of Nature, Iucn Species Survival Commission, International Union for Conservation of Nature, and Natural Resources. Species Survival Commission (2001). *IUCN Red List Categories and Criteria*. IUCN.

ITHAKA (2015). *JSTOR Global Plants*. URL: http://plants.jstor.org/.

Jaiswal, P., S. Avraham, K. Ilic, E. A. Kellogg, S. McCouch, A. Pujar, L. Reiser, S. Y. Rhee, M. M. Sachs, M. Schaeffer, L. Stein, P. Stevens, L. Vincent, D. Ware, and F. Zapata (2005). "Plant Ontology (PO): A Controlled Vocabulary of Plant Structures and Growth Stages". In: *Comparative and Functional Genomics* 6.7-8, pp. 388–397. ISSN: 1531-6912. DOI: 10.1002/cfg.496. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2447502/.

Jones, E., T. Oliphant, P. Peterson, et al. (2001–). *SciPy: Open Source Scientific Tools for Python*. URL: http://www.scipy.org/.

Joppa, L. N., D. L. Roberts, and S. L. Pimm (2011). "The Population Ecology and Social Behaviour of Taxonomists". In: *Trends in Ecology & Evolution* 26.11, pp. 551–553. ISSN: 0169-5347. DOI: 10.1016/j.tree.2011.07.010. URL: http://www.cell.com/trends/ecology-evolution/abstract/S0169-5347(11)00208-4.

Kelling, S., W. M. Hochachka, D. Fink, M. Riedewald, R. Caruana, G. Ballard, and G. Hooker (2009). "Data-Intensive Science: A New Paradigm for Biodiversity Studies". In: *BioScience* 59.7, pp. 613–620. ISSN: 0006-3568, 1525-3244. DOI: 10.1525/bio.2009.59.7.12. URL: http://bioscience.oxfordjournals.org/cgi/doi/10.1525/bio.2009.59.7.12.

Kluyver, T., B. Ragan-Kelley, F. Pérez, B. Granger, M. Bussonnier, J. Frederic, K. Kelley, J. Hamrick, J. Grout, S. Corlay, P. Ivanov, D. Avila, S. Abdalla, and C. Willing (2016). "Jupyter Notebooks – a Publishing Format for Reproducible Computational Workflows". In: *Positioning and Power in Academic Publishing: Players, Agents and Agendas*. Ed. by F. Loizides and B. Schmidt. IOS Press, pp. 87–90.

Knapp, S. and D. Wright (2010). "E-Publish or Perish". In: *Systema Naturae 250-the Linnaean Ark*. Ed. by A. Polaszek. London: Taylor & Francis, pp. 83–93.

Knapp, S. (2010). "Four New Vining Species of Solanum (Dulcamaroid Clade) from Montane Habitats in Tropical America". In: *PLoS ONE* 5.5, e10502. DOI: 10.1371/journal.pone.0010502. URL: http://dx.doi.org/10.1371/journal.pone.0010502.

Koning, D., I. N. Sarkar, and T. Moritz (2005). "TaxonGrab: Extracting Taxonomic Names From Text". In: *Biodiversity Informatics* 2. ISSN: 1546-9735. DOI: 10.17161/bi.v2i0.17. URL: https://journals.ku.edu/jbi/article/view/17.

Leary, P. R., D. P. Remsen, C. N. Norton, D. J. Patterson, and I. N. Sarkar (2007). "uBioRSS: Tracking Taxonomic Literature Using RSS". In: *Bioinformatics* 23.11, pp. 1434–1436. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btm109. URL: https://academic.oup.com/bioinformatics/article/23/11/1434/201163.

Lindon, H. L., L. M. Gardiner, A. Brady, and M. S. Vorontsova (2015). "Fewer than Three Percent of Land Plant Species Named by Women: Author Gender over 260 Years". In: *TAXON* 64.2, pp. 209–215. ISSN: 1996-8175. DOI: 10.12705/642.4. URL: https://onlinelibrary.wiley.com/doi/abs/10.12705/642.4.

