A lexicographic approach to profiling nomenclature usage in the biodiversity literature

Sandra Young

School of Computing, Mathematics and Engineering
University of Brighton

Principal supervisor: Roger Evans (University of Brighton)
Secondary supervisors: Gulden Uchyigit (University of Brighton),
Harald Schneider (Natural History Museum)

A Thesis Submitted for the Degree of
PhD in Computational Linguistics

· 2020 ·

Abstract

The digital age brought with it many opportunities for data analysis, as well as many challenges for data integration and management. Ontologies are a popular data representation structure because of their inference properties, used in searching and analysis. However, ontologies must assume a defined view of the world, or a domain, which may ignore the information stored within data or could even impose an unsuitable structure (conceptual model) onto the information. The area of biodiversity has a very specific problem in this regard. Biological taxonomy is, by nature, fluid, changing and multiple. Gaps in knowledge, evolution and differences of opinion as to the classification of species mean that there is no single agreed taxonomy, and inconsistent scientific nomenclature usage is widely tolerated in the biodiversity literature. The importance of the nomenclature and taxonomies for accurately communicating biodiversity information, coupled with the difficulty of modelling such information means that there are numerous efforts to create comprehensive ontologies and other knowledge representation resources of taxonomic and other biodiversity data. However, despite these efforts many of the resources are still fragmented, incomplete and work on a premise of imposing a single, external hierarchy onto the data mapped.

The literature review has revealed that, despite continued recognition of both the inconsistency and plurality of the scientific nomenclature, and of the importance of a proper understanding of the intended meaning of these terms when used, there has been no systematic empirical analysis of nomenclature usage in the biodiversity literature to profile meaning. My research project has applied a combined design science and corpus lexicographic approach to the problem, based on the "Word Sketch" analysis technique provided by the "Sketch Engine" lexicographic analysis tool. This research study has adapted Word Sketches to define a method by which nomenclature usage can be mapped and compared against ontological or other knowledge representation resource information, and across corpora to check for stability of usage and meaning. The method was first developed and tested with two test corpora (on the subject of freshwater fish) against an authoritative knowledge representation resource and was then evaluated through application to three nomenclature profile studies.

The method developed aims to serve people working in biodiversity by helping them to choose a suitable knowledge resource onto which to map specific bodies of data, to identify issues when integrating data, and to identify problems or inconsistencies in data or in knowledge representation resources that need to be reviewed, as well as mapping nomenclature use change

across language, domain, time, author, publication, etc. It could also be developed into a tool to aid novices in taxonomy to identify where multiple variants refer to the same species.

> *A word is an arbitrary label - that's the foundation of linguistics. But many people think otherwise. They believe in word magic: that uttering a spell, incantation, curse, or prayer can change the world. Don't snicker: Would you ever say, 'Nothing has gone wrong yet' without looking for wood to knock?...*
>
> Steven Pinker

Contents

List of Figures

List of Tables

Acknowledgments

First and foremost, I would like to give my very sincerest thanks to Dr Roger Evans, my primary supervisor, for his continued support and patience over the last four years. His detailed and precise notes, thoughtful comments and calm manner guided me throughout the doctoral process, and would steer me back on track on the numerous occasions I was ready to walk away. I would also like to thank my other supervisors, Gulden Uchyigit and Harald Schneider for their input and support throughout the doctorate.

Thank you to Myrsini, Ayad and Nuha and the other PhD students of Cockcroft 4.11 for the companionship when it seemed like submission time would never come, for being a circle with whom I could share my doubts, fears and frustrations and for making me realise I was not the only one.

Thank you to my housemates Sophie, Simon, Joy and Eddie for putting up with my craziness particularly over the first months of 2020 and for helping me eat the cakes I baked in moments of frustration, and to Gonul for your company on walks when I couldn't socialise in any other way. Thank you to Gabi for her thoughtful proofreading and presentation practice and Jamie for getting me off the blocks with the programming. That obstacle seemed insurmountable at the beginning of this journey and now seems so long ago. Dan, thank you for your patience although I know at times you didn't think this was worth it. And finally, thank you to all my family, and other friends not mentioned here but without whom I wouldn't have got here in the first place. And now to say, it is (nearly) done, I am back.

Chapter 1

Introduction

By way of introduction, it seemed appropriate to explore the idea of conceptual relativity and its relevance to the problem addressed in this thesis, by means of a discussion about the concept of species. On the surface, and in the way we describe species such as “humans”, “elephants” or “trout” in everyday conversation, species seem to be well-defined, discrete entities. However, the reality is considerably more blurred, for both biological and philosophical reasons. In reality, there is no universally accepted definition of a species, nor has there ever been. Schulz [191] describes the classification process of Linnaeus, the father of scientific nomenclature, as being driven principally by the “criterion of similarity between organisms and organisms groups” [191]. So essentially, a species can be thought of as a group of organisms that are sufficiently different from other organisms to be defined as such; but how different is “sufficiently different”?

In Linnaeus’ time, the idea of “sufficiently different” was based on nothing more than physical traits, the collection of which were considered by one taxonomist or another to be indicative of a different species. The differences in appearance between different specimens of the same species were thought to be “accidents” of life or nature. Nowadays, with genetic analysis, we know otherwise. Genetic material comprises a continuum, rather than discrete sets that can be easily separated. As Sandra Knapp, tropical botanist and taxonomist at the Natural History Museum, London, described on “In Our Time” at the end of 2019, species can be thought of as the bumps in a carpet, relating to “our best estimate of distribution of variation in nature” [182]. Steve Jones, Senior Research Fellow in Genetics at University College London, provided a very visual image of this by describing species before and after genetic analysis as starting by counting peas and ending up with pea soup [182]. Essentially, there is no discrete dividing line between one species and another. Species must be defined from a particular perspective, which results in multiple possible definitions.

With this in mind, there are a number of different approaches or perspectives to defining what a species is. The biological species concept is perhaps the most commonly known definition of a species. This concept, as defined by Mayr [142], encapsulates a group of organisms that can reproduce and produce fertile offspring. However, this definition is limited to looking at organisms that are alive at the same time, ignores issues with infertile members and is sometimes hard to corroborate [98, 139].

The evolutionary quality of species means that every single example of an organism is slightly different, something that with the development of genetic analysis technology became much more apparent. The phylogenetic species concept describes the "concept of a species as an irreducible group whose members are descended from a common ancestor and who all possess a combination of certain defining, or derived, traits" [54]. The phylogenetic species concept is less restrictive than the biological concept as it does not have the same restrictions as regards breeding. These are just two definitions from many more: Mayden [141] identified 24 different definitions. Now, in 2020, still there is no one definition that everyone agrees on.

As we have seen, the term "species" is a very good example of a word used to describe a real-world thing that at first sight appears to be a relatively static, clear and defined concept, yet the reality of the science is far from that. Genetic analysis has identified to what extent this delimitation of species is a human construct. The question of what a species is represents a suitable starting point to look at the issue of classification, and how arguably the same thing can be classified in different ways depending on the perspective from which the classification is made.

It must be emphasised that none of the different classifications are necessarily "invalid", certainly not in a communicative sense. In everyday language the term species is commonly understood, and is useful in speech and to share meaning. The term is fit for purpose, even if when we talk about species we are not necessarily thinking about the nebulous nature of species on a scientific level.

To demonstrate this point, we can consider some examples of classification. In much of our everyday language the categorisation of things seems fixed, obvious, without thinking about the implicit choices we have made to categorise things according to the context in which they are operating. Yallop [226] draws our attention to the relativity of classification by continuing with the species analogy, when he highlights that the scientific nature of the classification of species into genera, families, orders and higher is based on specific qualities (genes, or appearance, or reproductive powers) which have been scientifically defined according to specific requirements. He counters this with our categorisation of species in our normal everyday conversation, giving the example of the term "pines". This term is used frequently for trees that are not classified under the Pinus genus but that have the same characteristic pine needles that to a layperson look like pines. This vernacular usage may in fact originate from an original, erroneous, classification, which was maintained in everyday language because of habit and for its communicative usefulness. In fact, as we shall see, the scientific classification is also very fluid, for the reasons outlined above, and shall be a focus of this thesis. Straying from the subject of species but staying within biology, tomatoes scientifically are a fruit, but in our day-to-day lives we usually categorise them as a vegetable or salad item, depending on how they are going to be eaten. Although more or less scientifically accurate, these classifications are no less valid when taken in the context in which they are used. Effective communication is the key, which is achieved when using words appropriately to the context. These few examples can serve to provide an understanding as to how the classification of entities (species, words, meanings) can be lumped, split or changed according to the perspective from which we approach a problem.

The concept of species is just one of millions of other examples. Sand, for example, could be classified as a construction material or eroded rock, or an element of a beach. In each instance different ideas spring to mind about the same physical object (in the examples given). We often think of words having specific, fixed definitions but meaning is a lot more elusive than that. Even the definition of what a "word" is can be complicated [87]. Meaning is inherently context-based and a lot more fluid than we generally think. It depends on the boundaries that we set according to said context. Lexicographers and other linguists became acutely aware of this with the arrival of computational corpus linguistics, as it allowed them to analyse how words are actually used in different contexts [19, 62, 69, 115]. The capacity to analyse large amounts of data revealed that words did not emanate discrete, fixed meanings, but instead that meaning depended on context and the patterns identified existed on a continuum, which had to be divided from a particular perspective to delineate separate senses within a word's meaning. The concept of splitting and lumping meaning is common in lexicography, in which decisions are made as to how detailed the division of meaning should be [62].

As humans we traverse these different contexts and nuances in meaning generally with ease, understanding that tomatoes can be at once considered a salad item or vegetable in many culinary contexts, while scientifically being a fruit. Humans are good at abstraction and common sense reasoning, which is why these context-dependent classifications are easy for humans, but not so for computers. Computer processing has become increasingly important because of the capacity of computers to analyse large bodies of data [135, 201]. However, the ambiguities of human language cause a lot of problems for the accurate integration and analysis of said data. A computer cannot reason to truly understand whether one word means one thing or another in a given context, it has to be told. The difficulties computers face in activities such as natural language processing or machine translation are related to their inability to reason in this way.

There are different approaches to how computers are "instructed" to understand natural language. Artificial Intelligence (AI) tries to emulate a human way of thinking through the use of neural networks in which the computer itself devises rules according to patterns it identifies by way of processing huge amounts of statistical data. Neural networks allow for reasoning on a certain level in this way, but they still develop a very narrow intelligence: highly adapted to the specific domain or task at hand [227]. The problems with this approach lie in their susceptibility to bias where there have been imperfections in the collection of the training data, or will tend to exaggerate bias where there are biases in our society [126]. There is also the risk of so-called false discoveries [134], in which testing appears to give positive results, only to have identified an "interesting" pattern where there is none. This highlights the importance of transparency in data being used in automatic integration tasks. When using computers to process natural language it is essential that there is an understanding of what is contained in the data and how the models are formed: otherwise the black box in which they function hangs a large, worrying question mark over the validity of the results.

Given the need to understand natural language, ways of standardising and labelling human language documents have been developed within the scope of standardisation frameworks and within Natural Language Processing (NLP), to the point of training models using algorithms that learn and improve and evolve. These approaches harness the powers of computers to identify patterns and use these to their advantage. To then store and analyse this data, knowledge representation structures called ontologies have been developed. Ontologies allow computers to effectively reason across data sets by means of the logical structures that provide information about relationships between different classes and properties.

While this method of knowledge organisation aims to and often does overcome many issues in the context of natural language ambiguity, ontologies are still limited in scope in comparison with the vastness of natural language. Their explicit definition of a domain structure is both their advantage and disadvantage. Within this context the ambiguity of natural language becomes even more complex. Ontologies try to go beyond natural language but in the end a classification, as we saw at the beginning of this chapter, needs to take a perspective and each perspective will provide a different organisational structure. So, what happens if an ontology with one internal conceptual model is applied to information that is governed by another conceptual model? Would the result be the erroneous imposition of the ontology's conceptual model onto said information? Are there other cases in which ontologies are erroneously excluding data because of the defined conceptual model? To tackle these issues we need to identify methods by which to evaluate whether certain terms are 1) being used to express the same concept and 2) whether this concept changes between domains or other definition of units. There really are two questions in this regard:

- How can we be sure that our systems are accurately tagging or labelling data?
- How can we know that the ontologies are modelling things correctly/not imposing inaccurate information?

It is on this backdrop that the following PhD thesis is based. Increased automation of data analysis has enabled us to achieve things we never thought possible. However, we need to know what we are analysing is accurate and that the classification structure being applied is the right one, which is not a trivial task. The research project focuses on the domain of biodiversity, for reasons which will be explained in the following paragraphs, and in more depth in Chapter 2, the literature review.

Issues relating to accurate computer integration and analysis of data have left the biodiversity informatics community working to tackle the issues of knowledge fragmentation, dataset size and heterogeneity present in the field [77, 85, 149, 178, 216]. There are many initiatives underway and in use, focusing on the standardisation of concepts and vocabulary [15], the creation and integration of databases [178, 225] and the opening of data access [85, 170].

Two issues have been identified in this process that are of particular interest to this research. Both are related to questions of the identification and stability of concepts versus terminology usage. Firstly, the biodiversity informatics community is increasingly turning to ontologies to overcome the knowledge organisation and discovery problems that arise from the obstacles of fragmentation, inconsistent terminology usage and nomenclature issues [32, 137, 169, 216]. Ontologies aim to pin down the concepts at the heart of different domains and the relationships between them. However, as we saw in the first part of the chapter, the words we use to describe concepts are constantly changing, multiple and ambiguous [32, 41, 216]. As explained by Thessen [206], "clearly representing the natural language descriptions of phenotypes and environments with a set of ontologies is difficult, because natural language, while highly expressive, is often semantically ambiguous and reliant on context". The ClearEarth project [204], of which Thessen is a part, noted the difficulty in identifying different concepts, where specific terms were used across domains for different purposes [205]. The same was identified as regards the conceptualisation of the term ecological niche, which is used at different times and by different authors to represent different concepts [217]. These issues highlight problems in both terminology usage but also the difficulty in even arriving at a shared conceptualisation of a concept within a single domain, let alone across domains.

In other domains, to take a more empirical approach, corpus-based analysis has been used either alone or in conjunction with other statistical techniques in (semi-)automatic ontology learning to identify concepts and relations [4, 11, 27, 73, 194] from collections of texts themselves. Corpus-based analysis is also used in data-driven evaluation of ontologies, to check for coverage and accuracy of the concepts and relations within a specific domain [11, 26, 27]. In biodiversity the ClearEarth project is making moves towards the automation of ontology construction/population adapting existing algorithms used for the biomedical domain, but despite all the problems identified as regards conceptual stability, the researcher has found no efforts to perform a systematic empirical analysis of the literature to validate existing ontologies or to ascertain the conceptual stability of the terms in use.

The other issue is specifically related to the stability of concepts as regards scientific nomenclature and its vernacular equivalents. Many of the integration initiatives mentioned above use scientific names as the index for this information [109, 170, 203]. However, this causes problems of its own due to the multiple and fluid nature of biological taxonomies and, by extension, the scientific nomenclature. There is no one accepted biological taxonomy, but multiple ones which reflect not only different moments in time but also simultaneous differences in opinion [49, 68, 124, 170, 202]. The nomenclature is also multiple and reflects these changes [49, 109]. There have been numerous attempts at creating taxonomic databases that integrate the multiplicity of concepts versus names used to represent taxon concepts through the use of unique identifiers [61,124,153,172]. Different forms of knowledge organisation have also been attempted to overcome some of the difficulties this presents, such as ontologies [191, 192, 209] or forms of concept-oriented databases [17].

Despite all these efforts, and the recognised ambiguities in the use of the scientific nomenclature in the literature, there has been no empirical study to profile nomenclature usage in context. This is a big problem, as has been highlighted in the problems faced by those working to integrate taxonomic reference data [172], with issues including but not limited to incorrect assignment of names, incompleteness of names, synonymy, and even disagreements between experts about the identity of specimens in relation to the organisation of genera within the taxonomic hierarchy [49, 109]. If scientific names are being used as indexes for biodiversity data, misunderstandings about the concepts underlying these scientific names could be currently undermining these efforts. In [175], the author recognises the importance of mapping usage of the nomenclature in context are looking at semantically-enhanced journal articles [174]. However, to my knowledge there are no efforts underway to study this in unstructured legacy data.

Considering both the broader picture of knowledge integration and discovery across the biodiversity domain, and the more specific issue of using scientific names as the primary index for this information, being able to identify how these terms are being used is key. Developing a method by which the nomenclature usage within the literature could be systematically and empirically analysed would therefore be a great step towards a better understanding of commonalities and differences between the usage of these terms lie, as well as providing demonstrations of patterns of change, stability and ambiguity in their conceptualisation across resources and datasets.

The research aims to address the gap identified in the literature, to develop a method to perform a systematic, empirical analysis of the (in)consistency in the conceptualisation of the scientific nomenclature in the biodiversity literature. Such a method would make it possible to evaluate the compatibility of datasets, identify areas of greater intra- or inter-domain clarity or ambiguity and evaluate the validity of using ontologies in particular circumstances or for particular purposes.

To address this gap, the research takes a multidisciplinary perspective to extend the Word Sketch feature of Sketch Engine to profile the conceptualisation of species' names in the biodiversity literature. Word Sketches are a feature of the corpus query tool Sketch Engine, which provide a summary of a word's grammatical and collocational behaviour in context. Taking inspiration from McCarthy et al.'s [144] paper which added semantic annotation of WordNet to Sketch Engine's Word Sketches, this research will add appropriate semantic annotation as well as employ other aspects of corpus analysis and NLP tools readily available in order to extend the capacities of Word Sketches into this new field.

Lexicography has a well-founded history in the conceptual mapping of words in context for dictionary making and which makes it a suitable candidate to explore this avenue. Word Sketches were considered a suitable tool to employ given the empirical nature of the process, in which concepts are identified through their description in the body of data, rather than through an externally imposed hierarchy such as an ontology. This facilitates the identification of the hierarchy as it exists within the corpora for comparison with existing resources. We have seen how hierarchies in data can be identified through Word Sketches in the results of the Ecolexicon project [57, 128].

Corpus-based analysis more generally is also often used in automatic and semi-automatic ontology creation [4, 11], which supports the use of this methodology in the research. However, to my knowledge it has not been used to create contrasting profiles of the hierarchy of species based on empirical data to perform an analysis of the conceptual (in)stability of these terms. I have also found no evidence of using these techniques to compare conceptual stability across different corpora in other domains. Corpus analysis is also used as an ontology evaluation technique [11]. Corpus-based analysis is applied in this thesis to profile the hierarchy of species in one or more corpora, which then forms the basis for comparison, which provides an empirical evaluation of the use of scientific nomenclature and vernacular variants in context, evaluates conceptual stability across corpora and in comparison with ontologies to provide evidence as to whether specific datasets are suitable for integration, or whether a particular knowledge representation resource follows an appropriate conceptualisation for mapping the data.

The aim of the research was "to employ computational lexicography and natural language processing techniques to identify, extract and group nomenclature according to its usage in the biodiversity literature and use contrasting corpora and existing knowledge representation structures to perform a systematic empirical analysis of these conceptualisations". This aim arose from the identification in the literature review of issues relating to the usage of nomenclature as a result of the multiple and changing condition of biological taxonomies. This has been framed within the issue of automatic integration of data, or knowledge, and queries as to the problems that may be perpetuated should incorrect data be integrated or data excluded erroneously.

At the beginning of the PhD, the aim included exploring the characterisation of trophic interactions ("what eats what") in the biodiversity literature, through the application of these same techniques. This aim was originally explored in the pilot project in relation to trophic interactions ("what eats what", but was removed from subsequent steps of the research because of the complexity of achieving the aim in relation to the relations between species' name mentions specifically. Further work into interactions will be considered in the future. The exact wording of the aim also evolved throughout the course of the research, although the underlying aim did not change, more an understanding of how to present the problem. The evolution of the aims, objectives and research questions can be found in the appendix [add to appendix when finalised].

The research questions related to the final aim of the research were therefore:

1. How does empirical corpus-based analysis use the linguistic evidence in the biodiversity literature to model the hierarchical relationship between species?

2. How does the knowledge representation model extracted in research question one compare with other knowledge representation approaches currently being employed?

3. How do conceptualisations between different corpora vary quantitatively (number or trends of mentions) and qualitatively (contextually or links between different mentions)?

The objectives of the research are:

- model the hierarchy of relations between nomenclature reference/units of nomenclature as used in a specific corpus (by extracting the relevant information) (RQ1)

- create a graph/tree hierarchy image of this model to compare to the ontological structure for validation and evaluation purposes (RQ2)

- produce a technical validation and evaluation method to compare the relations identified for precision, recall (quantitative measures) and differences (quantitative and qualitative measures) between the different expressions of knowledge (RQ2)

- perform comparisons between the hierarchies extracted between different corpora and ontologies of choice to evaluate the conceptual stability of nomenclature usage (RQ3)

The method developed aims to serve people working in biodiversity by helping them to choose a suitable knowledge representation resource onto which to map specific bodies of data, to identify issues when integrating data, and to identify problems or inconsistencies in data or in knowledge representation resources that need to be reviewed, as well as mapping nomenclature use change across language, domain, time, author, publication, etc. It could also pave the way for future work in other areas relating to avoiding erroneous data integration and could serve to shed light on specifics relating to nomenclature usage yet to be captured in integrated databases. Finally, the method could be used in adapted forms to identify new terminology, used to map conceptualisations within corpora that could be used either for the basis of ontology-building or an alternative form of knowledge representation. As identified in the expert evaluation phase of this research, these techniques could serve as a tool for people new to the area of taxonomy, to help marry variants and raise awareness about the multiplicity of variants in existence for different taxa, and therefore minimise confusion. Furthermore, the techniques described in this thesis, while being adapted specifically to scientific nomenclature and other variants, could also be applied to other domains, such as inter-lingual terminology and other specialist domains.