Lughadha, E. N. (2004). "Towards a Working List of All Known Plant Species". In: *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 359.1444, pp. 681–687. ISSN: 0962-8436, 1471-2970. DOI: 10.1098/rstb.2003.1446. URL: http://rstb.royalsocietypublishing.org/content/359/1444/681.

Macklin, J., R. Rabeler, and P. Morris (2006). "Developing a Framework for Exchange of Botanical Specimen Data to Reduce Duplicate Effort and Improve Quality Using a 'Filtered Push'". In: *Botany 2006*. California State University–Chico. URL: http://www.2006.botanyconference.org/engine/search/index.php?func=detail&aid=587.

Marsland, S. (2014). *Machine Learning: An Algorithmic Perspective*. 2nd ed. Chapman and Hall/CRC. 457 pp. ISBN: 978-1-4665-8328-3.

Martin, S., W. M. Brown, R. Klavans, and K. W. Boyack (2011). "OpenOrd: An Open-Source Toolbox for Large Graph Layout". In: *Visualization and Data Analysis 2011*. Visualization and Data Analysis 2011. Vol. 7868. International Society for Optics and Photonics, p. 786806. DOI: 10.1117/12.871402. URL: https://www.spiedigitallibrary.org/conference-proceedings-of-spie/7868/786806/OpenOrd-an-open-source-toolbox-for-large-graph-layout/10.1117/12.871402.short.

McDade, L. A., D. R. Maddison, R. Guralnick, H. A. Piwowar, M. L. Jameson, K. M. Helgen, P. S. Herendeen, A. Hill, and M. L. Vis (2011). "Biology Needs a Modern Assessment System for Professional Productivity". In: *BioScience* 61.8, pp. 619–625. ISSN: 0006-3568. DOI: 10.1525/bio.2011.61.8.8.

McKinney, W. (2010). "Data Structures for Statistical Computing in Python". In: *Proceedings of the 9th Python in Science Conference*. Ed. by S. van der Walt and J. Millman, pp. 51–56.

McNeil, J., F. Barrie, W. Buck, V. Demoulin, W. Greuter, D. Hawksworth, P. Herendeen, S. Knapp, K. Marhold, and J. Prado (2012). *International Code of Nomenclature for Algae, Fungi, and Plants (Melbourne Code)*. Regnum Vegetabile 154. 232 pp. ISBN: 978-3-87429-425-6.

Meineke, E. K., A. T. Classen, N. J. Sanders, and T. J. Davies (2019). "Herbarium Specimens Reveal Increasing Herbivory over the Past Century". In: *Journal of Ecology* 107.1, pp. 105–117. ISSN: 1365-2745. DOI: 10.1111/1365-2745.13057. URL: https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/1365-2745.13057.

Michener, C. D., J. O. Corliss, R. S. Cowan, P. H. Raven, C. W. Sabrosky, D. S. Squires, and G. W. Wharton (1970). *Systematics in Support of Biological Research*. Washington D.C.: Division of Biology and Agriculture, National Research Council. 25 pp.

Michener, W. K. and M. B. Jones (2012). "Ecoinformatics: Supporting Ecology as a Data-Intensive Science". In: *Trends in Ecology & Evolution*. Ecological and Evolutionary Informatics 27.2, pp. 85–93. ISSN: 0169-5347. DOI: 10.1016/j.tree.2011.11.016. URL: http://www.sciencedirect.com/science/article/pii/S0169534711003399.

Mittermeier, R. A. and C. G. Mittermeier (1997). *Megadiversity: Earth's Biologically Wealthiest Nations*. CEMEX. 501 pp. ISBN: 978-968-6397-50-5.

Nadeau, D. and S. Sekine (2007). "A Survey of Named Entity Recognition and Classification". In: *Lingvisticae Investigationes* 30.1, pp. 3–26. DOI: 10.1075/li.30.1.03nad.

Oliphant, T. E. (2006). *A Guide to NumPy*. Vol. 1. Trelgol Publishing USA.

Orrell, T. M. (2019). *NMNH Extant Specimen Records. Version 1.21. National Museum of Natural History, Smithsonian Institution. Occurrence Dataset*. DOI: 10.15468/hnhrg3. URL: https://www.gbif.org/occurrence/1317392082.

Page, L. M., B. J. MacFadden, J. A. Fortes, P. S. Soltis, and G. Riccardi (2015). "Digitization of Biodiversity Collections Reveals Biggest Data on Biodiversity". In: *BioScience*, biv104. ISSN: 0006-3568, 1525-3244. DOI: 10.1093/biosci/biv104. URL: http://bioscience.oxfordjournals.org/content/early/2015/08/06/biosci.biv104.

Page, R. D. (2008). "LSID Tester, a Tool for Testing Life Science Identifier Resolution Services". In: *Source Code for Biology and Medicine* 3.1, p. 2. ISSN: 1751-0473. DOI: 10.1186/1751-0473-3-2. URL: https://doi.org/10.1186/1751-0473-3-2.

— (2013). "BioNames: Linking Taxonomy, Texts, and Trees". In: *PeerJ* 1, e190. ISSN: 2167-8359. DOI: 10.7717/peerj.190. URL: https://peerj.com/articles/190.

— (2016). "Towards a Biodiversity Knowledge Graph". In: *Research Ideas and Outcomes* 2, e8767. ISSN: 2367-7163. DOI: 10.3897/rio.2.e8767.

Parr, C. S., R. Guralnick, N. Cellinese, and R. D. M. Page (2012). "Evolutionary Informatics: Unifying Knowledge about the Diversity of Life". In: *Trends in Ecology & Evolution*. Ecological and Evolutionary Informatics 27.2, pp. 94–103. ISSN: 0169-5347. DOI: 10.1016/j.tree.2011.11.001. URL: http://www.sciencedirect.com/science/article/pii/S0169534711003247.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay (2011). "Scikit-Learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12, pp. 2825–2830.

Penev, L., W. J. Kress, S. Knapp, D.-Z. Li, and S. Renner (2010). "Fast, Linked, and Open – the Future of Taxonomic Publishing for Plants: Launching the Journal PhytoKeys". In: *PhytoKeys* 1.0. ISSN: 1314-2003, 1314-2011. DOI: 10.3897/phytokeys.1.642. URL: http://www.pensoft.net/journal_home_page.php?journal_id=3&page=article&type=show&article_id=642&abstract=1.

Penn, M. G., S. Cafferty, and M. Carine (2018). "Mapping the History of Botanical Collectors: Spatial Patterns, Diversity, and Uniqueness through Time". In: *Systematics and Biodiversity* 16.1, pp. 1–13. ISSN: 1477-2000. DOI: 10.1080/14772000.2017.1355854. URL: https://doi.org/10.1080/14772000.2017.1355854.

Peterson, A. T., S. Knapp, R. Guralnick, J. Soberón, and M. T. Holder (2010). "The Big Questions for Biodiversity Informatics". In: *Systematics and Biodiversity* 8.2, pp. 159–168. ISSN: 1477-2000. DOI: 10.1080/14772001003739369. URL: http://dx.doi.org/10.1080/14772001003739369.

Pitman, N. C. A. and P. M. Jørgensen (2002). "Estimating the Size of the World's Threatened Flora". In: *Science* 298.5595, pp. 989–989. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.298.5595.989. URL: http://science.sciencemag.org/content/298/5595/989.

Purves, D., J. P. W. Scharlemann, M. Harfoot, T. Newbold, D. P. Tittensor, J. Hutton, and S. Emmott (2013). "Ecosystems: Time to Model All Life on Earth". In: *Nature* 493.7432, pp. 295–297. ISSN: 0028-0836. DOI: 10.1038/493295a. URL: http://www.nature.com/nature/journal/v493/n7432/full/493295a.html.