The thesis is set out as follows:

Chapter 1: Introduction This chapter.

Chapter 2: Background Chapter 2 consists of an overview of the literature in relation to the thesis. It considers the issue of data in the digital age, the opportunities and the challenges that this presents, particularly as regards the accurate integration of data and the interplay between data and knowledge. It goes on to look at existing cross-domain efforts to respond to these challenges, along with persisting obstacles there. The focus then turns to the issue specifically within the domain of biodiversity, firstly considering challenges and existing efforts relating to the domain and data integration in general, then specifically focusing on the issue of scientific nomenclature within this equation. The final section of the chapter sets out previous research in the areas of corpus linguistics and lexicography, describing the history of the fields, and related previous research to make the argument as to the suitability of this approach to address the problem identified.

Chapter 3: Methodology and methods Chapter 3 sets out the methodological basis for the research and then the methods used to perform the research described in this thesis. As further described in the chapter, a research design method approach was taken in this research to allow for iterative learning throughout the research process, which was used to feed back into the development of a methodology, and improve and guide the development of the research. For this reason the results sections are split into four: one relating to the preliminary exploration of the data and approach (Phases 0 and 1), another which describes the technical evaluation of the method designed (Phase 2), followed by the application of the method (Phase 3) and finally an external evaluation of the method by domain experts (Phase 4). The argument for using this type of research design is set out here, along with the considerations necessary when designing research within the field of corpus linguistics and lexicography. The methods describe the design of the different phases of the research and any choices made as regards tools.

Chapter 4: Phases 0 and 1 - Relation hierarchy model and data framing Chapter 4 sets out the preliminary results of the research, which focused on the pilot exploration of the data, and then subsequent efforts which looked at how different filtering parameters, and different ways of framing the taxonomic references (considering multi-word terms as multiple entities or as a single, unified entity) affected the output from the dataset. This constituted the main part of the design cycle of the research.

Chapter 5: Phase 2 - Method validation and technical evaluation This chapter sets out the technical validation and evaluation of the methods developed, to assess the validity and evaluate the technical efficacy of the methods applied. This was necessary given the methodological focus of the research, to ensure that the method developed was indeed capable of answering the questions it set out to do.

Chapter 6: Phase 3 - Nomenclature profiling studies Having evaluated the method, the method was applied to a real-life situation in which various knowledge representation resources were evaluated and compared. Then the method was applied to specific species' profiles according to these knowledge representation resources across two different corpora. The method developed in this research was used to evaluate the corpora against the existing knowledge resources and each other to provide evidence as to the stability or lack thereof of concepts and terms by means of the profile studies performed and develop guidelines for anyone using the methods developed in this thesis in the future.

Chapter 7: Phase 4 - Outreach and focus group To provide further weight to the validity of the research, a focus group/outreach session was held with experts who use scientific nomenclature in their working life. The session focused on exploring ideas relating to term ambiguity and clarity, use of knowledge representation resources and also presented the research described in this thesis for feedback and comments. Full analysis of the outcomes is provided in this chapter. The evaluation was intended to better understand the data involved, my interpretation of my results and possible applications of the method from the perspective of domain experts and relates to the rigour cycle in the design science structure.

Chapter 8: Discussion and conclusions The discussion and conclusions chapter brings together the findings outlined in previous chapters and discusses their applications and future work, ending with the final conclusions resulting from the research and aligns them with the aims and objectives of the thesis, drawing the thesis to a close.

Chapter 2

Background

The introduction has provided a background as to the relativity of meaning in language. Meaning may often seem fixed and clear, but it is actually often ambiguous and fluid. While humans can process information in a way which overcomes this ambiguity, applying our knowledge and deductive powers to accurately process information in a contextually appropriate way, computers need different sorts of clues when dealing with ambiguous data. The definition of meaning and classification of objects must be made explicit, which means that certain choices must be made. Neither the definition of meaning nor the classification of objects are trivial tasks. The subsequent efforts in integration further complicate this matter. This chapter will look specifically at the research pertaining to the problem tackled in this thesis, firstly on a general level, then focusing on the domain of biodiversity. The second half of the review will present the background relating to the approach taken in this thesis to tackle the issues identified, and the argument as to why this is a suitable choice of approach.

2.1 Data in the digital age

In the digital age, the amount of data we have available is growing exponentially [11, 50, 135]. Today this data firstly must be properly archived and subsequently shared to make the most of the information held within [45]. For this reason, the importance of data management, and perhaps more importantly, knowledge management, is key to being able to make the most of it. As described by [7], knowledge is data plus an interpretation of the meaning of said data. Otherwise there is a risk of being overwhelmed by data we cannot interpret or utilise correctly. While this task used to be primarily performed by humans, because of the sheer amount of data produced these days, it is a task that must be performed by computers. This has resulted in efforts in many different fields that work towards the successful processing of data in order to produce knowledge. However, the management, interpretation and representation (classification/categorisation) of this data is not a trivial task. Data is not produced in one homogeneous, easy to integrate style. Firstly it is necessary to understand the different sorts of data that might be integrated, then other aspects of the data that needs to be standardised to be able to do so.

2.2 What data?

Firstly it is necessary to define what we mean by data and some different sorts of data that might be there. As described in the introduction, while humans use their knowledge of natural language to process information, computers need algorithms or data structures to instruct them on how to categorise information and process it to discern meaning [2,37,50,72,135]. For these purposes, data is often divided into three types: structured, semi-structured and unstructured. As the introduction described in relation to other aspects of our world and language, there is no universally accepted definition of exactly what constitutes one type of data or another. However, for the purpose of this thesis the following definitions apply:

- Structured data is the easiest form of data for computers to handle. Structured data usually comes in the form of tables or databases, in which all the information is clearly categorised for processing.

- Semi-structured data comprises, as the name suggests, data which possesses some form of structure. Semi-structured data has some form of markup that instructs computers as to the meaning of sections of the data. Further examples will be given in subsequent sections.

- Unstructured data comprises all other data. It can include natural, narrative language or formats such as videos or pictures. In the case of this thesis it only refers to natural language texts. This data is the most difficult and expensive data for computers to process. Unstructured data comprises about 80% [21,50,72] of data in existence today.

2.3 Data to knowledge

The introduction to this chapter highlighted the issue of converting data into knowledge. Unprocessed data alone means nothing - it must be interpreted in some way to derive meaning and be useful. All the above data types require some form of processing to be understood by computers, on a sliding scale of least (structured) to most (unstructured), because the various types have differing levels of explicit instruction within the data that helps computers to define and classify the information being presented.

The difficulty experienced in converting unstructured data into knowledge is what is sometimes called the "knowledge acquisition bottleneck" [138,194], which acknowledges the difficulty in quickly and accurately processing the exponentially growing sources of unstructured data. These difficulties arise from the ambiguity of natural language. In the case of structured, and to some extent, semi-structured data, computers have a degree of guidance as to the "meaning" of information or data because of the way it is categorised within the data structure. In the case of unstructured data there is no such information.

Natural language processing (NLP) is a central part of this process, being the means by which unstructured data is categorised and annotated for computer processing. Put simply, NLP is the process by which natural (human) language can be converted into structured data with which computers can work [44]. However, language is multiple, changing and ambiguous [206]. Ambiguity means multiple choices for computers when classifying and categorising data, and the issues this causes can be seen in the difficulty to perfect NLP techniques. There are many different steps: there are those more related to the syntactic and grammatical features of language, which involve tasks such as part-of-speech (POS) tagging and parsing. These help with the categorisation of sentence structure. This is still important in more semantic NLP tasks, such as word sense disambiguation (WSD), because structure and form cannot be totally separated from semantics [81]. There are also tasks more specifically related to semantics. These tasks are arguably the most complex [157, 183] and involve tasks such as named entity recognition (NER), named entity disambiguation (NED), named entity normalisation (NEN) and relationship extraction (RE). NER is the act of identifying specific entities, such as cities or people within unstructured texts. WSD focuses on the separation of word senses, where the same word may be used in different contexts for different meanings, and NED is a subset of this disambiguation process. NEN is focused on identifying groups of words with similar meanings. It has been shown that the syntactic and grammatical steps improve information extraction from unstructured data [59]. This is no surprise given that meaning cannot be completely separated from form (i.e. semantic information can be derived from form) [81], and some aspects of WSD or entity normalisation can be supported by knowing the grammatical identity of a word in a particular context, for example.

NLP systems can be described as knowledge-based, supervised, unsupervised or semi-supervised [44, 133, 183]. Knowledge-based systems use, as the name suggests, knowledge bases such as terminologies, thesauri, or ontologies as a basis for the tagging, particularly as regards semantics issues. The interplay between ontologies and NLP will be further explored in the Ontology section later in this review. Supervised systems require manually annotated corpora which are then processed using machine learning (ML) techniques of various types to "learn" the appropriate tags and salient features of the target data set. The semi-supervised and unsupervised systems represent steps down in the amount of manual input that goes into the design of the system, until no annotated data is required. The flip side to reducing manual input in the form of annotated corpora is a need for increasingly large data sets, with increasingly large training cycles [176, 183, 198]. The size of the data sets required also which makes it unsuitable for many specialist domains [183]. These systems are also somewhat of a black box and are susceptible to biases in the training data [23]. In situations where access to the original data must be maintained for empirical checking or other purposes, these automated methods are not suitable.

Despite great advances in all the areas described above, systems are either very labour intensive (i.e. supervised methods, which require detailed, hand annotated training data, or the production of detailed ontologies in the case of knowledge-based systems), or are very computer intensive, require huge amounts of data and are not as accurate as those which have knowledge in the form of either knowledge bases or manually annotated data [190, 222].

The complexity, multiplicity and ambiguity of natural language, mean that NLP systems are highly-specific and difficult to adapt to new areas [46, 205]. Structural ambiguity also continues to present obstacles to identifying the semantic meaning of different terms within similarly phrased sentences [78]. Word Sense Disambiguation is also highly contextual and knowledge-rich [157]. Natural language processing has advanced incredibly over the years, but it still comes up against many obstacles in the accurate disambiguation of meaning of terms and often has problems adapting a domain-specific algorithm to another domain [135, 190]. This is the reason why unstructured data is still the biggest problem for work on big data analytics.

Some domains, namely bioinformatics, have managed to successfully create quite extensive, accurate and well-trained systems of NLP for their domain [79, 205, 216], even if not complete [222]. However, this requires huge investment that many domains simply do not have. The non-transferability of the systems is such a big obstacle that while information extraction (an end goal of many NLP processes) in research is heavily focused on statistical approaches, in industry rule-based information extraction is still very important because of the fact that they are understandable, adaptable, can be easily used to integrate domain information, among other things [38].

Having looked at the general problem of converting natural language to a form in which computers can read it, the next section will consider the issue of how to ensure that this information can be correctly categorised and interpreted to integrate. This presents further problems, some of which are technical and others semantic in nature.

2.4 Data standardisation and integration

Given the heterogeneity of the data, how can it be integrated? As explained in the introduction, computers need a lot of signalling to be able to recognise data as one thing or another and be able to assign to specific boxes [183]. Although structured data is easier to handle than unstructured data (in the case of this thesis, narrative text with no markup), integration nevertheless requires standardisation of data formats, terminology usage (or alignment), and so on and so forth.

If data is not correctly standardised and integrated, final interpretation of the data will, at best, be limited, or at worst simply wrong [79]. As a result there are many efforts in the areas of standardisation of vocabularies, formats and knowledge representation frameworks, and subsequent integration efforts to provide access to the depth and breadth of data available.

2.4.1 Standardisation

Standardisation is required at a number of different levels. Data is produced in a number of formats, so standardisation is needed across data formats to make them interoperable or domains need to use specific, consistent formats to allow for data sharing and integration. Metadata standardisation is also necessary to have the same format in which this supporting information is described so that computers recognise different files as being on the same or different topics, through the use of markup languages [52]. Finally, vocabulary standardisation, which sets out a specific way of categorising and terms to be used for themes and objects within a domain, is necessary for computers to accurately integrate data sets from multiple sources within the same domain.

When single organisations, domains, people or projects describe what they are doing, they use language which is usually very specific to the task, their organisation and so on. While humans in general can navigate these challenges (albeit standardised vocabularies are very useful for humans as well), computers need more instruction as to how to organise information. Therefore standardisation in data representation is essential to ensure that the analysis is accurate and complete. The efforts above aim to define a common vocabulary to use to overcome these issues, also allowing for the better interoperability and integration of existing and future efforts in all domains, as the idea is to achieve a common understanding of the underlying framework, and the interoperability of the formats themselves. As mentioned in the introduction, this is more than just defining different objects or ideas which exist within a certain domain, the way they are defined present a specific outlook. Therefore standardisation tries to come to a common understand of this outlook, to therefore avoid the ambiguity that could possibly arise otherwise.

The importance of standardisation became ever more apparent with the appearance of the internet and in 2001 Tim Berners-Lee [18], came up with the idea of the Semantic Web, to try to rein in the proliferation of unconnected, incoherent data.

2.4.2 Semantic Web

The Semantic Web was thought up as a way to overcome the obstacle of large amounts of unintegrated data in many domains [18]. It is based on the use of common data formats, models and mark-up languages to facilitate the sharing of data across the world. The Semantic Web essentially focuses on metadata that provides common link points to information on the same or similar subject to interlink the mass of information on any one subject across the web. The use of the Resource Data Framework (RDF, data model), Extensible Markup Language (XML) and Web Ontology Language (OWL) provide the framework to describe objects.

The idea is not only to have compatible formats, but also to define data in specific ways so that the data itself is clearly and coherently categorised for computers to process together. It is making certain choices explicit, providing metadata in a consistent way so as to make data sets interoperable. On a simple level this should work with all levels of data, even if some of the unstructured data cannot be searched in the same way. "The Semantic Web, proposed to address the integration problem, can improve information retrieval beyond simple keyword matching with its knowledge representation languages and reasoning. The improvements afforded by the Semantic Web are already helping researchers answer complex scientific questions spanning multiple scientific disciplines. This has made semantic interoperability a major research topic." [156].

Standardised vocabularies and standards can be seen as attempts to develop ways of organising data in a universal and agreed format to allow for the successful integration of data [201]. They specify definitions of meaning and categories which are necessary not only for humans to understand how a specific concept is being used in a specific instance, but particularly for computers to be able to categorise the information as instructed.

While standardised vocabularies are one essential part of this large framework, ontologies are considered an important part of the solution as mentioned in previous sections. Standards and standardised vocabularies are still exposed to ambiguities as people apply them differently and the definitions can be interpreted in different ways [189,216], as well as still existing in many cases in isolated silos [177]. They are also usually structured in a flatter, more one-dimensional way, which has limited descriptive power. This is why ontologies are seen to provide possible solutions to some of the persisting limitations identified with standards and have become an integral part of the semantic web infrastructure.

2.4.3 Ontologies

A central structure within the Semantic Web is that of the ontology. Ontology originally was a Greek concept for the philosophy of the nature of existence. In the 1970s and 1980s it began to be adopted as a form of knowledge representation or knowledge organisation system, “an explicit specification of a conceptualisation” [84], with others stressing the “shared” aspect of the conceptualisation [7]. Ontologies in data science, as discussed in the introduction, are a formal representation of domains for this creation of knowledge [84, 197]. They aim to help computers process meaning in ways through the definition of said concepts (or classes) and the relations between them. They help computers to “think”, or infer further information from what is explicitly presented in the data. The Semantic Web is heavily reliant on these data structures [121,138], because ontologies do make some progress in the attempts to go “beyond” words to get to the “real” meaning of something. The idea is to help computers transcend some of the issues in natural language and ensure the accurate integration of data by defining concepts and the relationships between them in a specific space (or domain), removing their definition, as far as possible from the terminology used to describe said concepts [76]. This allows computers to seemingly make abstractions like humans but according to these fixed data structures [194]. Inference is possible because of the relations defined between one concept and another within each ontology. This is why ontologies have become such an integral part of the semantic web and its aims of being able to access and query across the whole body of data that is the semantic web.

Ontologies, as we have seen, are also important in the annotation of unstructured data. Well-defined, extensive, ontologies can be used in knowledge-based or hybrid NLP systems to facilitate accurate knowledge extraction domains, as has been seen in the bioinformatics domain [205]. As discussed in the NLP section, the issues of WSD and name normalisation present large obstacles to the correct identification, disambiguation and grouping of words according to their meaning in context. The structure of ontologies should help with that issue [136, 216].

However, ontologies face a number of hurdles. They, like the NLP systems that they often accompany, are very expensive and time-consuming to develop [27, 194], and require multiple eperts from different domains to be involved in the process. Ontologies can be used to improve NLP (hence why they are so desired for information extraction) due to their powers of inference, helping in WSD and name normalisation. However, NLP can also be used to try to identify candidates for ontologies. This demonstrates the nearly symbiotic relationship of ontologies with NLP processes.

There are areas in which ontologies have been developed and have quite extensive success, such as biomedicine [76, 110]. Also, to reduce the time burden and high cost of hand-crafted ontology creation, (semi-)automatic ontology creation has become an important research focus in recent years [4, 138, 147]. These take a number of approaches: statistical, linguistic or logical (often in hybrid systems which involve two or more of these approaches) [4, 27, 111]. These approaches vary in their automation levels [138, 161], whether existing knowledge sources are used to bolster the information gathered [111, 138, 147] and the purposes of the work.

Despite these efforts, many domains suffer from many, isolated and incomplete ontologies [11, 88, 110]. These are not easy to integrate because of a lack of common understanding when they were created. As mentioned in the introduction, because a concept is based within a specific conceptualisation, these structures may exclude information should it not comply with the specific conceptualisation presented by those who design the original ontology [194]. In fact, there are arguments as to the difficulty of coming to a "shared conceptualisation" even within a small unit of people within a domain, or a small domain itself [27, 194]. These realities hark back to the issue that ontologies face - they aim to go beyond words to arrive at the true essence of a concept, but in order to create a classification of any kind, as seen in the introduction, a perspective must be taken, choices must be made.

These choices may erroneously impose an interpretation when mapping to natural language, or may lead to a lot of information being left out because of apparent inconsistency or incoherence with the logic or conceptual model of the ontology [88, 100]. These same issues occur in attempts to align different ontologies. Going back to the discussion in the introduction, a choice has to be made as to how to define concepts, and this marks the structure. The ways these can be defined will change and always diverge unless there is a common understanding from the beginning. These issues have been described by Huang [100]. Huang describes four types of inconsistencies (in the formal sense of existence of contradictions) that can arise from multiple sources in the semantic web:

- Misrepresentation of defaults
- Inconsistency caused by polysemy
- Caused by migration from another formalism
- Caused by multiple sources

As regards the misrepresentation of defaults, the inconsistency provided by Huang [100] is a suitable one: that of defining penguins in a bird ontology as birds which cannot fly, while the default definition for birds is "are animals that can fly". The inconsistency caused by polysemy could, for example, be the representation of window which then can represent a physical or figurative window. The migration from another formalism refers to when ontologies are created by different systems which are not restricted by the same logical constraints, so concepts have been organised in different ways. This can cause problems because concepts may appear in two different places which are considered disjunctive within the ontology. Finally, where ontologies have been designed by a number of different sources, there may be differences in the conceptualisation of classes and their relations, so these will have to be considered.

Huang goes on to argue that there are two ways of dealing with inconsistencies in ontologies: to diagnose and repair the inconsistency [188]) or to avoid the inconsistency and applying different reasoning to get an answer that makes sense. The latter process is based on creating subsets of an ontology and testing to local soundness (consistency), and including it in the reasoning only if soundness is ascertained. As we have seen in the introduction, these sorts of ways of dealing with information, should we be using ontologies to extract information from natural language, alert us to the possibility of, at best, simply ignoring large amounts of data or, at worst, incorrectly imposing a structure/classification system or theory on natural language that was written from a perspective that differs from the one that the ontology creators had in mind. This is the issue. No matter how well curated an ontology is, it is then being used to map meaning onto natural language. This may work at times, but on other occasions it will either leave out lots of information, or may mistakenly impose a structure, philosophy or ordering of the world that was not intended by its author. The idea of automatically populating ontologies from data is also presented as a way of looking at how people who implicitly/unconsciously have a domain model in mind describe and write within a specific domain [11]. That same caveat would then apply to the automatically created ontology when being applied to other unstructured data that was not the focus of the original project. This issue will be looked at further in the later part of the review when looking at the domain focus of this research.

The first part of the literature review has focused on the issue of data and the processing of said data in general. The next section will look more specifically look at the domain focus of the PhD thesis: biodiversity.

2.5 Biodiversity data and knowledge

We have seen that the issue of knowledge extraction and representation is far from resolved on a general level. However, the biodiversity domain is a particularly interesting and complex case for many reasons. It has a very extensive history, with taxonomy sometimes being declared the "world's oldest profession" [39, 97]. The sheer wealth of legacy literature available, makes it a nigh on boundless task looking backwards, as well as forwards [168]. Even though initiatives such as the Biodiversity Heritage Library [85] are digitising this data, there is an overwhelming quantity of data (both in paper and originally digital format). For the data that is in the original format of paper, automatic OCR often fails to produce a suitable result for successful use with natural language processing (NLP) and machine learning (ML) annotation schema. Even for that originally in digital format there is still no widespread use of an NLP and ML framework such as the one in use in bioinformatics [205]. This means that despite all the work being done with a focus on the legacy literature, efforts in biodiversity informatics are, for the most part, an exercise in improving systems for current and future practice [174], because there simply is not the capacity to properly digitise and process all the wealth of information out there.