Rees, T. (2014). "Taxamatch, an Algorithm for Near ('Fuzzy') Matching of Scientific Names in Taxonomic Databases". In: *PLoS ONE* 9.9, e107510. DOI: 10.1371/journal.pone.0107510. URL: http://dx.doi.org/10.1371/journal.pone.0107510.

REFLORA (2017). *REFLORA Programme*. URL: http://reflora.jbrj.gov.br.

Remsen, D., K. Braak, M. Döring, and T. Robertson (2017). *Darwin Core Archives – How-to Guide*. Copenhagen: Global Biodiversity Information Facility. URL: https://github.com/gbif/ipt/wiki/DwCAHowToGuide.

Richards, K. (2010). *TDWG GUID Applicability Statement, Version 2010-09*. Biodiversity Information Standards (TDWG). URL: http://www.tdwg.org/standards/150.

Robertson, T., M. Döring, R. Guralnick, D. Bloom, J. Wieczorek, K. Braak, J. Otegui, L. Russell, and P. Desmet (2014). "The GBIF Integrated Publishing Toolkit: Facilitating the Efficient Publishing of Biodiversity Data on the Internet". In: *PLOS ONE* 9.8, e102623. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0102623. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0102623.

Roskov, Y., L. Abucay, T. Orrell, D. Nicolson, C. Flann, N. Bailly, P. Kirk, T. Bourgoin, R. DeWalt, W. Decock, and A. De Wever (n.d.). *Species 2000 & ITIS Catalogue of Life, 2016 Annual Checklist. Digital Resource At*. Species 2000: Naturalis, Leiden, the Netherlands. ISSN 2405-884X. URL: http://www.catalogueoflife.org/annual-checklist/2016.

Rousseeuw, P. J. (1987). "Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis". In: *Journal of Computational and Applied Mathematics* 20, pp. 53–65. ISSN: 0377-0427. DOI: 10.1016/0377-0427(87)90125-7. URL: http://www.sciencedirect.com/science/article/pii/0377042787901257.

Schalk, P. H. (1998). "Management of Marine Natural Resources through by Biodiversity Informatics". In: *Marine Policy* 22.3, pp. 269–280. ISSN: 0308-597X. DOI: 10.1016/S0308-597X(98)00013-X. URL: http://www.sciencedirect.com/science/article/pii/S0308597X9800013X.

Schreiber, J. (2018). "Pomegranate: Fast and Flexible Probabilistic Modeling in Python". In: *Journal of Machine Learning Research* 18.164, pp. 1–6.

Schuettpelz, E., P. Frandsen, R. Dikow, and L. Dorr (2017). "Applications of Deep Convolutional Neural Networks to Digitized Natural History Collections". In: *Biodiversity Data Journal* 5, e21139. ISSN: 1314-2828. DOI: 10.3897/BDJ.5.e21139. URL: https://bdj.pensoft.net/articles.php?id=21139.

Scoble, M. J. (2008). "Networks And Their Role In E-Taxonomy". In: *The New Taxonomy*. Ed. by Q. J. Wheeler. Systematics Association Special Volumes. CRC Press, pp. 19–31. ISBN: 978-0-8493-9088-3.

Scotland, R. W. and A. H. Wortley (2003). "How Many Species of Seed Plants Are There?" In: *TAXON* 52.1, pp. 101–104. ISSN: 1996-8175. DOI: 10.2307/3647306. URL: https://onlinelibrary.wiley.com/doi/abs/10.2307/3647306.

Secretariat of the Convention on Biological Diversity (2010). *Guide to the Global Taxonomy Initiative*. URL: https://www.cbd.int/doc/publications/cbd-ts-30.pdf.

Sekara, V., P. Deville, S. E. Ahnert, A.-L. Barabási, R. Sinatra, and S. Lehmann (2018). "The Chaperone Effect in Scientific Publishing". In: *Proceedings of the National Academy of Sciences* 115.50, pp. 12603–12607. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1800471115. URL: https://www.pnas.org/content/115/50/12603.

Seltmann, K. C., Z. Pénzes, M. J. Yoder, M. A. Bertone, and A. R. Deans (2013). "Utilizing Descriptive Statements from the Biodiversity Heritage Library to Expand the Hymenoptera Anatomy Ontology". In: *PLoS ONE* 8.2. Ed. by C. S. Moreau, e55674. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0055674. URL: http://dx.plos.org/10.1371/journal.pone.0055674.

Shorthouse, D. and R. Page (2019). "Quantifying Institutional Reach Through the Human Network in Natural History Collections". In: *Biodiversity Information Science and Standards* 3, e35243. ISSN: 2535-0897. DOI: 10.3897/biss.3.35243. URL: https://biss.pensoft.net/article/35243/.

Siracusa, P. de, L. M. R. Gadelha, and A. Ziviani (2018). "On the Social Structure behind Biological Collections". In: *bioRxiv*, p. 341297. DOI: 10.1101/341297. URL: https://www.biorxiv.org/content/early/2018/06/08/341297.

Smith, L. and R. Smith (1967). "Itinerary of William John Burchell in Brazil, 1825-1830". In: *Phytologia* 14.8, pp. 492–505. URL: https://www.biodiversitylibrary.org/page/14903934.

Stafleu, F. and R. Cowan (1976). *Taxonomic Literature: A Selective Guide to Botanical Publications and Collections with Dates, Commentaries and Types. Vol. 1–7*. 7 vols. Regnum Vegetabile. Utrecht: International Association for Plant Taxonomy.

Suhrbier, L., W.-H. Kusber, O. Tschöpe, A. Güntsch, and W. G. Berendsohn (2017). "AnnoSys—Implementation of a Generic Annotation System for Schema-Based Data Using the Example of Biodiversity Collection Data". In: *Database* 2017. DOI: 10.1093/database/bax018. URL: https://academic.oup.com/database/article/doi/10.1093/database/bax018/3074788.

Sweeney, P. W., B. Starly, P. J. Morris, Y. Xu, A. Jones, S. Radhakrishnan, C. J. Grassa, and C. C. Davis (2018). "Large–Scale Digitization of Herbarium Specimens: Development and Usage of an Automated, High–Throughput Conveyor System". In: *TAXON* 67.1, pp. 165–178. ISSN: 1996-8175. DOI: 10.12705/671.10. URL: https://onlinelibrary.wiley.com/doi/abs/10.12705/671.10.

SYNTHESYS - an Integrated European Infrastructure for Researchers in the Natural Sciences (2019). URL: https://www.synthesys.info/about-synthesys.html.

Taxonomic Names and Concepts Interest Group (2006). *Taxonomic Concept Transfer Schema (TCS), Version 1.01*. Biodiversity Information Standards (TDWG). URL: http://www.tdwg.org/standards/117.

Thessen, A., M. Woodburn, A. Ariño, C. Flann, N. Nicolson, D. Shorthouse, and D. Koureas (2016). *Joint RDA/TDWG Working Group on Metadata Standards for Attribution of Physical and Digital Collections Stewardship Case Statement*. URL: https://www.rd-alliance.org/group/metadata-standards-attribution-physical-and-digital-collections-stewardship/case-statement.

Thiers, B. (2018). *The Worlds Herbaria 2017: A Summary Report Based on Data from Index Herbariorum*. New York Botanical Garden. URL: http://sweetgum.nybg.org/science/docs/The_Worlds_Herbaria_2017_5_Jan_2018.pdf.

Thiers, B. (continuously updated). *Index Herbariorum: A Global Directory of Public Herbaria and Associated Staff. New York Botanical Garden's Virtual Herbarium.* New York. URL: http://sweetgum.nybg.org/science/ih/.