Biodiversity is also a particularly heterogeneous field. The very types of information collated and the way this information is presented [149,216] are particularly heterogeneous. The physical documents, their layout and the structure in which they are presented are all very diverse. Biodiversity research, by its nature, is also very fragmented, with research being described as existing in domain "siloes" [177, 178], each with their specific cosmovision, focus and emphasis [119]. There is also a high degree of semantic heterogeneity in the domains that come under the umbrella of biodiversity [41], which makes it difficult to navigate across data sets. For this to be possible, integration is needed. In fact, Konig [119] stresses the importance of thoughtful, not "naive" accumulation of data, in order to enhance possibilities of integration in current and future work. However, there are hundreds of years of legacy data to deal with too.

The Global Biodiversity Informatics Facility, speaks to the importance of integration in biodiversity research when it said, "the problem is not our lack of data but our lack of access to it, in an integrated way" [77]. This is particularly relevant in research looking at ecological systems, because it is necessary to span investigations across various domains in search of different pieces of information [149]. Integration is at the heart of the work of ecologists and biodiversity scientists, but as Kenall et al. [113] note, "Ecological and evolutionary data is typically very difficult to standardise since it can be highly heterogeneous. The diversity of subfields collecting data on very different scales of grain, extent, and time—from marine microbes to whole terrestrial ecosystems—make these highly challenging disciplines to integrate."

As described in the introduction, while humans can navigate the obstacles of different data sets, which use inconsistent terminology, different granularity of meaning and different focuses, analysing this information using computers poses a large problem. Nuances of meaning, different terminology or granularity of data must be standardised for automatic integration and processing of data by machines to be possible, as described in previous sections of this chapter and the introduction.

Biodiversity is also special because the backbone of its research could be said to be the biological taxonomy, being the structure on which other information hangs [171]. Furthermore, throughout this long history, the science of species and our understanding of them and the taxonomy into which they are organised has evolved dramatically. The taxonomy, which at first sight could be considered an ideal data form for such taxonomic or ontological data structure, suddenly seems much complex when you look beneath the surface. To better understand the problem it first is important to explain the theory of the biological taxonomy and make the distinction between that and the scientific nomenclature, the linguistic representation of said taxonomy.

2.5.1 Biological taxonomy versus scientific nomenclature

The difference between the biological classification of species and their linguistic representation is so important because it highlights the difficulty in pinning down a concept, defining it and communicating it. It highlights the blurred lines between the physical world and language and also highlights the ambiguities within both of these realms, despite our efforts to classify both.

As explained in the introduction, exactly what a species comprises is not clear, even to experts. There are various different proposed definitions with divided opinion [141, 142]. The biological classification of species (the biological taxonomy) is based upon the idea that all species evolved from one original organism, in accordance with Darwin's theory of evolution. In theory, there is only one taxonomy, however, no one agrees on a single representation of this taxonomy [68,124,170,202]. This is in part because there are still so many gaps in the taxonomy, leaving much uncertainty [94]. The evolutionary nature of species also means they can be hard to pin down and define with 100% certainty over time. As a result there are multiple existing biological theories of how species should be classified within the taxonomy at any one time. As a result, the definitions of species, as well as the labels given to species, "evolve" over time with expanding knowledge or trends in understanding [124, 202].

Taxonomists do not use the word species in their work because of the ambiguity mentioned in previous sections of this thesis. When looking at the taxonomy question it is helpful to understand the how taxonomists treat taxa. A taxon is the name given to a group of organisms classified within a taxonomy. The definition of a taxon is based on physical specimens, all of which should be used to formulate the concept described in the circumscription of a taxon. Taxonomists treat this description as a theory that is subject to change according to any new evidence that may arise, when considering the naming of said specimen in accordance with scientific labelling rules as to biological taxa.

These labels, otherwise known as taxon labels, or names, are the nomenclatural assignment of a group of organisms (taxon). There has been a structured naming tradition since Linnaean times, which provides a framework for taxonomic research and has done so for over 250 years [65]. When talking about the scientific nomenclature we are talking about this naming system, formalised through the nomenclature codes [102, 210]. These labels are the linguistic representation of the taxonomic hierarchy and classification of species, through genera, families, orders and kingdoms (and sometimes sub- levels of said rankings). They represent the decision made by the taxonomist as to where in the hierarchy of the biological taxonomy they think that specific group of specimens fits.

The scientific name given, abiding by the rules of scientific nomenclature, only alludes to the definition of the taxon. The term can be ambiguous for a number of reasons. Multiple taxa can be associated with a specific name (Melpomene refers to a plant genus (grammitid ferns) but also a name of a spider. Searching for melpomene may actually produce a result of Heliconius melpomene (a butterfly, also called common postman). It is also possible, indeed common, for there to be multiple circumscriptions, or definitions for the same taxon. This goes back to the issue in the introduction of how items are always categorised according to certain features. In taxonomy, these features are defined but even so there are disagreements in the categorisation of specimens according to these features, which lead to different circumscriptions and nomenclature linked to these.

The possibility of having multiple circumscriptions or taxonomic definitions of a species with one scientific name was one of the arguments for designing the taxonomic concept scheme. This is where the concept arises from a "classification of a group(s) of organisms by a person (taxonomist) at a given time" [200]. There are various studies that have looked this idea of taxon concept stability or instability compared to their nomenclatural stability, which look at the differing concepts behind the labels used, by comparing different taxon circumscriptions to define their congruence or not. A study on German mosses revealed 55% concept stability of taxa but only 17% nomenclature stability [122]. This investigation studied the relationship between different names and a certain taxon by defining relationships between the circumscriptions as congruent, included in, overlapping or excluding each other. This means that there are multiple potential taxa linked to scientific names, or even one potential taxon linked to multiple scientific names. This may cause problems when it comes to the integration of different data sets if the conceptual and terminological divergence is not properly dealt with. Another piece of research in the area of concept stability focused on the checklists of North American birds over the last 127 years, mapping the lumping (joining multiple taxa as one) and splitting (splitting a single taxon into two or more taxa) of taxa that has occurred to better understand the changes and the continuation of the taxonomic process [211].

Despite this ambiguity, these "labels" serve as the framework from which we hang our biodiversity knowledge. This compounds the issue, as it forms the hierarchy by which we understand ecosystems, and as a result is the heart of many efforts in trying to model this data adequately.

The examples in the preceding paragraphs hopefully help to demonstrate the complexity of the issue. While this thesis is focusing on the usage of the scientific nomenclature, to understand the complexities of its usage it is essential to understand the nature of what this nomenclature is used to describe. Taxonomists use the nomenclature to provide a label for a taxon, which they treat as a hypothesis that is there to be tested and changed as appropriate [16]. This clashes with the rest of the world who tend to use these names as if they were fixed entities. The taxon concept represents a valiant effort to deliver greater clarity in such an environment, however, it is not always applied or used consistently as a result of differences in opinions among taxonomists at any one time, and differences in levels of data sharing across the globe [36].

This demonstrates the complexity of the ambiguities surrounding scientific nomenclature usage in context. The choices made by taxonomists when they produce checklists: decisions to split or lump species (with the accompanying specimens and taxonomic circumscriptions), the placing of one taxa in a particular genus or another, represent differing perspectives as to how to present the same information [175, 211]. These differing perspectives represent the multiplicity of biological taxonomies. The names assigned in these documents are then used by a whole panorama of different people, experts from a plethora or domains, and also laypeople of all kinds. The names are applied ambiguously, without their "concept identifiers", that is the author plus date of the circumscription, using ambiguous synonyms, or even common names [192]. For those attempting to standardise and integrate biological/biodiversity data, this is a daunting, if not impossible task to do manually, let alone using computers. The importance of being able to track all this information, and access its actual usage in the context of narrative journals is highlighted by [175]. Despite this, as far as I am aware, there has been no attempt to map scientific nomenclature usage empirically across legacy narrative data sets (e.g. journal articles).

Having looked at the qualities of the data within the domain of biodiversity, the complications of standardisation and integration and the importance of correctly integrating said information, we shall now look at efforts to try to achieve this.

2.5.2 Standardisation and integration efforts in biodiversity

Standardisation efforts in biodiversity first focused on developing standardised vocabularies for the domain. These efforts are still ongoing and are very much a part of the biodiversity informatics infrastructure. To try to develop a common language through which scientists describe their work, there are a number of standards that have been developed. The Biodiversity Information Standards TDWG (Taxonomic Databases Working Group) have developed the Darwin Core (DwC) standard [218], which governs biodiversity-related issues, as an extension of the Dublin Core standard [48]. Other standards include the Ecological Metalanguage, used for recording information about ecological data sets and there is also the BioCase standard and repository used primarily for biodiversity collections in museums, to give some examples. These all include standard vocabularies of terms to be used, to encourage a common understanding between different data sets.

Biodiversity experiences the same problems of limitations identified with standards as other domains, in that it lacks the “kind of semantics or knowledge modelling needed for robust logical inference” [216]. Ontologies, as in other areas, are thought to be able to provide a more complex, multi-dimensional infrastructure because of their capacity to go beyond a taxonomic hierarchy structure to map multiple types of more complex relationships - interconnecting different types of data according to the concepts that they represent [66, 216]. They are also considered to be better at identifying the concepts beyond the word used to describe them [99,149,216]. As mentioned in the previous section, the infrastructure that ontologies provide, which allows for inference, is seen as a particular advantage because it opens the way for many possibilities regarding the complex querying of data. Ecologists, who work across many different domains to answer queries, as do many other biodiversity experts, currently have to perform many different searches in different areas to access the data they require to perform analyses. Use of overarching ontologies is hoped to streamline this process.

Despite this, biodiversity suffers the same issues with ontologies as other domains [110,216]. There are, in biodiversity as in other areas, many ongoing efforts to expand existing ontologies and introduce infrastructures to expand the capacities of ontologies and make them cross-operable. An important aspect of this, going back to what was mentioned as regards the semantic web, is having overarching guidelines as to how to describe certain things, and general, high-level ontologies that more domain-specific ontologies can then feed into. In the case of biodiversity, the OBO Foundry guidelines are those that are commonly used to “break down the barriers among data silos, enhancing the value of biodiversity data by allowing researchers to query across data sets” [216], by being able to implement different ontologies side by side.

As regards the issue of fragmented data in ontologies, continuing efforts can be separated into three broad categories: manual curation and expansion of existing ontologies, the compilation of repositories comprising existing knowledge bases, ontologies and terminologies and finally semi-automatic ontology population and biodiversity data annotation. These efforts will now be described in further detail, analysing any achievements and the obstacles that persist in the domain.

There are various efforts to expand existing ontologies in a mostly manual way, involving groups of experts working in affected fields. These efforts also aim to align work within the OBO Foundry guidelines and align multiple ontologies where necessary for the integration of data to facilitate searches in the ways described necessary in previous sections, such as work to expand and integrate the PCO and BCO [216, 217], the ENVO [31, 32], and the Plant Phenology Ontology [199]. These projects, and resources such as the Extensible Observations Ontology (OBOE) [137] aim to overcome issues of potential logical conflicts between different ontologies [149].

Another approach tries to harness a broad range of existing resources, and there are a number of examples of hybrid integration systems, which support a variety of different structures. The Global Biotic Interactions (GloBI) database [178] focuses on the aggregation of interactions between species. As the researchers responsible for the platform state, "with a detailed understanding of these interactions, ecologists and biologists can make better informed predictions about the ways different environmental factors will impact ecosystems. Despite the abundance of research data on biotic and abiotic interactions, no comprehensive and easily accessible data collection is available that spans taxonomic, geospatial, and temporal domains" [178]. The initiative uses a combination of standard taxonomies, ontologies, vocabularies and structured data repositories to integrate siloed information that is hard or impossible to cross-reference on its own. Input consists of interaction data sets in the form of structured and semi-structured data, which are integrated using ontologies and vocabularies in a semi-automatic way, to ingest and standardise data sets to provide an overall picture of the information provided.

The AgroPortal [110] is another such example of such a hybrid system. It describes itself as a repository for ontologies, standardised vocabularies and other resources relating to the agronomy field. The resource contains features such as automatic annotation which can simultaneously consult all the available taxonomies and ontologies for key terms, and automatic mapping features which try to map concepts from one ontology or resource to another to account for overlap. The annotation feature uses natural language processing techniques and ontologies to annotate data and then map said data onto concepts within free-flowing text [110]. This initiative aims to facilitate experts working in the domain of agriculture to access and mine information relating to this field in the smoothest way possible.

Finally, the ClearEarth project is looking to adapt the NLP and ML algorithms used currently in bioinformatics contexts and retrain for use on domains of geology, ice and snow and biology, in an example of efforts to populate ontologies semi-automatically. To do this they have worked intensively with expert annotators (both domain and linguistic specialists) to develop annotation guidelines for each area to manually annotate training corpora. They proceeded to retrain the aforementioned algorithms using data relevant to the above domains and held a Hackathon in Summer 2017 to test and evaluate the success of the algorithms so far at extracting data relating to "morphologic and behavioral traits, trophic relations, habitats". A recent conference proceeding publication describes the work they are doing in the area of ecology using the ClearEarth annotation tool (trained firstly by expert annotators as described above) to extract terms and relations which are going to be used to feed into the Ecocore ontology [204]. A paper they published about the annotation success of their retrain algorithm boasts overall precision of 85.56% and recall of 71.57% for the named entities selected [204]. The ClearEarth project demonstrates what can be achieved with resources and time. But the battle is far from over.

Considering the entirety of these efforts, it is possible to identify continuing obstacles in the domain. The overview of all these projects shows that, despite work towards the integration and alignment of ontologies, the work on this is still patchy. Most ontologies are very specific in focus, vary in granularity and, despite efforts to the contrary, continue to have massive gaps in the content covered. This provides barriers to the large-scale integration of data [5, 32, 216].

The process of developing ontologies requires collaboration between experts from various different domains in the stages of their development [31]. In the case of biodiversity, it involves scientific experts who provide information about the concepts and relations that need to be defined, and the hierarchies within these. NLP and ML experts are then needed to use this information to create algorithms, identify the relevant data in large corpora and extract these concepts. Finally, ontology creation experts, who are capable of translating this into OWL or another ontology language, are needed to create the structure of the ontology. For example, the ClearEarth project made use of hackathons to first annotate training data using scientific domain experts, followed by NLP and ML experts using this training data to annotate the test corpora and then ontology experts in the definition of the ontology structure and creation (project ongoing) [204]. As we can see, given the complexity of the process, it is an exceptionally time-consuming task [216] and ontologies are very costly to develop [32]. Automatic ontology creation does not seem to be seen as developed/accurate enough to be a major focus in this area. ClearEarth are making headway but it still requires the initial effort of manually annotating the training data sets.

All these projects highlight some of the difficulties firstly in the population of ontologies, the definition of classes and organisation of the conceptual model, and then also on how these ontologies map to data sources themselves. The complexity relating to the hand-crafted identification of concepts and the relationships between them arises repeatedly [32, 194, 216]. While this is particularly seen as a problem cross-domain [216], there are a number of instances even intra-domain in which the definition of concepts it considered to cause substantial issues [194]. Even when human beings are discussing conceptual definitions, it is not always easy or possible to come to a final agreement that works for every context. For example, the Population Ontology Community (PCO) workshop described in Walls et al [217], working towards extending the usage and capacities of the ontology, came across some conceptual hurdles. Specifically, the term “ecological niche” was recognised as complex as “different ecologists have formulated niches in different ways [12-14]. Some more focused on spatial ecological meaning, some in community ecological frames, and yet others related directly to species physiological tolerances. Previous work on the Environment Ontology (ENVO) had conceptualised the niche as an environment that would allow a given species to maintain and expand its population” [217]. In the end, the decision made during the workshop was not to have a single class relating to this one apparent “concept” in a single domain, as really it seems that it can be defined as a term, which has different concepts attached to it. The ENVO ontology project [31], for example, recognises the difficulty in certain terms such as “biome” and “habitat”. In this case, the team chose to present the terms in a looser way, to allow for differences in how they are used across different domains. These sorts of issues highlight the reality of needing to make explicit decisions which affect the conceptual model of one ontology or another. In the case of ecological niche, a detailed separation of concepts seems to have been followed, whereas in the case of biome the curators seem to, at the moment, have tried to maintain the idea of one, multi-faceted concept to facilitate integration across domains. This is very demonstrative of the issues faced in building ontologies, and also the caveats in mapping unstructured data using said data structures (in being sure they are a suitable model). This is a clear demonstration of that described in the introduction and earlier in the literature review. To try to get to the essence, or the inherent truth of something does not really exist: it is dependent on the context, or the perspective from which something must be presented.

The researchers on the ClearEarth project pay heed to conceptual difficulties in a paper [205] which discusses the issues they faced in developing the annotation guidelines for their project. During the process, there were many obstacles in deciding how to assign annotations, where to delimit the annotations. In this case it would seem that differences in opinion of classification of annotations was across expertise boundaries. However, the project worked with two domain experts and two linguists only. These problems still highlight the inescapable choices in these tasks. There is never one way of modelling information, which can cause problems for ontology integration - or using as a model for loads of unstructured data without knowing the ontological model chosen is suitable for the data.

When looking at difficulties perceived in some aspects of automated ontology production one of the obstacles seems to be the limited nature of the knowledge bases used for the learning of concepts and relations in the first place, in contrast with the variety of natural language (linked to the fragmentation and incompleteness of ontologies). The question of the quality of the training data sets, also arose. The researchers recognise that “The natural texts were not written with the ontologies at-hand. Rather, dictionaries and thesauri, or popular usage are the basis - particularly for some genres. It has to be allowed that many written texts will deviate from the highly technical precise definitions in the ontologies, and those deviations could degrade the NLP result.” [205]. Conversely, this is one of the arguments given by [11] for the automatic creation of ontologies from natural language texts, describing it as the reverse process of traditional ontology population.

As regards using ontologies to map to data (structured and unstructured), various projects have also highlighted difficulties in doing this correctly. This is linked to the difficulties found in word sense disambiguation (WSD) as discussed earlier in the chapter, considering the complexities of natural language. The AgroPortal has an automatic mapping feature, which has some limited functions [110]. Limitations occur as a result of concepts in different ontologies that appear analogous actually refer to slightly different concepts, or ontological classes with different names may refer to the same concepts. When looking at AgroPortal's annotation feature, it was interesting to note that it would incorrectly annotate "order", referring to a taxonomic rank, when the semantic meaning of this instance was "in order to" as a set phrase. This shows that the mapping does not or at least sometimes fails to take into account the context in which the word is found.

The Environment Ontology project also came across similar issues, with their automatic mapping feature, in which they identified "simple false positives, ambiguous class labels, and text-mining routines which only account for the basic structure of the ontology" [32] as the error categories. These issues can be traced back to lack of sufficient NLP processing in data, and also refer back to the same sorts of issues that ClearEarth found in deciding on annotations as for the adequacy of class labels, and how to present these.

The Plant Phenology Ontology project, which focuses on being able to integrate structured data, also identifies that even after integration, a lot of the time data are not then ready for each analysis, because of differences in the conceptual models of the data (data resolution, methodology, etc.) [199].

This leads to a more fundamental issue identified across the different projects. ClearEarth even argues that ontologies are "not exactly what is required for the operation of processing natural, people-authored texts" [205]. This links back to their argument about the deterioration of the quality of the final result of the ontology classes, but also highlights the issue of then mapping legacy data (particularly unstructured, but all types) to said ontologies. Ontologies are made with their specific world view, they are defined and therefore unambiguous. However, if the data being mapped does not fit with said conceptual model, it may have said worldview erroneously imposed on it, or simply be left out. While understanding the need to avoid the integration of data which is in fact incorrect, it can also be argued that being able to identify the characteristics and conceptual model on which the data itself is guided is an essential tool to be able to evaluate the validity of said data and to potentially uncover new and valuable knowledge.

Having looked at some more general projects looking at the standardisation and integration of information in the area of biodiversity, we shall now look at some of those initiatives with a specific focus on biological taxonomy and scientific nomenclature.

2.5.3 Standardisation and integration efforts relating to nomenclature and taxonomy

Having seen the efforts on a broader scale in the domains of biodiversity, issues relating to scientific nomenclature (naming) and biological taxonomies (classification) will described separately, considering how key this information is to how biologists, conservationists, ecologists and other biodiversity experts communicate and interpret data. While the focus of this thesis is on the usage of the scientific nomenclature, scientific nomenclature cannot be completely separated from the taxonomy. They are two different things, but they are the labels used by experts when communicating about species, or taxa. The complicated relationship between the two is described in the section on biological taxonomy versus scientific nomenclature. At first sight the biological taxonomy and the scientific nomenclature may at first sight appear to be an ideal place to start as regards taxonomic or ontological data structures but the evolutionary and multiple nature of the taxonomy mean that they are very challenging to tackle in this way [66]. This can all be linked back to the discussion in the introduction about the decisions on how to classify information according to purpose. Therefore this section will look at efforts both to model scientific nomenclature and biological taxonomies.

The ambiguity of the nomenclature was mentioned in a previous section of this chapter. For this reason, in the standardisation and integration of this information, a key concern is the erroneous integration of data, particularly in the context of scientific nomenclature. Concerns include the ambiguity of scientific names, making it impossible to unequivocally link them sometimes to specific organisms due to the multiplicity of names being used to refer to a single concept, or vice-versa [61, 109, 172]. To overcome this in the past, the idea of the taxonomic concept was born [63, 66, 109, 168]. To clarify this in a computational way, unique identifiers (LSIDs, UUIDs) are a common method by which to overcome the issues of multiple synonyms and homonyms by linking them under one unique identifying number that represents a single concept [172]. However, as with many initiatives, there are multiple identifier systems and this can only be performed with new data, or legacy data with sufficient information to link to one or other of these identifiers.