Triebel, D., P. Scholz, T. Weibulat, and M. Weiss (2011). "An Online Thesaurus for Standard Bibliographic Data on Exsiccatae in Botany and Mycology". In: Biodiversity Informatics Standards. New Orleans, US.

Tucker, A. and D. Kirkup (2014). "Extracting Predictive Models from Marked-Up Free-Text Documents at the Royal Botanic Gardens, Kew, London". In: *Advances in Intelligent Data Analysis XIII*. Ed. by H. Blockeel, M. van Leeuwen, and V. Vinciotti. Lecture Notes in Computer Science 8819. Springer International Publishing, pp. 309–320. ISBN: 978-3-319-12570-1 978-3-319-12571-8. URL: http://link.springer.com/chapter/10.1007/978-3-319-12571-8_27.

Tulig, M., N. Tarnowsky, M. Bevans, Kirchgessner, Anthony, and B. M. Thiers (2012). "Increasing the Efficiency of Digitization Workflows for Herbarium Specimens". In: *ZooKeys* 209, pp. 103–113. ISSN: 1313-2989. DOI: 10.3897/zookeys.209.3125. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3406470/.

Turland, N. (2013). *The Code Decoded a User's Guide to the International Code of Nomenclature for Algae, Fungi, and Plants*. 1st ed. Regnum Vegetabile. Königstein: Koeltz Scientific Books. 169 pp. ISBN: 978-3-87429-433-1 3-87429-433-1.

— (2019). *The Code Decoded*. 2nd ed. Vol. 1. Advanced Books. DOI: 10.3897/ab.e38075. URL: https://ab.pensoft.net/article/38075/.

Turland, N., J. Wiersema, F. Barrie, W. Greuter, D. Hawksworth, P. Herendeen, S. Knapp, W.-H. Kusber, D.-Z. Li, K. Marhold, T. May,

J. McNeill, A. Monro, J. Prado, M. Price, and G. Smith, eds. (2018). *International Code of Nomenclature for Algae, Fungi, and Plants*. Vol. 159. Regnum Vegetabile. Koeltz Botanical Books. ISBN: 978-3-946583-16-5. DOI: 10.12705/Code.2018. URL: https://www.iapt-taxon.org/nomen/main.php.

Utteridge, T. M. A. and R. P. J. de Kok (2006). "Collecting Strategies for Large and Taxonomically Challenging Taxa". In: *Reconstructing the Tree of Life: Taxonomy and Systematics of Species Rich Taxa*. Ed. by T. R. Hodkinson and J. A. Parnell. 1st ed. Boca Raton: CRC Press, pp. 297–304. ISBN: 978-0-429-12809-7. URL: https://www.taylorfrancis.com/books/9780429128097.

Van Rossum, G. and F. L. Drake Jr (1995). *Python Tutorial*. Centrum voor Wiskunde en Informatica Amsterdam, The Netherlands.

Van Oever, J. P. and M. Gofferjé (2012). "From Pilot to Production: Large Scale Digitisation Project at Naturalis Biodiversity Center". In: *ZooKeys* 209, pp. 87–92. ISSN: 1313-2989. DOI: 10.3897/zookeys.209.3609. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3406468/.

Walls, R. L., J. Deck, R. Guralnick, S. Baskauf, R. Beaman, S. Blum, S. Bowers, P. L. Buttigieg, N. Davies, D. Endresen, M. A. Gandolfo, R. Hanner, A. Janning, L. Krishtalka, A. Matsunaga, P. Midford, N. Morrison, É. Ó. Tuama, M. Schildhauer, B. Smith, B. J. Stucky, A. Thomer, J. Wieczorek, J. Whitacre, and J. Wooley (2014). "Semantics in Support of Biodiversity Knowledge Discovery: An Introduction to the Biological Collections Ontology and Related Ontologies". In: *PLOS ONE* 9.3, e89606. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0089606. URL: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0089606.

Wang, Z., H. Dong, M. Kelly, J. Macklin, P. Morris, and R. Morris (2009). "Filtered-Push: A Map-Reduce Platform for Collaborative Taxonomic Data Management". In: *2009 WRI World Congress on Computer Science and Information Engineering*. Vol. 3, pp. 731–735. DOI: 10.1109/CSIE.2009.948.

Waskom, M., O. Botvinnik, P. Hobson, J. B. Cole, Y. Halchenko, S. Hoyer, A. Miles, T. Augspurger, T. Yarkoni, T. Megies, L. P. Coelho, D. Wehner, cynddl, E. Ziegler, diego0020, Y. V. Zaytsev, T. Hoppe, S. Seabold, P. Cloud, M. Koskinen, K. Meyer, A. Qalieh, and D. Allan (2014). *Seaborn: V0.5.0 (November 2014)*. Zenodo. DOI: 10.5281/zenodo.12710. URL: https://zenodo.org/record/12710.

Wheeler, Q. D., S. Knapp, D. W. Stevenson, J. Stevenson, S. D. Blum, B. M. Boom, G. G. Borisy, J. L. Buizer, M. R. De Carvalho, A. Cibrian, M. J. Donoghue, V. Doyle, E. M. Gerson, C. H. Graham, P. Graves, S. J. Graves, R. P. Guralnick, A. L. Hamilton, J. Hanken, W. Law, D. L. Lipscomb, T. E. Lovejoy, H. Miller, J. S. Miller, S. Naeem,

M. J. Novacek, L. M. Page, N. I. Platnick, H. Porter-Morgan, P. H. Raven, M. A. Solis, A. G. Valdecasas, S. Van Der Leeuw, A. Vasco, N. Vermeulen, J. Vogel, R. L. Walls, E. O. Wilson, and J. B. Woolley (2012). "Mapping the Biosphere: Exploring Species to Understand the Origin, Organization and Sustainability of Biodiversity". In: *Systematics and Biodiversity* 10.1, pp. 1–20. ISSN: 1477-2000. DOI: 10.1080/14772000.2012.665095. URL: http://dx.doi.org/10.1080/14772000.2012.665095.

Wieczorek, J., D. Bloom, R. Guralnick, S. Blum, M. Döring, R. Giovanni, T. Robertson, and D. Vieglais (2012). "Darwin Core: An Evolving Community-Developed Biodiversity Data Standard". In: *PLoS ONE* 7.1, e29715. DOI: 10.1371/journal.pone.0029715.

Wilkinson, M. D., M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. d. S. Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. G. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. C. Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, and B. Mons (2016). "The FAIR Guiding Principles for Scientific Data Management and Stewardship". In: *Scientific Data* 3, p. 160018. ISSN: 2052-4463. DOI: 10.1038/sdata.2016.18. URL: https://www.nature.com/articles/sdata201618.

Willis, C. G., E. R. Ellwood, R. B. Primack, C. C. Davis, K. D. Pearson, A. S. Gallinat, J. M. Yost, G. Nelson, S. J. Mazer, N. L. Rossington, T. H. Sparks, and P. S. Soltis (2017). "Old Plants, New Tricks: Phenological Research Using Herbarium Specimens". In: *Trends in Ecology & Evolution* 32.7, pp. 531–546. ISSN: 0169-5347. DOI: 10.1016/j.tree.2017.03.015. URL: http://www.sciencedirect.com/science/article/pii/S0169534717300939.

Winker, K. and J. J. Withrow (2013). "Natural History: Small Collections Make a Big Impact". In: *Nature* 493.7433, pp. 480–480. ISSN: 0028-0836. DOI: 10.1038/493480b. URL: http://www.nature.com/nature/journal/v493/n7433/full/493480b.html.

Zika, P. (2014). "A New Species of Stonecrop (Sedum Section Gormania , Crassulaceae) from Northern California". In: *Phytotaxa* 159.2, pp. 111–121. ISSN: 1179-3163. DOI: 10.11646/phytotaxa.159.2.5. URL: https://biotaxa.org/Phytotaxa/article/view/phytotaxa.159.2.5.