There are many integrated taxonomic resources in existence, such as the Catalogue of Life [225], the Encylopedia of Life [170], the Plant List [203]. There are also a number of efforts to provide taxonomic data in an ontological format. Ones that provide information about nomenclature assignment include the Vertebrate Taxonomy Ontology [150] and the NCBI (National Center for Biotechnology Information) organismal classification. These resources focus on nomenclatural relationships, not taxonomic ones. Franz [66] stresses the fact that this is sufficient and all that is necessary for many specialists working with scientific nomenclature, and even supports the creation of a comprehensive ontology which maps the evolution of nomenclature.

Modelling the biological taxonomy and the scientific nomenclature together is a complicated task. Many of the combined resources are what [175] calls a “backbone name-based taxonomy [...] a single, monolithic hierarchy in which any and all conflicts or ambiguities have been pragmatically (socially, algorithmically) resolved, even if there is no clear consensus in the greater taxonomic domain". While recognising the need to be dynamic they are actually singular instead of multiple in their presentation of the taxonomy. Therefore they are effectively choosing, only able to present an image of a specific moment in time, from a specific point of view. This is a problem if this is to be used to map scientific nomenclature usage because it may impose a structure that clashes with that which was intended. Changes to the imagined organisation of the biological taxonomy, or different perspectives, can impact on scientific nomenclature usage. Despite this they are partially separate, and at times seemingly contradictory [66, 191]. There are various efforts to try to overcome these issues. It is argued that there are "two fundamentally different models to create ontological representations of taxonomy; viz. strictly nomenclatural and full-blown taxonomic representations" [66]. Given the number of different purposes of the nomenclature and biological taxonomy, each of these are useful for different communities for their varied purposes.

There are efforts that try to overcome this issue. Firstly there are efforts to create overarching infrastructures which integrate these other resources and help encompass larger and larger scopes, some with multiple different taxonomic backbones, such as the Global Names Architecture [179], which is part of the Global Biodiversity Information Facility [70], an aggregating platform that concentrates primarily on primary source data, including species observations, distributions and reconciliation of species names. However, currently, the research that is looking to overcome this modelling issue tries to separate scientific nomenclature from taxonomy in a variety of ways.

There are recent efforts that try to model the scientific nomenclature separately to the taxonomic process, and model the taxonomic process specifically instead of a moment in time, to account for these issues. Schulz [191] intends to create an ontology of biological taxa that is"neutral to the different and conflicting species conceptualizations. It departs from the principle that biological taxa are something that regardless of its existence in nature or its (fiat) attribution by biologists has a highly-ranked importance in biology and therefore requires to be accounted for in biomedical ontologies" [191]. This is indeed true, but it fails to address the issue of looking at unstructured literature to identify usage of specific terms and patterns of usage to investigate whether there are differences, or inconsistencies in said usage.

TaxMeOn [209] is another such approach, in this case a meta-ontology that tries to separate the taxonomic concept from the scientific nomenclature used to describe it. It also includes a common names section on the understanding that these names are used even more ambiguously and that this changes across different countries and geographies dramatically. OpenBioDiv-O [192], is probably the most complete effort in the works so far. The OpenBioDiv project presents a "dynamic representation of the scientific process of biological taxonomy and not of any particular state of knowledge" [192]. It has produced an ontology that can be aligned with other important domain ontologies. In contrast with other previously mentioned taxonomies, “multiple hierarchies of taxonomic concepts may exist [...] it is possible according to the ontology to have two sets of taxonomic concepts (even with the same taxonomic names) with a different hierarchical arrangement” [192].

On top of the OpenBioDiv-O, a knowledge system has been created, that is called the OBKMS (OpenBio Knowledge Management System) [175]. The system aims to overcome the barriers of having siloed taxonomic information: that locked in taxonomic treatments, all the changes and multiplicity of terms linked to said taxonomic treatments and all the information in different types of scientific literature. It tries to collate all this information together to facilitate searching and tracking of said data through time. The OBKMS in fact uses data from journal articles to track usage of scientific nomenclature in context.

None of these initiatives provide a complete picture of the taxonomy or of scientific nomenclature [102,210]. OpenBioDiv knowledge management system is the only initiative that explicitly describes an aim of looking at how scientific nomenclature is actually used in context, and this is through semi-structured data [175]. This is despite repeated and multiple recognition of the ease of erroneously understanding scientific nomenclature in scientific literature because of the issues explained up until now, also the recognition that nomenclature is used inconsistently for these reasons. For this reason a way to evaluate usage and map the hierarchy of said usage in unstructured bodies of literature (scientific or otherwise), without imposing an external hierarchy represents the gap identified here.

In this first part of the literature review we have seen how researchers and professionals in the areas of biodiversity, informatics and archiving are trying to overcome issues of standardisation and integration in the domain of biodiversity, the successes and continuing obstacles. We have also considered the specific case of biodiversity as regards the interplay between biological taxonomy and scientific nomenclature. The importance of conceptualisation is central throughout throughout: how to pin down a specific concept and how this is linked to the linguistic representation(s) of said concept. The issue of usage is also central: how stable is said concept or, in this case, how stable is the usage of said linguistic representation. As we have seen in the first part of the literature review, to my knowledge there has been no empirical review of the biodiversity literature to profile nomenclature usage in context in a way in which current knowledge representation resource can be evaluated, or in which different corpora can be compared for their characterisation of the nomenclature, despite a recognition that this is important for accurate access to the information locked inside the wealth of unstructured data in existence.

The second part of the literature review will focus on the approach being taken in this research project to tackle the gap identified: a method by which to profile the usage of scientific nomenclature and common names in context to determine the stability of the concept it is used to represent. The method aims to forge the way to evaluate the completeness and accuracy of existing ontologies and other knowledge representation resources, as well as make judgements as to the stability of conceptual usage of terms across corpora or against existing knowledge representation resources, to aid in the accurate and reliable integration of data. In this thesis I applied a lexicographic approach to the problem. The area of lexicography is dedicated to identifying the concepts behind word use in context, which is arguably ambiguous and undefinable in the strictest sense [69]. The history of lexicography and corpus linguistics and related research to support the line of investigation taken will now be explored.

2.6 Corpus linguistics

Corpus linguistics is the basis for the approach currently taken in lexicography, and so the theory and practices behind corpus linguistics will be set out here.

The techniques applied in corpus linguistics, that of language analysis using naturally-occurring, real life texts such as books, newspapers or letters, can be traced back to biblical and literary scholars as early as the 13th Century [163]. Samuel Johnson, in fact, could be considered the first corpus linguist due to the way he used index cards and real examples for his dictionary [62]. Later, at the beginning of the 20th Century, corpus techniques were used in lexicographic studies such as the Oxford English Dictionary, in which bodies of real-life texts were gathered in concordances to perform studies of language, as well as dialectal studies and other such pursuits [58].

The seeds for modern corpus linguistics were first sown around the 1950s, driven by a desire to collect real, true data to analyse [127], but it was not until the 1980s and 1990s that corpus linguistics really started to come into its own as it is conceived today. These days corpus linguistics, more specifically corpus-based analysis, is understood to be a methodology which allows for the empirical computational analysis of large, principled collections of texts [19]. Such analysis is both quantitative and qualitative in nature. The aim behind corpus linguistics is to examine the distribution of word usage over a large body of text, to find patterns in usage – whether focusing on semantic, lexical, grammatical or syntactic features. The corpus must aim to be representative of a specific language, or aspect of language, to get a faithful representation of the language being studied.

Corpus-based analysis uses frequencies of words or phrases, along with concordances, to draw conclusions about language use in context, being able to make judgements on what actually appears in real texts. Gries [83] argues that the formal differences in patterns of language usage, whatever the unit of language being analysed, can be attributed to functional differences. Therefore, what Gries is saying is that corpus linguistics can be used to identify frequencies of patterns of co-occurrence of specific linguistic features, and that these patterns or variations in the frequency of these patterns across difference data sets will tell us about variations in language use. These frequencies, alongside comparative analyses of variations between data sets are what can be used to draw conclusions about language use and change.

The idea is that one can use quantitative methods to draw assumptions about specific features of language. The patterns revealed by frequency data in context can provide information about the functional differences between languages used in different domains or any groups of text which can be separated on the basis of any specific, measurable feature(s). Corpus analysis has the benefit of allowing one to take a step back, looking at the quantitative data to see patterns in usage over large data sets, but also allowing one to consider the data qualitatively by delving into the texts themselves. This permits a better understanding of the qualitative differences in the contexts in which words are used and the effect of this on the concepts they represent. This supports the idea that context is everything – whether the lexical, grammatical or syntactic - because changes at any of these levels can indicate changes in conceptual meaning.

Computational corpus linguistics varies from other, more traditional, forms of linguistic analysis because of its emphasis on real-life examples and also the quantitative side to the analysis. Corpus-based analysis is used to analyse large bodies of language data to identify formal patterning, repeated events and other insights into word behaviour. It is an empirically-based methodology, which makes use of both quantitative and qualitative analysis techniques. It is typically used in the domain of lexicography to identify the actual use of words in context but is also applied to discourse analysis, and as a translation and language-learning aid. It has been essential to the evolution of lexicography, which has relied on the computational nature of this approach to analyse large swathes of documents. The mix of qualitative and quantitative techniques means that frequency and co-occurrence analysis can be used to identify potentially interesting phenomena, which experts can then analyse fully through qualitative techniques. This thesis focuses on a data set which is widely recognised to be heterogeneous and extensive. The thesis aims to develop a method for extracting and conceptualising various nomenclature references across dataset(s), to create a profile against which to compare existing knowledge representation resources in the same subject. Corpus linguistics, because of this flexibility, and because of the proximity to the raw data, make it a suitable approach for this task.

Most forms of linguistic analysis are wholly qualitative in nature. They do not support generalisations of findings because they focus on specific examples and performing in-depth analysis on these. An example of this could be during the time of the structuralists such as Saussure (1857-1913) and Benjamin Lee Whorf (1897-1941), who were interested in how the structure of language affects how we see the world, therefore these differences between languages shape our perceptions. Much research at the time focused on travelling to different parts of the world to study specific populations and interview people [213]. There are pros and cons to any method, and corpus linguistics is sometimes accused of steering too much on the side of quantitative data, ignoring some of the in-depth analysis allowed by other linguistic techniques. However, other linguistic techniques are more susceptible to bias, given the personal and specific nature of the data accumulated and lack of transparency in the theories derived from any analysis. It must be emphasised that corpus linguistics is not immune to bias. One must be particularly careful when compiling corpora, as this can skew results.

Corpus analysis is used in a variety of fields, particularly that of language learning to find patterns of use in learners of a language or as a teaching aid for learners of a language, translation, terminology, lexicography. It can also be applied to different fields such as sociology.

2.6.1 Lexicographic approach

The lexicographic approach is the description of the methods applied in modern lexicography to identify the different senses (concepts) words are used to express. The approach applies corpus analysis to analyse word use in context, separating out the different meanings as required to create definitions for words in dictionaries or to organise words into thesauri as a categorisation tool. As described in the Routledge Handbook of Applied Linguistics, “lexicography is an area of applied linguistics that focuses on the compilation of dictionaries (practical lexicography) as well as on the description of the various types of relations found in the lexicon (theoretical lexicography)” [62]. This thesis sets out research which aims to apply and extend the techniques used in this approach to profiling the scientific nomenclature.

Lexicography has historically been based on the idea of defining the different senses words are used to express. In the past dictionary making was “traditionally carried out by writing examples on cards indexed by the word of interest, with the examples being found by long and extensive reading, and relying on the instincts and intuition of readers” [117]. At that time, the prevailing position was that words possessed “correct” senses that were somehow integral to their existence. The first dictionaries were seen as something that would serve as “a prescriptive and normative authority which would serve to establish a standard of correctness” [62]. This perception of dictionaries and definitions follows the logical definition tradition of Aristotle, in which “a ‘definition’ is a phrase signifying a thing’s essence” [10], representing an intensional understanding of the subject.

Corpus linguistics was first applied to finding all the possible mentions of a specific term or phrase. Lexicographers quickly adopted it as a way to analyse the use of language and words based on real, naturally-occurring text, in a methodical and more complete way, rather than relying solely on intuition as before.

Corpus-based analysis allowed lexicographers to go further and empirically analyse large bodies of data, looking at the different senses expressed by the same words, in different contexts, without missing examples or having one’s own personal biases on meaning colour the analysis (or at least mitigating this bias). As such, corpus linguistics revolutionised lexicography and people started to realise that words were contextually founded, that their meaning varied depending on context. “Corpus linguistics makes it possible to identify the meaning of words by looking at their occurrences in natural contexts, rather than relying on intuitions about how a word is used or on incomplete citations collections” [19].

With advent of computers and the arrival of computational lexicography, different forms of empirical analysis such as concordances showed that both word use and meaning are highly dependent on context, actually representing a very fluid idea of meanings words are used to express, instead of the idea that a word has a correct and well-defined meaning.

Abraham Solomonick in 1996 declared that should “our rules for defining words and other lexical units in dictionaries must be severed from the rules of logicians [. . .] because logical definitions are aimed at defining things and phenomena in reality [. . .] and lexicographic definitions – at defining units of a linguistic system, called language” [196]. This is referring to the fact that actual things and the language used to describe them are actually very different.

As the Lexicography chapter in the Routledge Handbook of Applied Linguistics highlights, when defining a dictionary entry, one must impose boundaries between word senses. However, when looking at corpus data one sees that the senses actually overlap, there are no clear boundaries just different uses that can be used to see patterns on a large scale [62]. This is why the advent of computational linguistics marked such a shift in positioning as regards definition, as for the first time these patterns (or lack thereof) could be observed over large swathes of natural language text. This is very similar to the need to impose a boundary on one species to the next for communicative purposes, or the need to define a concept from a perspective in an ontology.

As a demonstration of this, lexicographers are often faced with decisions relating to lumping or splitting senses. As Kilgarriff mentioned in his paper “‘I don’t believe in word senses’”, the organisation of the information one is presented with depends on the context in which one is working, and the purposes of one’s work. Much work on dictionary entries is to do with the “splitting” or “lumping” of senses [62]. These choices are governed by many factors, such as the length or size of the dictionary, the target audience, whether the dictionary is a general one, or a specialist one. However, as argued by Halliday [87], there is no clear defining line between each way of presenting this information, more of a continuum along which a decision must be made. As we have seen in the previous sections, the concept of “splitting” and “lumping” senses has its parallels in taxonomy as well. Going back to the North American bird checklist article [211], there have been patterns of “splitting” or “lumping” of species concepts according to new data, changing traditions within the field of taxonomy, which shows the fluidity of taxon concepts in the same light as those of definitions.

There are also those who claim that word senses as such do not exist. Adam Kilgarriff, in

1997, wrote a paper in which he claimed, “the corpus citations will be clustered into senses according to the purposes of whoever or whatever does the clustering. In the absence of such purposes, word senses do not exist” [115]. He defends that meaning is a construct of the context in which it is placed and the person interpreting that context. In fact, “many lexicographers today therefore reject the use of definitions (both the term and that which it stands for)” [69].

This shift from the idea of a logical definition, in which there is a true essence in the meaning of a word, to the idea that the senses of words only exist within the context they are being described, is exactly why the lexicographic approach is a suitable approach to look into the use of terms and scientific nomenclature in the biodiversity literature. The biodiversity domain is faced with many of the issues that dictionary makers are faced with. In the area of taxonomy, taxonomists try to define concepts based on specific specimens and their characteristics, but then to be able to discuss this concept it is given a linguistic label, that of scientific nomenclature, that is used within the literature to communicate. This is similar to the role of dictionaries to create well-defined limits where there are none in the definitions of words, to be able to categorise and give clear definitions to serve as a guideline for people to communicate, whether in everyday life or as regards the area of biodiversity. Returning to the discussion in the introduction of how we classify information according to features that are relevant in the particular context in which they are being used also finds its parallels here. This is relevant in the context of ontologies because if they are to be used to organise information, it must be clear that the information being mapped follows the same conceptual model.

Given the obstacles identified in using formal ontologies for mapping the scientific nomenclature, this approach offers opportunities to validate and evaluate current knowledge representation systems because it allows for inconsistencies and contradictions in the source data and includes these in the word characterisation so that 100% of the available evidence can be considered [62,115]. The lexicographic approach will provide an empirical way to see how the terms are presented in different data sets and move from there, instead of the other way around. To the best of my knowledge there has been no empirical study of the use of scientific nomenclature in the literature, nor an evaluation of this type of existing ontologies in this domain.

2.7 Word and relation characterisation

Throughout the literature review, a number of different approaches to different levels of ontology creation have been presented and discussed. The methods employed range from manual curation to the use of statistical, linguistic and logical techniques. Many employ a combination of these, given the complexity of the task. Said ontologies also take various forms, from the most simple, taxonomy ontology, to fully-fledged formal description logic ontologies. Word and relation characterisation is essential in the development of ontologies, these being key units to their structure.

This section will briefly outline the usage of word embeddings in word relation and characterisation, before outlining the lexicographic approach to this and making the argument for choosing to employ the latter in this thesis.

2.7.1 Word embeddings

In recent years, word embeddings, the distributional representation of words as vectors, have been shown to be very successful at identifying word similarity and classification [152,154,176]. The vector representation is calculated according to multiple features relating to a word or phrase according to its context. Word embeddings are calculated using statistical methods that learn these features through the processing of huge quantities of data [224]. Relations between words are seen by their proximity or distance from each other in the vector space. They have also been shown to tend to form semantic groups in the vector space, showing patterns according to relations such as city to country, or man to woman [126].

These qualities have made word embeddings hugely popular in some information extraction tasks in recent years and even have been used in some ontology population tasks [101,108,173, 224].

However, word embeddings experience some complications. Domain specific (DS) embeddings are a more complex issue for two reasons: the amount of data required to train on such data sets is not available, and also embeddings trained on general data sets do not tend to give particularly good results because domain-specific terms have different meanings to the ones the word has in a general context. There has been a considerable amount of research in this area [73, 160, 185, 224], which take different approaches to adapting word vectors to domain specific terminology. However, these approaches require either very large data sets, or the use of knowledge resources to support the creation of said word embeddings. The desire to compare relatively small corpora of different types limits the possibility of using word embeddings in this research.

Word embeddings are also not easily applied to words with multiple meanings, as they usually produce only one vector which is an output relating to the weighted importance of all possible meanings of a word [152, 176]. There has been work into producing vectors to distinguish between different meanings of a word to account for polysemy and homonymy [223]. However this relies on WordNet as a resource to identify known different senses of words. This would not be applicable in this case, as by imposing a semantic framework on the data, it would undermine the empirical nature of the approach.

In relation to this thesis, all of the above approaches suppose a distancing from the text itself and focus on domain specific terminology, specialised language in which the meaning of a word differs from its meaning in a general context, not on terminology the meaning of which may vary within similar contexts. While word embeddings do work on empirical data, the processes involved in creating word embeddings mean that it is impossible to go back to the data to see why a certain embedding has been produced. This is because of the size of the data, the number of features and the lack of transparency as to what these features are. In the case of the scientific nomenclature, where usage is so internally ambiguous, this does not appear to be an appropriate first step to profile usage across different corpora.

2.7.2 Lexicographic approach

The lexicographic approach employs the use of corpus query tools to perform the analyses described in summary in the first part of this section. It is described in more detail in the methodology chapter.

As mentioned earlier in the chapter, lexicographic analysis is used principally in the production of dictionaries, being used to identify the different contexts in which words are used to create definitions of the different senses each word can be used to represent. This can be performed using concordance analysis, using keyword searches to identify terminology particularly for specialist domains, and also using both statistical and linguistic association measures which identify specific collocations within different corpora. These are the measures used to look at collocational patterns of words in context, to discern different groups of meaning which are split or lumped according to the requirements of the dictionary. In looking at the creation of a dictionary entry, however, there is one feature that stands out against others, which will be described in the next section.

The lexicographic approach can be used to extract the hierarchy as it exists within the test corpus itself, without imposing an outside interpretation. Aside from ascertaining word meaning through context, corpus linguistics and lexicographic analysis can be used to identify taxonomic and other relationships between words. Hearst [90] defined so-called Hearst Patterns, which are common linguistic patterns which denote specific taxonomic and other relations between words such as the head modifier principle [95]. These have been used in various areas to identify taxonomic relations and create taxonomic structures of lexicons, as well as when corpus-based analysis has been used either alone or in conjunction with other statistical techniques in (semi-)automatic ontology learning to identify concepts and relations [4, 11, 27, 73, 194]. Corpus-based analysis is also used in data-driven evaluation of ontologies, to check for coverage and accuracy of the concepts and relations within a specific domain [11, 26, 27].

The advantage and difference in the approach being taken in this research project is the aim to look empirically at the data with all possible perceived occurrences of inconsistency or incoherence to serve as a validation tool for existing ontologies, because of the characteristics of this approach. The next section will describe the corpus query tool applied throughout this thesis and research surrounding the Word Sketch feature, a specific focus of the thesis.

Sketch Engine and Word Sketches

The Sketch Engine [116] was designed as a corpus query tool for use by lexicographers. It is described in more detail in the methodology design section. A corpus is a collection of natural language texts, which in today's world are found in machine-readable form. As well as traditional corpus analysis features such as concordances and word lists, Sketch Engine has a feature called Word Sketch [114], which produces a one-page statistical overview of a word's grammatical and relational behaviour in a large corpus (collection of texts) in a simple and easy to digest way. Word Sketches are traditionally used as a basis from which to draft a dictionary entry, being used to evaluate the most salient uses of a word, along with typical collocations and contexts in which said word is used, identifying the key concepts this word is used to represent [116]. The Sketch Engine aims to access the senses (concepts) each word is used to represent, show the multiplicity of this and also how different senses might be linked to specific contexts or constructions. This is particularly apparent when looking at Word Sketches which give a summary of frequency lists of word use in context – showing the patterns of word use in context, to give us an idea of the concepts behind the word. The concepts can be identified by the mix of different semantic contexts in which the word is used, and the different grammatical roles it takes.

Word Sketches have also been used to tackle research questions, mainly in the contexts of terminology and semantics. In the area of semantics and word characterisations, the research project focused on adding semantic annotations to the study corpus [144]. There has been research into extending the use of Word Sketches to categorise words according to semantic features as well as syntactic and grammatical ones. McCarthy et al [144] started to research this possibility by semantically annotating the UKWaC corpus with the WordNet lexical database supersenses [60]. This preliminary research into adding extra features to the Word Sketch tool demonstrates the possibilities for further extending the feature for use within the linguistic community by using this resource to semantically categorise the use of words by their contextual attributes.

McCarthy [143] then went on in 2016 to look at the clusterability of word senses and how this was related to how well-defined differences between the word meanings in specific contexts are. This showed that in cases in which the different meanings a word is used to represent were well-defined that the clusters were in turn stronger and better defined. This means, for example, that a word which is used in very specific, distinct contexts to represent different meanings would cluster more strongly. The issue of "ecological niche" mentioned in the literature review is an example of a potentially interesting issue to investigate, looking for the presence of patterns in the specific contexts in which the concept changed, or to see if the conceptual basis was more grounded in an author's preference or a specific period of time. It is also on the basis of a comparative approach that investigations into the conceptual stability and profiling of species names will be studied.

If we consider other research to extend the Word Sketch feature, the other adaptable variable focuses on extending the Sketch Grammars, the rules used to identify grammatical and syntactic relations between words. This is the variable which can be used to identify semantic meaning and relations between concepts. This research has been performed by the Ecolexicon group, a research group at the University of Granada [128]. In contrast with the work done by McCarthy et al. [144], the Ecolexicon research looks at how the relations between words can inform us about semantic meaning, such as Hearst Patterns [90] can be used to do so. This is something that seems particularly relevant looking at the area of biodiversity, considering the hierarchy the tree of life is generally accepted to follow, and so represents a useful addition to the research thinking about the aims of the project here. In the first of the Ecolexicon pieces of research "preliminary results indicated that hyponymy subtypes were constrained by the ontological nature of concepts, depending on whether they were entities or processes" [74]. The idea of knowledge patterns [14] is central to their work. Knowledge patterns are short, domain-independent phrases that are used in formulaic ways to describe conceptual relations between words, based on Hearst Patterns and other syntactic patterns found in naturally occurring human language. They are clearly applicable to the present research, given the different relations between concepts of species and other entity mentions and also the interactions between them. This is one of the areas in which I believe I can overcome some of the issues the knowledge organisation initiatives are having in the area of biodiversity informatics.

Word Sketches, in a sense, could be considered human-readable word embeddings. Instead of the thousands or millions of features defined for word embeddings, the features in Word Sketches comprise the co-occurrence data of words in specific grammatical/syntactic contexts (please see the Methodology chapter for a more detailed description of the Word Sketch). This allows lexicographers to breakdown the different senses of a word in a particular corpus.

This chapter has set out the current research in the areas of data and knowledge management and representation. It has then proceeded to analyse the specifics of the biodiversity domain, in which it highlighted the particular nature of the domain as heterogeneous and reliant on integration for proper analysis as underlying reasons for the focus on this area. The specific qualities of the scientific nomenclature, the ambiguities in its use as a result of the complexities of biological taxonomy and classification and the use of this naming system as an index for biological data identify a real need to develop a method by which unstructured legacy data can be profiled and evaluated for integration. While current efforts attempt to develop means in the future to do this with semi-structured data, to the best of my knowledge, there are no efforts in place to do this with the legacy literature, or to look at actual usage of the terms themselves. The decision to apply a lexicography approach is linked to the need to be able to identify differences across different data sets, which may be not be expected. Precisely because of the changing and multiple positions on biological taxonomy, which impacts on scientific nomenclature usage choices, the entirety of the relations within a data set should be identified and any inconsistencies examined, instead of imposing one or other perspective or simply ignoring the data. The next chapter will provide details of the methodology and research design of the overall project.

Chapter 3

Methodology and general research design

The literature review outlined previous work in knowledge management and representation, firstly from an overall perspective and then more specifically in the fields that biodiversity encompasses. It outlined the role of ontologies in the field of biodiversity and obstacles relating to their application. Finally it presented an argument for taking a lexicographic approach applying corpus linguistics in response to the issues highlighted. It also set out existing research into extending Word Sketches, a central part of the research design for this thesis. The methodology chapter further explores the theoretical basis for the methodology, providing support for this choice. It also provides a breakdown of the theoretical and practical choices that must be made when pursuing research from this methodological standpoint, as well as an overview of different techniques used in corpus-based analysis. The second half of this chapter sets out the tools chosen on the basis of these considerations, and is followed by the methods applied throughout the thesis.

The research carried out in the development of this thesis draws from the design science field of thinking. Principally used in Information Systems research, design science applies an iterative method through which IT artefacts can be innovated and improved by learning about the environments within which they exist. Hevner [92, 93] describes the process of design science through "three closely related cycles of activities" [92]: the relevance cycle, the rigour cycle and the central design cycle. Design science essentially wants to improve the reality for people working in a certain area by introducing new and innovative processes or systems. The relevance cycle represents the application domain, the reality in which the application exists. In the case of my research this could be understood to be the study of biodiversity and the reality of many integration processes and initiatives that identify the problem of integration on the basis of the scientific nomenclature. The rigour cycle represents the scientific foundations and experience that informs the research project, which in this thesis is related to not only the detail of the processes underway within the area of biodiversity and data integration, but also the knowledge of corpus linguistics and lexicography applications, current practices and applications. The relevance and the rigour cycles iteratively feed into the design cycle, which subsequently feeds back into the previous cycles. The design takes place in the design cycle. The design science approach to the research design was chosen because of the way it allows for a data-driven approach and for input from the different aspects of the environment to be included: the data itself, the domain field and the approach field, which in this case consists of the biodiversity literature, the domain of biodiversity and biology, and the approach field which is lexicography and corpus linguistics.

This approach has been chosen to study nomenclature usage within the biodiversity literature because of the ambiguities outlined in the multiplicity of the nomenclature [66,68], coupled with a lack of a method by which to empirically analyse its usage. As has been outlined in the literature review, corpus linguistics and lexicography is used to look at word meaning in context [19]. Computational corpus linguistics and lexicography has highlighted the relativity of meaning in relation to its context and even questioned the premise of well-defined, clear distinctions between different senses or meanings [115, 163]. Corpus linguistics, in its capacity to search over large amounts of data to find patterns relating to semantic prosody or the syntactic limits of certain phrases or word clusters [163] has revealed some surprising results. The lexicographic approach, based on corpus linguistic methods, has been proposed because of its capacity to map taxonomic structures within empirical data (narrative texts). A research design structure has been employed because of the exploratory nature of the development, of, in this case, a method. The iterative cycles that feed back into the process, from the exploration of the data, development of the method, application of the method and then both technical and expert evaluation to validate and identify areas for further work. These steps will be described in Section 3.5. The next sections will look firstly at the technical aspects of corpus building and then secondly at choices relating to the analysis techniques available when embarking upon a research project in corpus linguistics or lexicography.

3.1 Lexicography and corpus linguistics method: creating a corpus

This section outlines the different considerations necessary when creating a corpus for a study. It is split into two parts: one, which focuses on the collection of data for analysis, and two, which focuses on how to process the data appropriately prior to the study for subsequent analysis. Corpus linguistics itself was introduced in Section 2.6.

3.1.1 Data collection and cleaning

A fundamental question when looking to embark upon a corpus linguistics project is how to build the corpus. There is no "one size fits all" when building a corpus. Depending on the intention behind a corpus, it will need to be larger or smaller. Dictionaries require a very large corpus because they focus on semantic meaning, and also aim to identify all possible meanings of a word. In contrast, corpora intended to study common grammatical phenomena or the terminology of a specialist domain, can be smaller [184]. To be a corpus, the text collection also has to be classed as "representative". Representativeness is dependent not only on size, but content. The aims of the analysis are essential in deciding these factors. If a study aims to investigate language usage in a niche domain, a sample of documents from one journal over a specific time frame may be suitable. If, in contrast, the aim is to compare American and British English in the academic literature, it may be necessary to gather together all possible documents from a variety of journals, to be able to compare usage between different publications and locations [158]. It is important that corpus creation is conscious and explicit to be able to balance the bias that is inherently present because in the end a corpus is (nearly) always a sample of a domain.

Increasing digitisation and the internet has resulted in a rise in the number of very large, web-scraped corpora. These aim to give insight into language usage unveiling patterns that simply would not be visible on smaller corpora. Web-scraped corpora, because of their size and collection method, follow slightly different rules. Their size is thought to outweigh some of the bias issues that may cause problems for smaller corpora if not meticulously compiled. However, large corpora also represent issues because of the variety of situations in which a phenomenon might appear, which might affect the results [118]. The decision on whether to build a small corpus or opt for a larger, web-scraped corpus depends very much on the defined aims of the analysis.

Having clearly defining the required size and content of the corpus, copyright and permissions issues must be addressed by properly researching and abiding by the rules in the reference country or countries [158]. What format the corpus in and what metadata to store in the files for reference during the research study are also important concerns that affect how you will be able to leverage the information within the corpus during the study itself.

It may be possible to collect texts which are already clean text files, but often they come in other formats (PDF, HTML). In the case of the former, documents that have been processed with OCR (optical character recognition) are usually messy and therefore should be cleaned to ensure accuracy of the OCR job, particularly on smaller corpora. Also, when producing web-scraped corpora the HTML should be removed to prevent the skewing of results. Repeated tags would skew the results, overly emphasising their importance within the dataset. There are now various tools that can help with this job, such as jusText. When creating a large corpus then it is also important to remove possible duplicate texts and anything else that may create bias in the corpus [112].

In short, when building a corpus you must ensure that you abide by the law, that the formats you use allow you to carry out your study effectively and that the choices as to the contents and extension of the corpus are defined and explained to ensure the data to be studied is collected in a scientific way. This will also allow for the identification of any potential biases or to help explain any phenomena throughout the study that may result from the choice of texts.

3.1.2 Corpus pre-processing

Corpus pre-processing involves different steps in which the corpus is annotated to provide information about its contents, which involves different types of processing, explained in the following sections.

Grammatical and syntactic natural language processing (NLP)

Before analysing the corpus in any way, some level of pre-processing must be applied. Lexicography typically requires basic-level NLP to be applied. This involves tokenisation (which splits the corpus into single units of words and punctuation), lemmatisation (which identifies and tags each word with a standard form of the word), sentence splitting and part-of-speech (POS) tagging (which tags each word with its POS in that specific context).

Semantic entity identification: named entity recognition (NER)

Some, but not all, corpus linguistic projects will require some form of semantic entity recognition. This project will also employ other levels of NLP and pre-processing to extend the use of Word Sketches. Named entity recognition (NER) is one of them, and is a technique used in NLP to identify proper nouns, such as people, organisations and place names, to support computational analysis of language. It is used in information extraction and data mining to identify entities of interest [79, 207] and also in tagging large corpora based on existing ontologies for search purposes [167]. Some automatic ontology creation projects also employ NER to identify concepts of interest [193]. There are a number of different approaches taken in NER, each of which tries to tackle different aspects of the problem. They are explored further in Section 3.4.2.

3.2 Lexicography and corpus linguistics method: analysis techniques

3.2.1 Keyword identification

A common corpus analysis technique is keyword identification. Keywords are words which appear significantly more frequently in a test corpus in comparison with a reference corpus [56]. This is a common technique in terminology identification in which the normalised frequencies of words of a specialist corpus will be compared with those of a general corpus such as the British National Corpus [22]. This is a very useful technique for identifying terminology specifically related to a specific domain.

3.2.2 Frequency and dispersion

Much of corpus linguistics analysis is based around frequencies, so it is important to know which ones to use and when. There are various different levels of straight frequencies: raw frequencies, normalised frequencies and the comparative ranking of words according to these frequencies. The most frequent words in a corpus can give you a broad idea about subject matter and patterns of word usage, and ranking across two corpora can tell you about the differences and similarities in word usage and content when comparing corpora. If the test corpora are two specialised corpora in different domains it can provide information about how terminology differs between these corpora [30, 163].

To be able to accurately compare multiple corpora, normalised frequencies should be used [30, 163]. Corpora usually differ in size, for which reason, corpora can only be compared accurately through relative frequencies. Frequencies are often normalised to hits per million, or per thousand words.

Straight frequencies cannot, however, tell you about the distribution of words across a corpus. To understand whether a specific term is characteristic of a whole corpus or concentrated in one area, it is necessary to calculate the distribution, or dispersion of a word [82]. Dispersion can be calculated in a number of ways. Range$_2$ is described as a simple, but fairly "crude" [30] dispersion measure, as it does not account for the number of times a word appears in different sections of the corpus. Calculations that offer greater levels of detail include the coefficient of variation (which has a focus on variation of distribution) or Juilland's D (which has a focus on evenness of distribution) [30]. Gries [82] has also proposed the Deviation of Proportions (DP) formula to overcome some of the issues identified in the two former formulae for not taking good account of the corpus part size variation. DP considers the expected distribution of a word or phrase in comparison with the observed distribution per corpus part and then calculates over the whole.

In deciding which of these calculations to use, it is important to bear in mind the purpose of the dispersion calculation. These calculations are often used to analyse the homogeneity of a corpus, or to make judgements as to the level of specialisation of terms. Different calculations will be more appropriate in different contexts. It is also important to consider what information is available in relation to the separate parts of the corpus.

3.2.3 Collocations and association measures

Much analysis within corpus linguistics focuses on collocations. Collocations are "combinations of words that habitually co-occur in texts and corpora" [30]. They can be based on frequency or also association measures. Association measures (otherwise known as collocation measures) are statistical measures that calculate the strength of the relation between words based on different features of said co-occurrence. Common association measures include Mutual Information Score, Dice, LogDice, and loglikelihood. However, scores such as Mutual Information and loglikelihood are often criticised for over-emphasising infrequent hits [30, 117], which can distort the results. Statistics such as MI2 (an adaptation of MI), Dice and logDice rectify this by shifting the focus to the exclusivity of the collocation, instead of rarity per se [30]. Statistical collocation measures on their own are limited because they tend towards a bag of words (BOW) approach, which does not take into account the importance of the surrounding syntax. As described in their paper, Bridging Collocation and Syntactic Analysis [214], "syntax-based approaches to collocation extraction focuses on the accurate selection of the candidate dataset in the first place [...] optimising the haystack and transforming it into a much smaller pile" (p.25). This is a very important consideration when choosing association measures, besides the semantic information that can be involved in syntactic patterns.

Word Sketches

The Sketch Engine corpus query tool (described in the tool section 3.4.1) has a special collocation feature called Word Sketch [117]. Word Sketches are a central focus of this research. They produce a one-page statistical overview of a word's grammatical and relational behaviour in a corpus by combining frequency and logDice statistics of collocations between word pairs, grouped by grammatical relation.

This feature is typically used by lexicographers as a basis from which to draft a dictionary entry, being used to evaluate the most salient uses of a word, such as typical collocations and contexts in which the word is used. It can be used to identify the key concepts a word is used to represent in specific contexts. This feature is particularly useful because of the combination between collocation association measures and syntax, the importance of which we have just explored, and the frequencies in which these collocations appear. The Word Sketches are produced by using Sketch Grammars. Both Word Sketches and Sketch Grammars will be further explored in Section 3.4.1 on tool selection.

3.3 Lexicography and corpus linguistics method: evaluation

There is a dual aspect to the evaluation in this thesis, firstly to validate the method and then to gain expert support for the relevance of the approach to real-life problems. In the case of this thesis, the evaluation forms part of the design science model, the relevance cycle. This applies both to the technical validation and evaluation, as in the feedback relates directly to the relevance of the method design and if it achieves what it sets out to do, and secondly the external, expert evaluation is used to evaluate the relevance and possible applicability in relation to the outside world, in this case the biodiversity domain.

3.3.1 Validation and technical evaluation methods

Precision and recall is a typical method for evaluating results in the computer sciences, for example the areas of information extraction and natural language processing [13, 159]. This approach focuses on the accuracy of the results provided by a computer analysis of some kind, taking into account the accuracy (% of right answers, precision) and the sensitivity (% of correct answers captured, recall). However, there are various instances in which this is not a suitable method or in which the method should be adapted to one's specific needs. A. Kilgarriff, Kovář, and Krek [114] used a variation of the precision and recall technique in their evaluation of Word Sketches, to make a comparison between the output of their Word Sketches and the actual dictionary entries from the Oxford English Dictionary (OED). The OED had used the Sketch Engine corpus query tool to mine for concordances of words in the edition of interest, but had not employed the use of Word Sketches. For the evaluation, the researchers applied precision in the typical way, automatically evaluating the senses that the Word Sketch had correctly identified. However, they used an expert to manually evaluate recall. Automated recall was not suitable because correct answers that were not provided in the Word Sketch version of the dictionary entry had to be considered.

Evaluation techniques used in automatic and semi-automatic ontology creation and other NLP tasks are also relevant here. These can be split into four main types, according to [11]:

- Gold standard-based evaluation
- Application-based evaluation

- Data-driven evaluation
- Human evaluation

A gold standard-based evaluation consists of comparing the creation (in this case an ontology) with an existing "gold standard". The created ontology could be tested against said gold standard for precision and recall (as described in the previous paragraph) of classes and instances, for example [148]. In this context, recall is sometimes called coverage [11,194]. Coverage focuses on whether there are concepts not covered in the corpora in the ontology and vice versa. The evaluation could also include concept alignment between the two ontologies. While applying some of the aspects of this approach could be useful in this research project to assess the preliminary results, there is no gold standard ontology for the data under scrutiny. There is no one accepted biological taxonomy. Checklists and other forms of knowledge representation are constantly evolving, as referenced in the literature review [175, 211]. One of the objectives of the PhD was to produce a semi-automatic evaluation method to compare the relations identified for precision, recall (quantitative measures) and differences (quantitative and qualitative measures) between the different expressions of knowledge. The method chosen to do this would have to be explained within the domain context.

Application-based evaluation, or task-based evaluation is evaluation of the appropriateness of a product to fulfil a specific task. For example, in the case of this research, while not an ontology, the results of the research could be given to a specialist in freshwater fish to evaluate the use of this tool to identify species mentions and the links they have between each other for search purposes.

Data-driven evaluation of ontologies is the most similar to the basic premise of the research study at hand. This is a process in which domain-specific knowledge resources are used to assess coverage of an ontology in a specific domain. In this case the research is focusing on the field of biodiversity and wants to measure the coverage of ontologies relating to the scientific nomenclature in comparison with what can be found in domain-specific corpora and if there are any underlying differences.

Finally human evaluation of ontologies is where there are criteria defined to evaluate different aspects of an ontology, such as richness, accuracy etc. and be evaluated by humans. This is time-consuming and costly and is not commonly used these days [11].

In this PhD thesis, there are two sides to the technical evaluation. Firstly is the validation and subsequent evaluation of the method developed as part of this research project, secondly is the evaluation of existing ontologies on the basis of the method developed within the research project. None of the above techniques can be used without adaptation as the aim of the research is not to create an ontology, per se, rather produce a representation of the data within the test corpus, whether it complies with the logical constraints of ontologies or not. By creating such a representation, the intention is to provide an evaluation of existing ontologies, using a mixed evaluation technique that will be described in Section 3.5.3.

3.3.2 Expert evaluation

Research does not exist in isolation and it is always important to engage with domain experts in the area of biodiversity and potential beneficiaries of the research. Expert evaluation is used to receive feedback as to the validity of my analysis and the applicability of any method developed. The evaluation can take various forms, such as interviews or focus groups. It is important to give stakeholders a voice in these spaces, and some autonomy in directing the conversation. Without this, the approach taken by the researcher cannot be questioned or cross-analysed properly. The focus group approach allows for the researcher to observe and analyse contrasting and sometimes changing opinions expressed by people as they interact and discuss the topics from their perspectives [25], while ensuring the researcher can still guide the questions through the items of interest to be able to get feedback on specific aspects of the research that are relevant. Therefore semi-structured interviews or focus groups would be suitable options. Researchers are considered to be central in the process of interviews, which tend to be one-to-one, whereas focus groups can be used where the aim is to gain insight through the fruits of the discussion generated with a group of people together [25]. In interviews, which tend to be one-to-one, the researcher takes the role of "investigator". This may be more suitable in cases where specific information is required from one expert. In contrast, in a focus group discussion, the researcher takes more of a facilitation role. The participants have greater freedom in the direction the discussion takes, because they are the central participants in the discussion [164]. This is be more useful in situations where a variety of opinions may help to produce new ideas and thoughts. The size of focus groups can vary, one study found varying degrees of participants from 3-21 [164]. Smaller groups are appropriate when input to truly understanding a matter, or if the participants are considered experts in that field for feedback as such [123]. Focus groups do not provide information that can be generalised to the wider population, but can help to better understand a problem [123]. In the latter case, and where more in-depth information from specifically selected experts is required, sometimes what are called mini-focus groups are appropriate, which can comprise from 2-4 people [164]. If the people have been chosen for specific perspectives, having smaller, more directed focus groups can enhance the amount of input provided by each participant, therefore providing a richer basis of analysis from that specific viewpoint.

Focus groups and semi-structured interviews are principally qualitative in their analysis, as the questions should aim to be open-ended and stimulate discussion [34]. Interviews will need to be recorded in some way and subsequently transcribed and if any specific quantitative results are required to identify more specific patterns this should be obtained by an accompanying questionnaire or a template for the focus group. The analysis often takes some form of thematic analysis [91], either manual or through the use of a qualitative analysis programme such as NVivo [181]. The analysis can take the form of deductive or inductive reasoning, depending on whether the researcher wants to be led by the themes identified by themselves or whether what comes out of participant data should lead the analysis [24, 91]. Deductive reasoning means the researcher will stay aligned with the themes identified prior to the group but also means that it may exclude themes perceived as important by the participants, which would benefit from a more inductive approach. Often the analysis can be guided in part by both, which allows the researcher to both keep hold of themes deemed important and also allow new ideas to emerge from the participants. However the analysis is performed, it should follow the steps set out by Braun and Clarke [24] to ensure the flexibility of the approach is respected, while also ensuring that there is enough structure to make it a reliable approach. This is also to ensure that analysis actually takes place [24].

3.4 Methods: tool selection and description

This section outlines the selection process of different tools used throughout the course of the PhD. Where necessary, information is also provided about the features that are essential parts of the research design to give the reader a background in their characteristics and usage.

3.4.1 Corpus query tool: Sketch Engine

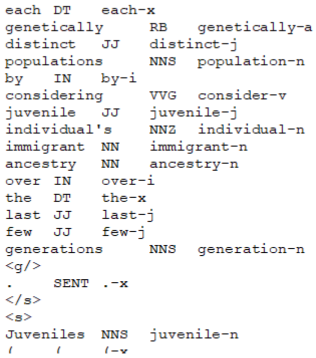
The Sketch Engine [116] was chosen as the corpus query tool for the project for a number of reasons. Sketch Engine performs the grammatical pre-processing as part of the corpus compilation process, whereas with other tools such as AntConc [9] the processing has to be performed separately, and is the most basic of corpus query tools presented here. #LancsBox [28] is another example of a corpus query tool, with more of a focus on statistical analysis and comparing different statistical measures. #LancsBox also has a feature which produces collocation graphs. However, this feature works only on specific words plus collocations or restricting for a specific part of speech. Sketch Engine was also the only corpus query tool identified with the Word Sketch function and the capability to add extra annotation to control relation extraction. This feature is a focal point of the research because of the way in which it combines different collocation association measures and word collocations.

The following sections explain different aspects of Sketch Engine and how they can be manipulated for use.

Text processing in Sketch Engine

The natural language processing (NLP) required for Word Sketches is tokenisation, lemmatisation, sentence splitting, POS tagging and lempos tagging. Lempos, which the lemma plus a hyphen plus a letter indicating its part of speech. Word Sketches use the lempos part of the file.

To compile a corpus in Sketch Engine, first upload the text files, at which point Sketch Engine automatically performs the above processing, ready for its subsequent analysis. Where necessary, as in the case of this research, the pre-processed corpus is available for download in a vertical file format for extra tagging to be added before analysis. This is a word per line (WPL) file with vertical, tab-separated columns for lemma, POS and lempos (see Figure 3.1).



```
each DT   each-x
genetically     RB   genetically-a
distinct   JJ   distinct-j
populations     NNS  population-n
by   IN   by-i
considering     VVG  consider-v
juvenile   JJ   juvenile-j
individual's    NNZ  individual-n
immigrant  NN   immigrant-n
ancestry   NN   ancestry-n
over IN   over-i
the  DT   the-x
last JJ   last-j
few  JJ   few-j
generations     NNS  generation-n
<g/>
.    SENT .-x
</s>
<s>
Juveniles  NNS  juvenile-n
(    (    (-x
```

Figure 3.1: Example of WPL file

Word Sketches

Word Sketches have briefly been introduced in Chapter 2 as well as earlier in this chapter (Section 3.2.3). Word Sketches form a central part of the research design and focus because of their unique combination of using collocation association measures combined with syntactical patterns and the frequency of these co-occurrences to produce summaries of word behaviour in context.

Word Sketches provide two numerical parameters in the measurement of a collocation between two words: frequency of hits and salience score. Frequency of hits refers to the number of times the relation between two words occurs in the corpus. The salience is the strength of the relationship between the two words in the context of the Word Sketch. As described in the Section 3.2.3, there are a number of statistical association measures that can be used to calculate the strength of a relation (collocation) between two words. Sketch Engine's Word Sketches used to be based on MI log frequency [186] and is now calculated with a version of the LogDice score [131], which is constructed on the basis of the Dice coefficient (one of the association measures mentioned in the methodology):

$$D = \frac{2fxy}{fx + fy}$$

where fxy represents the number of times two words (x and y) appear together, fx represents the number of times that word x appears, and fy represents the number of times that word y appears. The LogDice score definition is [186]:

$$LogDice = 14 + log_2 D$$

The plus 14 adaptation intended to increase the number size so the scores would be easier to handle [186]. The theoretical maximum and minimum are 14 and 0, respectively.

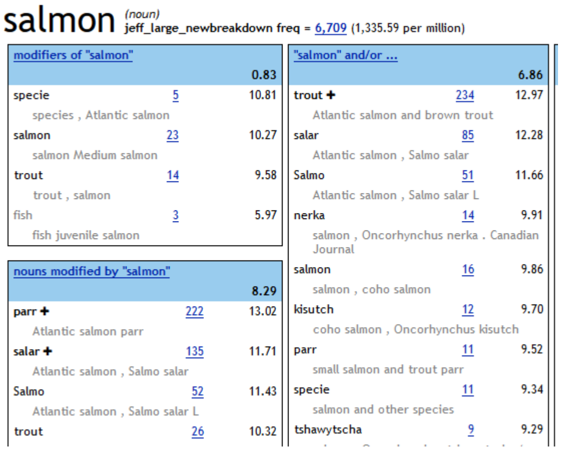
The equation works on the frequency of co-occurrence of the two words in question, divided by the sum of the frequency of each of them separately. The size of the corpus does not affect the outcome of the collocation score and so therefore can be used to compare collocation strength across various corpora [186]. It is claimed to be better that the MI (mutual information) score because it does not favour infrequent occurrences to the same degree. The fact that it can also be used to compare multiple corpora, means that in this research project it could also be used to look at consistency across different corpora of the strength of concordances or relations between different mentions.

However, as stressed in the methodology, statistical measures alone are not efficient at identifying collocations, as they only take into account statistical measures. Syntax is also an important variable when talking about collocation [214], and Word Sketches include this aspect through the Sketch Grammars (see the next section). In the case of this research, the focus is on specific collocations, which also follow specific syntactic patterns as a way of describing the semantic relationship between the two words.

Figure 3.2 shows a snippet of a typical Word Sketch, which is produced for manual evaluation by lexicographers when writing dictionary entries. The underlined numbers represent the frequency of hits, and the other column of numbers represent the salience of the relation, which will be explained in more detail later.

Sketch Grammars

To produce Word Sketches, Sketch Engine applies a rules-based method to identify grammatical and syntactic relations between words. Sketch Grammars are the name given to the file which defines these rules, and were developed as an alternative to full parsing. The rules are written in Corpus Query Language, which is a regular expression-type language used to perform complex



salmon (noun) jeff_large_newbreakdown freq = 6,709 (1,335.59 per million)

modifiers of "salmon"		0.83
specie	5	10.81
species , Atlantic salmon		
salmon	23	10.27
salmon Medium salmon		
trout	14	9.58
trout , salmon		
fish	3	5.97
fish juvenile salmon		

nouns modified by "salmon"		8.29
parr ✚	222	13.02
Atlantic salmon parr		
salar ✚	135	11.71
Atlantic salmon , Salmo salar		
Salmo	52	11.43
Atlantic salmon , Salmo salar L		
trout	26	10.32

"salmon" and/or ...		6.86
trout ✚	234	12.97
Atlantic salmon and brown trout		
salar	85	12.28
Atlantic salmon , Salmo salar		
Salmo	51	11.66
Atlantic salmon , Salmo salar L		
nerka	14	9.91
salmon , Oncorhynchus nerka . Canadian Journal		
salmon	16	9.86
salmon , coho salmon		
kisutch	12	9.70
coho salmon , Oncorhynchus kisutch		
parr	11	9.52
small salmon and trout parr		
specie	11	9.34
salmon and other species		
tshawytscha	9	9.29

Figure 3.2: Partial image of typical Word Sketch, which shows the different grammatical relations the keyword (salmon) has with other words in the corpus

queries on corpora [129]. Sketch Grammars identify rules which will extract word co-occurrence in particular grammatical relations, such as subject-verb, modifier plus modified noun, adjective plus noun. There are stock Sketch Grammars for many languages available on the web platform. Personalised Sketch Grammars can be created as required. An example given in the literature review would be of the Ecolexicon project [128], which employed Hearst Patterns [90] and other similar patterns to identify different types of semantic relations between words (see Section 2.7.2). The Sketch Grammar used in this research is based on the Ecolexicon grammar. Examples of both stock and adapted Sketch Grammars can be found in Appendix B.

Corpus Configuration Files

The Corpus Configuration file is what defines the qualities of the corpus, such as how many columns the vertical file will have and what the attributes of those columns will be. The stock corpus template, for example, accepts a vertical file with three columns: the word, the POS tag, and the lempos. The Word Sketch attribute (the attribute used as the condition for Sketch Grammar rules) is also defined in the configuration file. The Word Sketch itself will always return the lemma as the output (which is found in the lemma part of the lempos column). However, it is possible, for example, to add an extra column for personalised annotations to impose semantic or other restrictions on Word Sketch output. See Appendix B for configuration file examples, together with Sketch Grammars that illustrate that point.

3.4.2 NER tools options and selection

In choosing the NER tools, it was important to consider the different approaches available and the advantages and disadvantages of any approach.

Machine learning approaches Machine-learning methods use advanced statistical analysis techniques to automatically learn patterns from large collections of "training" data, to then detect new instances of those patterns in previously unseen data. They have the advantage of being quick and not requiring experts to study a dataset to develop specific rules. They can identify patterns that may be missed by a human and can discover new information which is not possible through dictionary approaches as a dictionary has just a fixed amount of information which is fed to the machine by the researcher. As with other methods, the results only match the quality of the data used, different types of data may be better or less suited to this approach, depending on how susceptible it is to algorithm analysis.

Dictionary-based approaches Dictionary-based approaches are knowledge-based and therefore accurate, as they rely on curated information provided researchers or domain-experts working on the project in question. Unfortunately, dictionaries are time-consuming to keep up-to-date and are limited to recognising entities that are included in the dictionary by doing pattern-matching. This means that any relevant entities within the corpus not included in the dictionary will not be recognised. The same occurs with spelling variations or differences - only those which exactly match the dictionary entries will be annotated. Dictionary look-up can be particularly good for entities that do not, for example, follow specific patterns or rules regarding their form, such as common names.

Rule-based approaches Rule-based approaches rely on human knowledge of patterns which are then used to apply algorithms according to these rules. They are also considered linguistic approaches because they are often based on linguistic patterns. The advantages to these approaches are that they can discover unknown information about the dataset (in contrast with dictionary-based approaches, so for example, names which follow particular patterns can be identified without knowing that that particular name appears beforehand). They do not require a training corpus and should be accurate as experts are involved in the development of the rules used to identify the entities of interest. However, rule-based approaches may accidentally exclude information through omission of relevant rules. They are also time-consuming to develop. In the field, this approach seems to generally be used in conjunction with other approaches.

Tool selection: Global Names Recognition and Discovery (GNRD) A number of tools were identified when searching for an appropriate tool to use in the research. NetiNeti [3] was identified as a tool that applies probabilistic machine learning methods to identify various ranks of scientific names, with a focus on recognising misspelled names, or those with OCR errors or other variations. In the area of biodiversity this capacity is important because the scientific nomenclature is commonly misspelled, and much of the legacy data has to be OCR'd to be digitised. However, unfortunately when trying to access this tool the link was broken, another of the issues with tools becoming inaccessible over time.

The Organisms and Species [166] tools is a dictionary look-up tool, which identifies taxonomic names and synonyms, binomials following Linnaean naming conventions, acronyms, common names, abbreviations. They also handle misspellings and other naming variations. However, I also discarded this tool because of the reliance solely on that included in the dictionary.

There are a number of tools that take a hybrid approach. TaxonGrab [120], is a tool based on using regular expressions, which flag up strings of two to three words that do not appear in the lexicon. The lexicon is a general language lexicon, excluding any scientific names as found in the Integrated Taxonomic Information System (ITIS) database [103]. Nomenclature rules according to Linnaeus are then applied. The advantage to this approach is that it can find names that do not appear in an existing database (or dictionary) but it will not identify any vernacular names in the texts. TaxonGrab has some multilingual functions but at the time of publishing the paper the functionality was limited due to using a limited dictionary look-up for languages other than English.

The Find All Taxon Names [187] tool also adopts a hybrid approach, using a mix of "structural rules, dynamic lexica with fuzzy lookups, and word-level language recognition" [187]. This tool gained the highest precision and recall of all the tools identified in [187] (over 99% for each), but in subsequent research on a different dataset (from a different domain to the original test data) both precision and recall suffered dramatically [3]. Reduced performance seemed to arise from misrecognising species' names with authorship and failing to recognise genus names, despite recognising the species' name, which discouraged me from using this tool in the research.

I finally decided to use Global Names Recognition and Discovery (GNRD) [179], a hybrid tool which combines TaxonGrab and NetiNeti. The tool only extracts scientific names but can be used on literature written in languages other than English. I chose it because of the access to the tool through an online web platform and an API through which you can upload text or PDF documents, which meant ease of access, as well as output as HTML or JSON list of all scientific names identified in said document, which meant that the information could easily be transformed for use within the investigation. While the tool only extracts scientific names, it can be used on literature written in languages other than English, which was thought to be useful should any non-English documents end up in test corpora. The tool was also mentioned frequently in the literature review and was the one chosen in at least one of the biodiversity informatics projects reviewed [208].

3.4.3 Network analysis and visualisation tools

This research looks at how Word Sketches can be adapted and then processed to access data relating to nomenclature usage in the biodiversity literature, transforming the data from a format in which it is manually evaluated to one which can partially automate the extraction of hierarchical structures and relations between these mentions.

In order to visualise the results in a way that would draw out network links through any hierarchy appearing, it was decided that a network graph would be a suitable way of visualising. This is similar to research in corpus linguistics that creates collocation graphs as a visualisation tool, because the modality of using collocation networks "indicates through different features of the visual display [...] the main properties of the relationship between the node and its collocates" [29], making it a more powerful tool than a collocation table. In this case to reveal the taxonomic relations between different nomenclature and common name variants in a corpus. Brezina [29] differentiates between the thesaurus network and these collocation networks (dictionary versus discourse). This research in this thesis sits somewhere between those two extremes, which will be discussed in the method design further.

To create these network graphs, a tool which included the ability to zoom in and out, to select or hide various nodes and relations according to different properties was needed. Two options were considered: Network X [86] and Cytoscape [165]. Network X is a popular library for network analysis in Python and appears in many of the data science courses looking at network analysis [6, 47]. However, while it is a very flexible library for network analysis, it is not recommended for good visualisation of networks as this is usually done through other programs.

Cytoscape [165] is a commonly used network data analysis tool for biological interaction networks. Cytoscape has features which allow you to filter data by edges, nodes, or characteristics of the former in order to analyse specific aspects of a network. Multiple different filter requirements can be applied simultaneously which is very useful when the analysis needs to focus on nodes and edges which are dependent on multiple criteria. There is also a feature through which it is possible to select nodes neighbouring other nodes, or edges adjacent to selected nodes, which can help to identify different patterns of connectedness through the network. This allowed me to manipulate the data easily in a way that would produce useful visualisations.

While this tool was originally designed for biological interaction networks, it can and is applied to general network analysis [195] and has been chosen for its capacity to discern and characterise the relations between species mentions and patterns within the corpora.

3.4.4 Tool selection conclusion

This section has described the tools chosen for the methods in this PhD thesis, as well as features of said tools, highlighting the aspects applied in this thesis to gain an understanding of adaptable features and requirements. The next section will describe the different stages of the final research design.

3.5 Methods: research design

The methods applied in the PhD thesis are iterative in nature. They have been informed by various iterations of research. These iterations were informed in structure by the design science methodology, which was described and justified at the beginning of the chapter and allowed for a process of learning in which different aspects fed into each other to produce a fit-for-purpose final product. In this case the process has only gone as far as to develop a method, there is no "product" as such. The literature review informed the first focus of study and the approach (part of the rigour cycle in the research design). This was then used as a basis to develop the subsequent phases, each of which was based on the findings of the previous phase.

3.5.1 Design overview

Phase 0 The pilot phase, part of the design phase, was used to explore the behaviour of different classes of terms within the first test corpus and to explore their behaviour in relation to trophic interactions. This phase represented work in preparation for responding to Objective 1 "model the hierarchy of relations between species that is identified within a specific corpus (by extracting the relevant information)" and Objective 5 "apply the above methods to interactions between species". After this initial exploration Objective 5 was not pursued further because the nomenclature issue was so complex as to constitute a thesis in its own right. The trophic interaction extraction constitutes further work beyond this thesis. The work flow presented in Figure 3.3 shows the process followed for Phase 0 of the project. The later phase built on this work flow, working to automate some of stages as necessary and add extra features to extend the scope of the research.

Phase 1 This was the main research design development stage. It focused on responding to Objectives 1 "model the hierarchy of relations between species that is identified within a specific corpus (by extracting the relevant information)" and 2 "create a graph/tree hierarchy image of this model to compare to the ontological structure for validation and evaluation purposes". This stage investigated how filter parameters and the separation or unification of nomenclature terms could affect the final output representation. The possibilities of using relation network graphs to disambiguate meaning was also explored. Figure 3.4 was developed through this stage for application.

Phase 2 Having developed the research design, it was necessary to evaluate its efficacy, which formed part of the relevance cycle as described earlier. Phase 2 focused on performing a validation and technical evaluation of the method design as presented here. This was to respond to Objective 3 "compare the relations identified for precision (quantitative measures) and differences (quantitative and qualitative measures" against an existing ontology. The evaluation was performed on the corpora processed as per the work plan in Figure 3.4, plus the transformation of an existing ontology into a suitable format to be compared against the test data extracted from the corpora.

Phase 3 The application phase, in which the methods developed were applied to a number of case studies. These were developed to respond to Objective 4 "perform comparisons between the hierarchies extracted between different corpora and ontologies of choice to evaluate conceptual (in)stability of nomenclature usage". These profiling studies incorporated a mix of more traditional corpus analysis methods relating to frequencies and dispersion, with the relation network representations developed throughout the rest of this project.

Phase 4 The final stage of the project, and also part of the relevance cycle. It involved a focus group comprising people who use nomenclature for various purposes in their professional lives. The focus group was used to gain insight into their understanding of ambiguity and usage of nomenclature, to contrast this with my interpretation of the data and also to gain insight into their opinions on my approach and interpretation of the data. The intention was to provide external validity to the project, on top of the internal validity assessed through the technical method evaluation, by speaking to domain professionals about the data extracted from my research and asking for feedback as to the accuracy of interpretations and potential applicability of the research in their work environments. It was also to provide outreach and contribute to the wider debate relating to nomenclature usage, and knowledge representation and integration, which is an essential part of any research.

3.5.2 Phases 0 and 1: Work flow development

As has just been mentioned, the pilot phase (Phase 0) follows the basic work flow found in Figure 3.3. All subsequent phases follow the basis work flow found in Figure 3.4, although the development of this work flow was fixed in Phases 0 and 1. The following will go into more detail about the different stages.

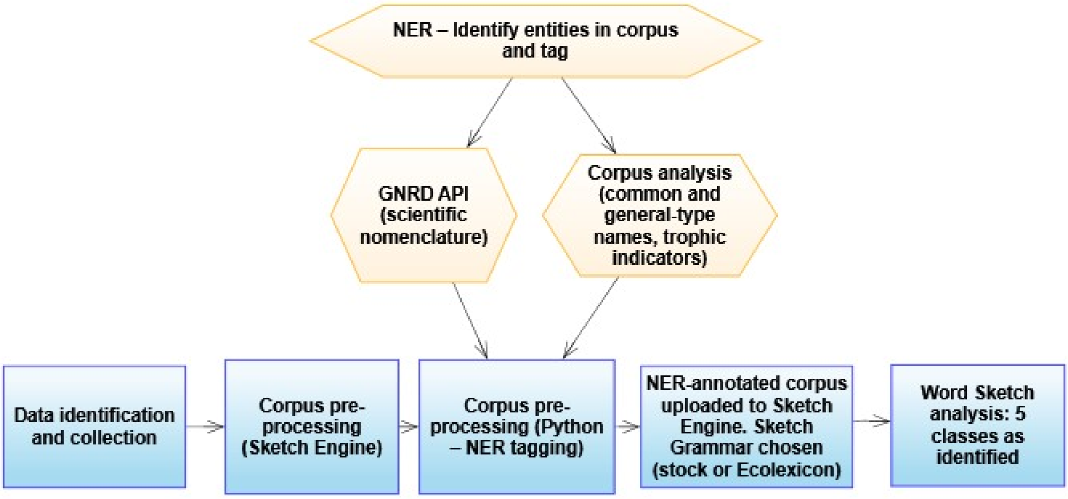

NER – Identify entities in corpus and tag
GNRD API (scientific nomenclature)
Corpus analysis (common and general-type names, trophic indicators)
Data identification and collection
Corpus pre-processing (Sketch Engine)
Corpus pre-processing (Python – NER tagging)
NER-annotated corpus uploaded to Sketch Engine. Sketch Grammar chosen (stock or Ecolexicon)
Word Sketch analysis: 5 classes as identified

Figure 3.3: Basic work flow - Phase 0

Data identification and collection

Phase 0 An academic corpus based on papers that fed into the Database of Trophic Interactions [80] was chosen for the pilot phase and was described on Zenodo, a data sharing platform (hereinafter called the Zenodo corpus). A total of 29 files were OCR'd and included in the corpus, which consisted of a total 351,435 words (types), 540,449 tokens (according to the Sketch Engine counter).

Phases 1, 2, 3 For Phase 1 and beyond, two corpora were identified and studied. Both of these corpora also have focus on freshwater fish. The first test corpus (JEFF corpus) comprised articles from the Journal of Ecology of Freshwater Fish (JEFF), was accessed through Wiley-Blackwell Journals and comprises 3,456,159 words (types), 5,023,230 tokens and 593 documents. It was collated using the Crossref TDM API, in accordance with the Wiley Online terms and conditions for text and data mining [220]. To make into a processable format it was OCR'd

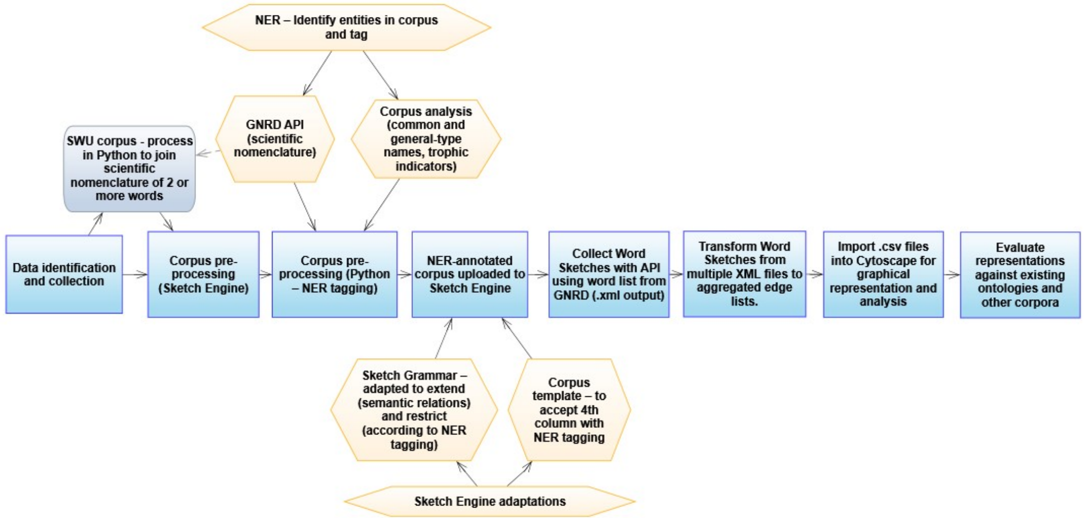

NER – Identify entities in corpus and tag
GNRD API (scientific nomenclature)
Corpus analysis (common and general-type names, trophic indicators)
SWU corpus - process in Python to join scientific nomenclature of 2 or more words
Data identification and collection
Corpus pre-processing (Sketch Engine)
Corpus pre-processing (Python – NER tagging)
NER-annotated corpus uploaded to Sketch Engine
Collect Word Sketches with API using word list from GNRD (.xml output)
Transform Word Sketches from multiple XML files to aggregated edge lists.
Import .csv files into Cytoscape for graphical representation and analysis
Evaluate representations against existing ontologies and other corpora
Sketch Grammar – adapted to extend (semantic relations) and restrict (according to NER tagging)
Corpus template – to accept 4th column with NER tagging
Sketch Engine adaptations

Figure 3.4: Basic work flow - subsequent phases

using the AntFileConverter PDF processing tool [8]. This corpus was used to analyse and develop the techniques relating to filtering and framing of the nomenclature profiles (Phase 1).

For the second part of this research, an extension of the JEFF corpus (called the WEB corpus) was created through web-scraping of seed words (see Appendix A.4) from the original JEFF corpus. The WEB corpus is 4,390,477 words (types) and 6,133,678 tokens. It contains many different sources found on the web, therefore it cannot be assumed to represent an academic, curated, well-informed authorship.

The WEB corpus was analysed alone and then used to compare and contrast with the JEFF corpus. Having two corpora allowed for comparative analysis in the technical method evaluation (Phase 2) to provide evidence of stability or lack thereof in the results produced as regards precision and patterns of differences. The JEFF corpus is comprised of purely published, academic journal articles while the WEB corpus could be comprised of anything matching the seed words published on the WEB. This also meant that in Phases 3 and 4 (application and evaluation phases), the analyses could serve to identify different patterns of behaviour or usage relating to the content of the corpora.

Corpus pre-processing: Sketch Engine

The research project required the corpus to undergo various stages of pre-processing. As mentioned previously, the pilot stage worked with the Zenodo corpus, the other stages of the research worked with the JEFF corpus to start and then incorporated the WEB corpus. Pre-processing of the corpora varied depending on each stage, outlined below:

Phase 0 (pilot phase) First stage processing in Sketch Engine as described in the tools section.

Phase 1, 2 and 3 Two versions of each corpus, to take into account the binomial quality of nomenclature. One version of the corpus was left in its original form, hereinafter called the original corpus. The other was adapted so that identified species' names were joined with an underscore (e.g. Anguilla anguilla is expressed as Anguilla_anguilla), hereinafter called unified term corpus. This meant that the corpus could be analysed both in its original form and with multiple word nomenclature being considered as single entities, to see how this impacted word characterisation.

When ready, each corpus was uploaded to Sketch Engine and processed as described in the Sketch Engine section above. It was then downloaded as a vertical WPL file (see Figure 3.1) to add additional semantic annotation as described below.

Corpus pre-processing: semantic annotation

Mark-up has been an important question within the research project, considering not only what entities to annotate but also a suitable schema and positioning for said annotation, in order to extend the Word Sketches to identify specific semantic concepts. The annotation requires first that the entities to be tagged be identified and second various choices have to be made as to the granularity, coverage, position of the annotation, as well as the methods employed to annotate the corpus.

The identification of entities to be tagged had two aspects: the identification of the scientific nomenclature in the test corpus and the identification of general-type and common names in the corpus. These were identified using different methods which were chosen because of the different coverage requirements in the research. These methods are described below.

Scientific named entity recognition Global Names Entity and Recognition (GNRD) was used in the research to identify the scientific nomenclature within the test corpora. These were accessed using the API function and were downloaded in JSON format for conversion into word lists in Python that could be used to annotate the test corpus in the lempos column (Phase 0) and fourth column (Phases 1, 2, 3) of the WPL corpus file.

Phase 0: GNRD was used as above to identify scientific names in the Zenodo corpus and then tagged accordingly.
Phase 1: GNRD was used as above to identify scientific names in the Zenodo corpus and then

Table 3.1: Phase 0: Annotation schema for the Zenodo Corpus

Common names (NCOM)	Life stage (NGENPRT)	Collective (NGENCOLL)
perch	nymph	species
trout	parr	specie
salmon	larvae	insect
waterstrider	larva	animal
strider	egg	fish
minnow		plant
roach		

tagged accordingly.

Phase 2: Main analysis based on tagging both the JEFF and WEB corpora with the list of names extracted from the JEFF corpus. This was to ensure that there would be sufficient names in common to perform the analysis.

A preliminary analysis was also performed of the precision and relations identified on each corpus according to the names extracted from the respective corpora themselves. Further work would constitute performing a full breakdown of the differences and similarities between the WEB and JEFF corpora when analysed through their respective lists.

Phase 3: The nomenclature profiling studies used the terms found in the ITIS, CoL and VTO, respectively (all scientific nomenclature variants and vernacular variants). This is to be able to make a comparison between the profile presented in the resource and that shown by the data.

Details of the name lists from each phase of the research can be found in Appendices A.5 and A.6.

Using corpus-based analysis to identify common names and general-type terms As mentioned in the Sketch Engine section, corpus-based analysis can be used to identify keywords in a corpus. This was performed to find common names and general-type terms in the corpora at different stages of the research, as will be described below, split by phase.

Phase 0: See Table 3.1.

Phase 1: See Table 3.2.

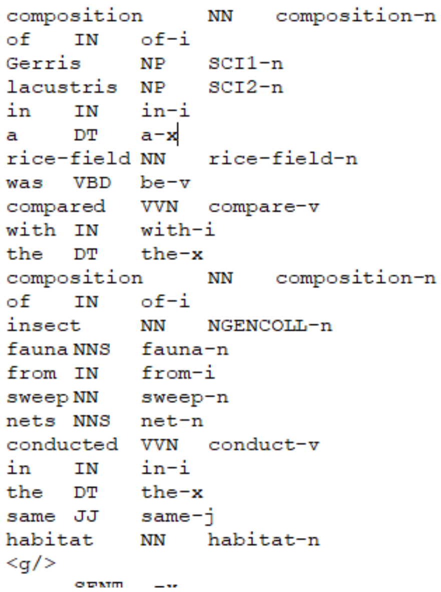
Phase 2: Only used scientific names as identified in GNRD and described in the previous section. The can be found in Appendix A.5.

Phase 3: The nomenclature profiling studies used the terms found in the ITIS, CoL and VTO, respectively (all scientific nomenclature variants and vernacular variants). These can be found collated in the Appendices A.6.

Table 3.2: Phase 1: Annotation schema for the JEFF corpus

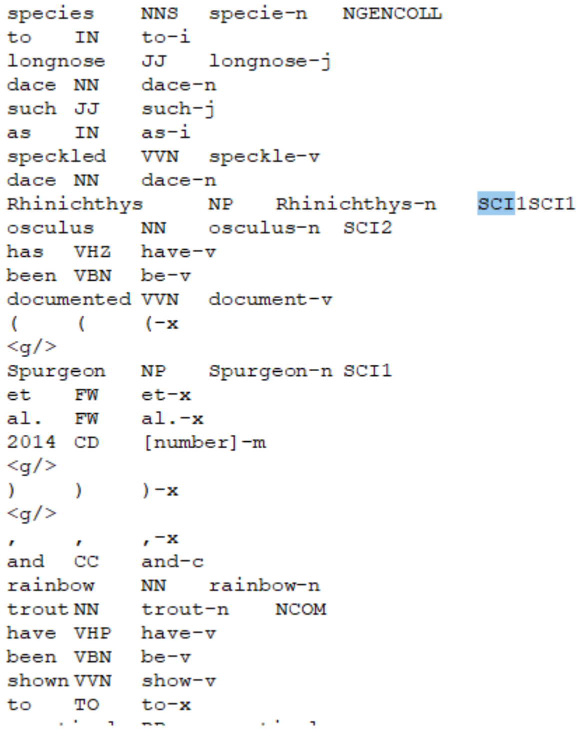
Common names (NCOM)	Life stage (NGENPRT)	Collective (NGENCOLL)
perch	nymph	species
trout	parr	specie
salmon	larvae	insect
eel	larva	animal
trout	egg	fish
chub		plant
stickleback		
goby		
whitefish		

Granularity As regards granularity, these entities were divided into five class-types: scientific names (two classes, for multiword terms), general-type words (collective and life stage classes) and common species' names (one class), which were chosen to be able to consider differing behaviour of terms from different class types. These tags were placed all in the same column as semantic markers where any of these entities were identified in the corpus, and were used by the Sketch Grammars to restrict Word Sketch output.



```
composition      NN    composition-n
of     IN    of-i
Gerris       NP    SCI1-n
lacustris  NP    SCI2-n
in     IN    in-i
a      DT    a-x
rice-field NN    rice-field-n
was    VBD   be-v
compared    VVN   compare-v
with  IN    with-i
the    DT    the-x
composition      NN    composition-n
of     IN    of-i
insect       NN    NGENCOLL-n
fauna NNS   fauna-n
from  IN    from-i
sweep NN    sweep-n
nets  NNS   net-n
conducted  VVN   conduct-v
in     IN    in-i
the    DT    the-x
same  JJ    same-j
habitat     NN    habitat-n
<g/>
       SENT   -x
```

Figure 3.5: Sample of tagging in Zenodo corpus in lempos column



```
species     NNS   specie-n    NGENCOLL
to     IN    to-i
longnose    JJ    longnose-j
dace   NN    dace-n
such   JJ    such-j
as     IN    as-i
speckled    VVN   speckle-v
dace   NN    dace-n
Rhinichthys      NP    Rhinichthys-n     SCI1SCI1
osculus     NN    osculus-n   SCI2
has    VHZ   have-v
been   VBN   be-v
documented VVN   document-v
(      (     (-x
<g/>
Spurgeon    NP    Spurgeon-n SCI1
et     FW    et-x
al.    FW    al.-x
2014   CD    [number]-m
<g/>
)      )     )-x
<g/>
,      ,     ,-x
and    CC    and-c
rainbow     NN    rainbow-n
trout NN    trout-n     NCOM
have   VHP   have-v
been   VBN   be-v
shown VVN   show-v
to     TO    to-x
```

Figure 3.6: Sample of tagging in JEFF corpus in fourth column

Annotation location In the pilot project the annotations were placed in the lempos column to identify behaviour patterns of the classes identified here. In this case, the class tag would replace the word originally in the corpus. All other words were left as they were (see Figure 3.5), so only obtained five Word Sketches were analysed in Phase 0 (to identify the patterns of behaviour of the class entities identified). In all subsequent stages of the research, these annotations were placed in a separate, fourth column where a keyword was identified, or blank where not (see Figure 3.6). This meant that classes could be used to restrict Word Sketch output (through relation rules) but obtained a fine-grained output in the Word Sketch. Word Sketches were then called according to the lemma (the individual nomenclature labels).

Tagging method The annotations were performed automatically through a Python script I developed. Manual annotation was not a feasible option on corpora of these sizes and as the annotation itself was based on a dictionary look-up script the miss-rate should be minimal.

NER-annotated corpus uploaded to Sketch Engine

After adding the semantic annotation, the annotated WPL file is then reloaded to Sketch Engine. This is where Phase 0 and subsequent phases diverge.

Phase 0 Semantically tagged WPL file is uploaded as a new corpus, choosing the Sketch Grammar as appropriate. In the beginning, the stock Sketch Grammar available as standard on the Sketch Engine website was used. Later in the pilot, the extended Sketch Grammar developed by the Ecolexicon research group [128] was used.

Phase 1, 2 and 3 Adapted Sketch Grammar used, which selects only the relevant relations from the Ecolexicon sketch grammar and also adapts to select only examples which have the appropriate annotation. Configuration file also adapted and uploaded to add the fourth column as an attribute so it is recognised in the Word Sketches.

See Appendix B for all Sketch Grammars and the Configuration File as adapted. Further information about Sketch Grammars and Configuration files can be found on the Sketch Engine page [130].

Sketch Grammar relations

Phase 0 In the pilot stage of the research all the Sketch Grammar relations were included in the analyses for a number of reasons:

- Both hierarchy of classes and trophic interactions were of interest (so both noun/adjective-noun relations and noun-verb relations)
- Part of the objective of the pilot project was to determine the relations of interest
- The analysis was being done manually

The pilot stage identified which of these grammatical relations could be mapped to ontological (hierarchical) relations for the purposes of the research. Table 3.3 shows the Sketch Grammar relations included in the analyses in subsequent parts of the research in this regard. In Phase 0 of the research, the relations identified in Table 3.4) were also included, as well as relations linked to trophic interactions (subjects or objects of particular verbs (subjects of "X" or objects of "X").

Phases 1, 2, 3 Phases 1, 2 and 3 included relations only shown in Table 3.3 because of the need for clarity as regards parent-child relations. The lemma formulation relation is too varied to be easily automated as it can represent many different types of relations.

Table 3.3: Use of Sketch Grammar relations and hierarchical meaning

WS gramrel as shown	Keyword (X)	Collocation
"X" is part of...	child	parent
modifiers of X	child	parent
nouns modified by X	parent	child
X has part...	parent	child
X is a ...	child	parent
X is a type of...	child	parent
X is the generic of...	parent	child

Table 3.4: Sketch grammar relations between nouns excluded in the 2nd part of the research

WS gramrel as shown	Meaning
X %(3.lemma) .../...	This relation is used to identify phrases involving the keyword and preposition. To be useful it would need to be made more specific.
X and/or ...	This represents where nouns are joined by "and" or "or" and can represent sibling relations but not exclusively.

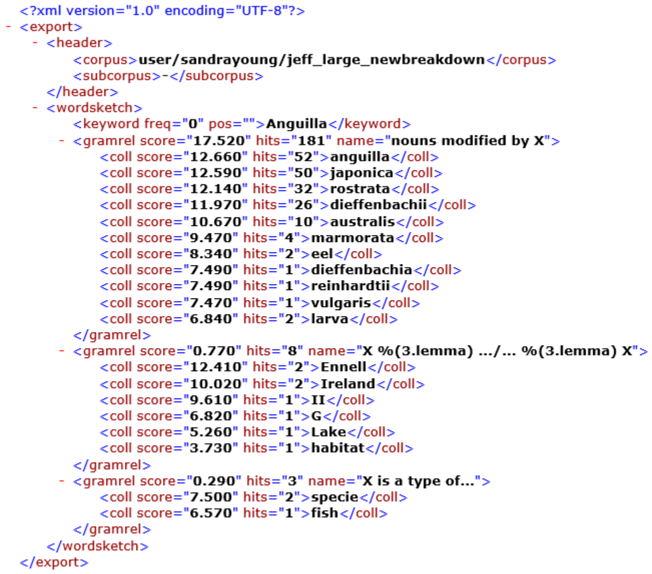
The binary parent-child relation was maintained throughout the whole project for clarity in identifying patterns and because the structures being studied involved the hierarchies within the scientific nomenclature, which are, at least on a superficial level, parent-child in nature. Further investigation into expanding the relations identified and incorporating more of the Sketch Grammar relations would constitute further work beyond the PhD thesis.

Collect Word Sketches through the API

Phase 0 Only five classes were analysed, so the Sketch Engine was queried manually to extract the Word Sketches needed.

Phases 1 and 2 The Sketch Engine can be queried through an API to collate Word Sketches, and in these phases, querying was performed through the API using a Python script developed for the project. The list of names collected through the GNRD tool were used to collate the Word Sketches of all scientific nomenclature reference identified in the corpus. These were downloaded in XML format. Figure 3.7 shows an example of a single XML for the keyword Anguilla in the JEFF original corpus.

In Phase 1 the JEFF corpus was analysed in two different scenarios (original JEFF corpus and unified JEFF corpus). The name lists used to pull Word Sketches varied in each case. The



```
  <?xml version="1.0" encoding="UTF-8"?>
- <export>
  - <header>
      <corpus>user/sandrayoung/jeff_large_newbreakdown</corpus>
      <subcorpus>-</subcorpus>
    </header>
  - <wordsketch>
      <keyword freq="0" pos="">Anguilla</keyword>
    - <gramrel score="17.520" hits="181" name="nouns modified by X">
        <coll score="12.660" hits="52">anguilla</coll>
        <coll score="12.590" hits="50">japonica</coll>
        <coll score="12.140" hits="32">rostrata</coll>
        <coll score="11.970" hits="26">dieffenbachii</coll>
        <coll score="10.670" hits="10">australis</coll>
        <coll score="9.470" hits="4">marmorata</coll>
        <coll score="8.340" hits="2">eel</coll>
        <coll score="7.490" hits="1">dieffenbachia</coll>
        <coll score="7.490" hits="1">reinhardtii</coll>
        <coll score="7.470" hits="1">vulgaris</coll>
        <coll score="6.840" hits="2">larva</coll>
      </gramrel>
    - <gramrel score="0.770" hits="8" name="X %(3.lemma) .../... %(3.lemma) X">
        <coll score="12.410" hits="2">Ennell</coll>
        <coll score="10.020" hits="2">Ireland</coll>
        <coll score="9.610" hits="1">II</coll>
        <coll score="6.820" hits="1">G</coll>
        <coll score="5.260" hits="1">Lake</coll>
        <coll score="3.730" hits="1">habitat</coll>
      </gramrel>
    - <gramrel score="0.290" hits="3" name="X is a type of...">
        <coll score="7.500" hits="2">specie</coll>
        <coll score="6.570" hits="1">fish</coll>
      </gramrel>
    </wordsketch>
  </export>
```

Figure 3.7: Sample of Word Sketch XML for Anguilla

original JEFF corpus treated each nomenclature reference (single word) separately, whereas the unified corpus combined nomenclature references to analyse nomenclature terms as one full nomenclature term (e.g. binomial nomenclature for a species name). When first analysing the original JEFF corpus, there was a lot of noise as a result of searching for Word Sketches based on the separate units that comprise a term in the scientific nomenclature. To overcome this issue, a subset of the full JEFF GNRD name list (see Appendix A.5 was used. This list collated names selected from the genus and species parts of the terms. When the analysis for the unified corpus was performed, the full name list (presented as unified nomenclature terms) was used for a number of reasons. Firstly, by design the name list would be different as in the unified corpus each nomenclature reference has been unified as a full nomenclature term (i.e. binomial nomenclature for a species name). Secondly, in the unified corpus, fewer relations were identified as a result of this unification. Unification resulted in more specific terms, which reduced the noise in the relations extracted. These practical reasons together with the exploratory nature of this phase meant that this was the decision taken for the information extraction in Phase 1.

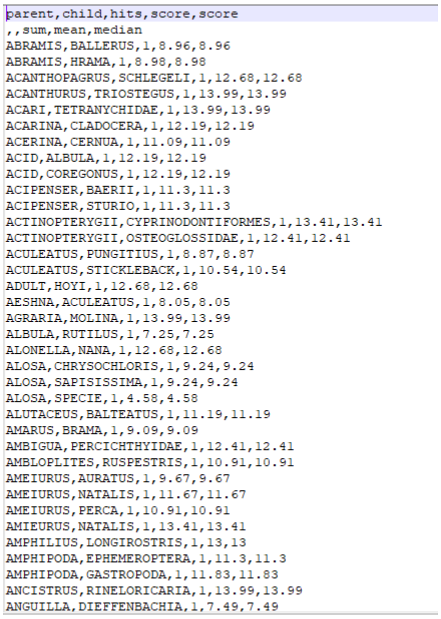
Phase 2 Word Sketch lists are described in full in Section 3.5.3.

Phase 3 The API was queried using the list of the names relating to the taxon/species in question (as identified through the VTO, ITIS and CoL), which included both scientific nomenclature and vernacular variants.

Transform Word Sketches

Phase 0 This step was not part of the pilot phase.

Phase 1, 2, 3 The development of a method to transform Word Sketches from the tables seen in Figure 3.2 to something that can be visualised in a graph structure was a focus of Phase 1. This process was then applied in Phases 2 and 3. Word Sketches were downloaded in XML format using the Sketch Engine API, then I developed a Python [180] script to convert them into comma-separated value (CSV) files. These files included the information of the keyword and collocate (defined as source and target, source referring to parent, target to child in the relation) and the number of hits and salience score of this relation. Figure 3.8 shows an example of the collated table, which is an aggregation of the relevant relations extracted from multiple Word Sketches like in Figure 3.7. These files were then used to create a Pandas DataFrame [146], (a table that can be manipulated within Python), which meant the data could easily be filtered as required for visualisation in Cytoscape.



```
parent,child,hits,score,score
,,sum,mean,median
ABRAMIS,BALLERUS,1,8.96,8.96
ABRAMIS,HRAMA,1,8.98,8.98
ACANTHOPAGRUS,SCHLEGELI,1,12.68,12.68
ACANTHURUS,TRIOSTEGUS,1,13.99,13.99
ACARI,TETRANYCHIDAE,1,13.99,13.99
ACARINA,CLADOCERA,1,12.19,12.19
ACERINA,CERNUA,1,11.09,11.09
ACID,ALBULA,1,12.19,12.19
ACID,COREGONUS,1,12.19,12.19
ACIPENSER,BAERII,1,11.3,11.3
ACIPENSER,STURIO,1,11.3,11.3
ACTINOPTERYGII,CYPRINODONTIFORMES,1,13.41,13.41
ACTINOPTERYGII,OSTEOGLOSSIDAE,1,12.41,12.41
ACULEATUS,PUNGITIUS,1,8.87,8.87
ACULEATUS,STICKLEBACK,1,10.54,10.54
ADULT,HOYI,1,12.68,12.68
AESHNA,ACULEATUS,1,8.05,8.05
AGRARIA,MOLINA,1,13.99,13.99
ALBULA,RUTILUS,1,7.25,7.25
ALONELLA,NANA,1,12.68,12.68
ALOSA,CHRYSOCHLORIS,1,9.24,9.24
ALOSA,SAPISISSIMA,1,9.24,9.24
ALOSA,SPECIE,1,4.58,4.58
ALUTACEUS,BALTEATUS,1,11.19,11.19
AMARUS,BRAMA,1,9.09,9.09
AMBIGUA,PERCICHTHYIDAE,1,12.41,12.41
AMBLOPLITES,RUSPESTRIS,1,10.91,10.91
AMEIURUS,AURATUS,1,9.67,9.67
AMEIURUS,NATALIS,1,11.67,11.67
AMEIURUS,PERCA,1,10.91,10.91
AMIEURUS,NATALIS,1,13.41,13.41
AMPHILIUS,LONGIROSTRIS,1,13,13
AMPHIPODA,EPHEMEROPTERA,1,11.3,11.3
AMPHIPODA,GASTROPODA,1,11.83,11.83
ANCISTRUS,RINELORICARIA,1,13.99,13.99
ANGUILLA,DIEFFENBACHIA,1,7.49,7.49
```

Figure 3.8: Sample of aggregated Word Sketches as parent-child relations

Visualise and filter edge lists

Phase 0 Word Sketches were manually obtained for the five classes of words described in the annotation section and manually evaluated the hierarchy revealed through the relations between the classes for further work.

Phase 1, 2 and 3 Word Sketches contain two numerical parameters, as described in Section 3.4.1, frequency and salience. The use of the LogDice calculation for the salience calculation, as described in the methodology section, avoids the issues related to over-emphasis of infrequent collocations, which are particularly relevant for lexicographic purposes. Lexicons are extensive, varied and heterogeneous, complying with what is known as Zipf's Law, which results in a long tail distribution [145]. Scientific nomenclature is no exception here. It was therefore hypothesised that salience, the association measure applied by Word Sketches, might help to highlight different or better collocations between different nomenclature variants than frequency alone.

Phase 1 focused first on identifying suitable methods to visualise the graphs. As described in Section 3.4.1, Cytoscape, a network analysis and visualisation tool [165], was chosen for this purpose. After having developed a method by which to create edge lists in the previous section, it explored the effect of manipulating frequency and salience filters on the output of the graphs. This is discussed in more depth in 4. Phases 2 and 3 then applied that learned in Phase 1.

3.5.3 Phase 2: Method validation and technical evaluation

Phase 2 evaluated the validity and success of the method design developed in this research through technical validation and evaluation techniques as set out in the work flow Figure 3.9. This process set out to respond to two questions:

1. Does the method developed in this research project do what it sets out to do?
2. How well does the method achieve its purpose?

The first question responds to the validation element of the process, whereas the second question represents the evaluation side of this technical process. The method is then applied through the nomenclature profiling studies in Chapter 6 and evaluated for usefulness by experts in Chapter 7.

The evaluation also looked to systematically analyse the difference between the frequency and salience filters, drawing on the preliminary results from Phase 1 and looking to draw some evidence-based conclusions as to the semantic qualities of the nomenclature pairs highlighted by each filter.

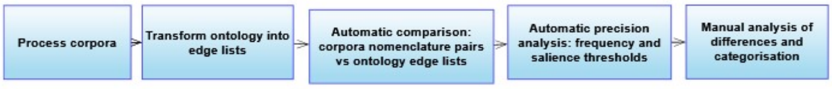

Process corpora
Transform ontology into edge lists
Automatic comparison: corpora nomenclature pairs vs ontology edge lists
Automatic precision analysis: frequency and salience thresholds
Manual analysis of differences and categorisation

Figure 3.9: Work flow: method validation and evaluation

This process involved using corpora annotated as per the steps described in previous sections of the research. It also required further processing steps in order to transform the ontological resource into a suitable format for comparison with the nomenclature pairs extracted from the test data.

The actual validation and evaluation analysis described in this chapter was set up in two stages. The first stage consisted of a precision analysis performed on both test corpora. To do this, I calculated the total number of nomenclature pairs (pairs of words identified as nomenclature that were related in the corpus) extracted from either the JEFF or WEB corpus that matched with a nomenclature term in the chosen ontology (precision). By comparing the extracted nomenclature pairs with the ontology I could evaluate the efficacy of the method developed in identifying nomenclature usage accurately in the test data. There is no gold standard ontology in the domain, nor are the test corpora expected to be representative of a specific domain, therefore the use of precision in the evaluation was supplemented with a detailed analysis and classification of the nomenclature pairs identified in the test corpora that did not appear as valid nomenclature references in the ontological resource. It cannot be assumed that differences between the ontology and the corpus data are errors on the part of the corpus or the method. Differences were therefore used to make assessments as to the completeness of ontological representations and also qualitative differences between different corpora and ontologies. Differences were evaluated according to defined criteria which will be outlined in Section 3.5.3 and on the basis of these differences an adjusted precision score, which took into account any false negatives, was produced. The details of all steps in this process will be outlined below.

Choices in the method validation and evaluation process

Choosing an ontology Finding a suitable ontology presented a number of challenges. The Journal of Ecology of Freshwater Fish covers many types of species, which when searching often were accounted for in different ontologies. This relates back to fragmentation issues highlighted in the literature review. Two ontologies were considered in the evaluation: the Vertebrate Taxonomy Ontology (VTO) [150] and the NCBI organismal classification ontology [162]. The VTO, as the name suggests, focuses exclusively on vertebrate species. The NCBI ontology includes all such organisms but has a focus on bacterial organisms and viruses, which are not a focus of this thesis. The bacteria and virus entries complicated mapping because of the variability of the scientific nomenclature in these other kingdoms, so the VTO was chosen because of the nomenclature coverage and the format it was available in. While the test corpora had many invertebrate references, fish (vertebrates) were the main focus of the content and therefore the main focus of the analysis.

The VTO [150] was chosen as a reliable resource, given the description of the resource in research included in the literature review. However, the plurality of the biological taxonomy and its linguistic representations means that there are many different characterisations of these terms in multiple different resources.

Test corpora In light of the fact that there is no one accepted representation [68,124,170,202] of the biological taxonomy or the nomenclature, the evaluation must reflect the descriptive nature of the research. One of the arguments for using the approach I have taken here is the ability to use empirical evidence to evaluate the usage of the nomenclature in real life, so two corpora were created with differing qualities to see if this demonstrated any differences in usage across different areas, namely a scientific, peer-reviewed corpus in contrast with a web-scraped corpus which may include texts from various different sources, perceived as less reliable than the peer-review corpus. A hypothesis was that the JEFF corpus would give more consistent, reliable results than the web-scraped one. The use of the lexicographic approach meant that while the methodology intended to automate part of the process, it was still always possible to return to the evidence and look at the reasons for the results, which provided a robust mechanism for ensuring that data was not just accepted at face value and reasons behind the numbers could be identified.

The use of two corpora also added weight to the validity of the evaluation process, because it allowed for a comparison and to evaluate cross-corporal stability in relation to the precision scores. Future work would constitute testing the methods on corpora from different domains or different families of species.

Corpora tagging and analysis scenarios Three analysis scenarios were applied to the corpora in the course of the analysis, each scenario applying the annotation and Word Sketch extraction schema as set out in Table 3.5. Overview analyses (described in Table 3.6) were performed on the corpora in all scenarios. These served to perform a validation analysis on the method in general and also select a scenario for the corpora tagged with the JEFF name list, but with the full Word Sketches for this list, and also the corpora tagged with their respective name lists. Then a full, detailed analysis (described in Table 3.7) was performed on the corpora tagged with the full JEFF name list, but with Word Sketches pulled from a subsection of this list. Links to the relevant name lists can be found in Appendix A.5.

The overview analyses served to provide an overview of the patterns of behaviour of the nomenclature pair references identified within and across the corpora, whereas the detailed analyses allowed for a more in depth look at where these differences were and permitted an evaluation of the method but also ideas as to where there may be points of interest for the application of the method.

Table 3.5: Corpus tagging and names

Scenario	**Analysis**	**Corpus**	**GNRD list**	**Word Sketch list**	**Corpus label**
1	Overview	JEFF	full JEFF	JEFF full	JEFF (JEFF, WS full)
1	Overview	WEB	full JEFF	JEFF full	WEB (JEFF, WS full)
2	Overview	JEFF	full JEFF	JEFF full	JEFF (JEFF, WS full)
2	Overview	WEB	full WEB	WEB full	WEB (WEB, full)
3	Overview	JEFF	full JEFF	JEFF subsection	JEFF (JEFF, WS subsection)
3	Overview	WEB	full JEFF	JEFF subsection	WEB (JEFF, WS subsection)
3	Detailed	JEFF	full JEFF	JEFF subsection	JEFF (JEFF, WS subsection)
3	Detailed	WEB	full JEFF	JEFF subsection	WEB (JEFF, WS subsection)

Table 3.6: Breakdown of overview analysis steps

Overview analysis
Number of relations identified overall
Number of relations identified: frequency filter
Number of relations identified: salience filter
Precision vs VTO: frequency filter
Precision vs VTO: salience filter

The detailed analysis was performed on the JEFF and WEB corpora which had been tagged with the JEFF GNRD list, pulling the Word Sketches for a subsection of that list. This was chosen for the detailed analysis to ensure sufficient crossover between the two corpora for a proper comparison, and also because it was identified that the subsection emphasised more frequent hits therefore there would be more reliable data to work with, along with a manageable amount of data to analyse manually. Detailed analyses of the corpora under the other conditions would constitute further work.

Transform ontology into edge lists

In this evaluation phase, the first step was to convert the ontology into the same format as the test data. In the research the edge list obtained from the VTO only included the basic backbone of the ontology, with no synonyms included. These synonyms have been accounted

Table 3.7: Breakdown of detailed analysis steps

Detailed analysis
Number of relations identified overall
Number of relations identified: frequency filter
Number of relations identified: salience filter
Precision vs VTO: frequency filter
Precision vs VTO: salience filter
Breakdown of differences (filter for 5 or more hits)
Breakdown of differences (filter for 4 or more hits and salience 9)
Breakdown of differences (filter for 4 or more hits and salience 10)
Breakdown of differences (filter for 4 or more hits and salience 11)
Adjusted precision for all filter thresholds analysed

for manually in the evaluation process and also used later in the study. Future work could look at including these synonyms in the automatic evaluation.

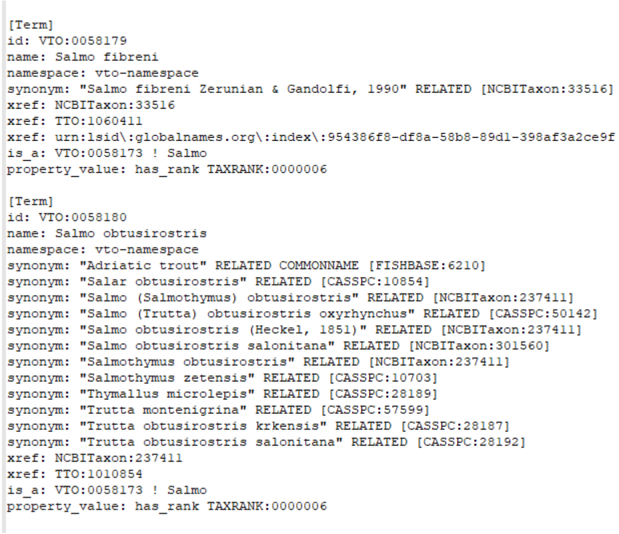
To convert the ontology, the Open Biomedical Ontologies (OBO) format file of the VTO (see Figure 3.10) was downloaded from the European Bioinformatics Institute website [53]. The OBO format is a popular formal ontology language and is used widely in biomedical and biodiversity domains. This version was chosen instead of the OWL (web ontology language) format because the latter uses links instead of terms within the actual document, which would complicate a direct comparison between terms.

This was then filtered using Python scripts to only include the sections "name:", "is_a" and "property_value", which was then used to convert to a CSV file which looked like Figure 3.11.

Finally, there were two more steps necessary. One to ensure that any two-word scientific nomenclature were separated, and were then labelled source (is_a) and target (name), respectively. Also the taxonomic rank was automatically converted to the string equivalent as detailed in the original ontology (see Appendix C). This was used to create a look up service so that source and target taxonomic rank of the matched pairs could be compared with the test corpus representation. This produced a final comparison, a section of which can be seen in Figure 3.12.

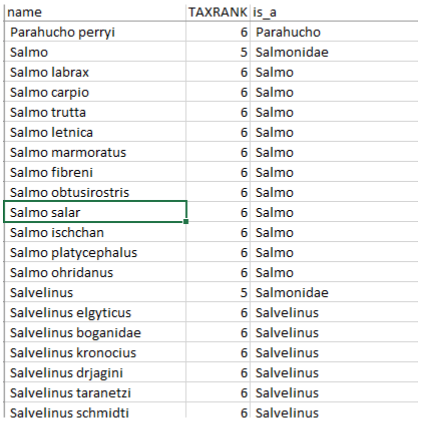
Main evaluation process: first stage

Comparison of extracted nomenclature pairs with ontology edge lists The comparison was based on matching edges (word pair relations) across datasets. These relations were extracted as described in method developed in Phase 0/1 (see Chapter 4). Figure 3.8 earlier in the chapter shows an example of the nomenclature pairs extracted and processed ready for comparison.



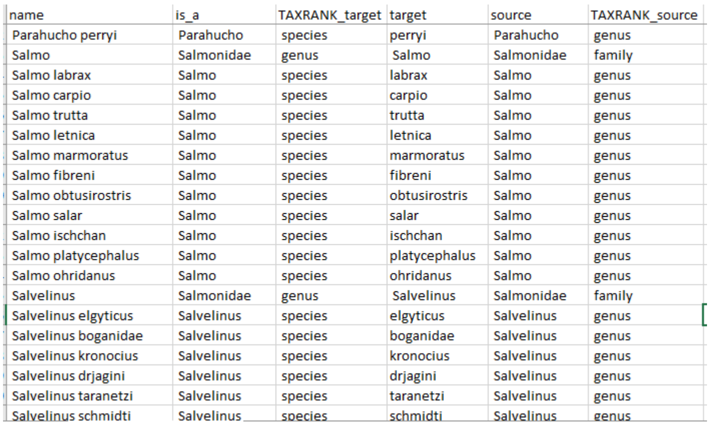
```
[Term]
id: VTO:0058179
name: Salmo fibreni
namespace: vto-namespace
synonym: "Salmo fibreni Zerunian & Gandolfi, 1990" RELATED [NCBITaxon:33516]
xref: NCBITaxon:33516
xref: TTO:1060411
xref: urn:lsid\:globalnames.org\:index\:954386f8-df8a-58b8-89d1-398af3a2ce9f
is_a: VTO:0058173 ! Salmo
property_value: has_rank TAXRANK:0000006

[Term]
id: VTO:0058180
name: Salmo obtusirostris
namespace: vto-namespace
synonym: "Adriatic trout" RELATED COMMONNAME [FISHBASE:6210]
synonym: "Salar obtusirostris" RELATED [CASSPC:10854]
synonym: "Salmo (Salmothymus) obtusirostris" RELATED [NCBITaxon:237411]
synonym: "Salmo (Trutta) obtusirostris oxyrhynchus" RELATED [CASSPC:50142]
synonym: "Salmo obtusirostris (Heckel, 1851)" RELATED [NCBITaxon:237411]
synonym: "Salmo obtusirostris salonitana" RELATED [NCBITaxon:301560]
synonym: "Salmothymus obtusirostris" RELATED [NCBITaxon:237411]
synonym: "Salmothymus zetensis" RELATED [CASSPC:10703]
synonym: "Thymallus microlepis" RELATED [CASSPC:28189]
synonym: "Trutta montenigrina" RELATED [CASSPC:57599]
synonym: "Trutta obtusirostris krkensis" RELATED [CASSPC:28187]
synonym: "Trutta obtusirostris salonitana" RELATED [CASSPC:28192]
xref: NCBITaxon:237411
xref: TTO:1010854
is_a: VTO:0058173 ! Salmo
property_value: has_rank TAXRANK:0000006
```

Figure 3.10: Snippet of the unaltered VTO OBO format file



name	TAXRANK	is_a
Parahucho perryi	6	Parahucho
Salmo	5	Salmonidae
Salmo labrax	6	Salmo
Salmo carpio	6	Salmo
Salmo trutta	6	Salmo
Salmo letnica	6	Salmo
Salmo marmoratus	6	Salmo
Salmo fibreni	6	Salmo
Salmo obtusirostris	6	Salmo
Salmo salar	6	Salmo
Salmo ischchan	6	Salmo
Salmo platycephalus	6	Salmo
Salmo ohridanus	6	Salmo
Salvelinus	5	Salmonidae
Salvelinus elgyticus	6	Salvelinus
Salvelinus boganidae	6	Salvelinus
Salvelinus kronocius	6	Salvelinus
Salvelinus drjagini	6	Salvelinus
Salvelinus taranetzi	6	Salvelinus
Salvelinus schmidti	6	Salvelinus

Figure 3.11: VTO filtered, and converted to CSV file



name	is_a	TAXRANK_target	target	source	TAXRANK_source
Parahucho perryi	Parahucho	species	perryi	Parahucho	genus
Salmo	Salmonidae	genus	Salmo	Salmonidae	family
Salmo labrax	Salmo	species	labrax	Salmo	genus
Salmo carpio	Salmo	species	carpio	Salmo	genus
Salmo trutta	Salmo	species	trutta	Salmo	genus
Salmo letnica	Salmo	species	letnica	Salmo	genus
Salmo marmoratus	Salmo	species	marmoratus	Salmo	genus
Salmo fibreni	Salmo	species	fibreni	Salmo	genus
Salmo obtusirostris	Salmo	species	obtusirostris	Salmo	genus
Salmo salar	Salmo	species	salar	Salmo	genus
Salmo ischchan	Salmo	species	ischchan	Salmo	genus
Salmo platycephalus	Salmo	species	platycephalus	Salmo	genus
Salmo ohridanus	Salmo	species	ohridanus	Salmo	genus
Salvelinus	Salmonidae	genus	Salvelinus	Salmonidae	family
Salvelinus elgyticus	Salvelinus	species	elgyticus	Salvelinus	genus
Salvelinus boganidae	Salvelinus	species	boganidae	Salvelinus	genus
Salvelinus kronocius	Salvelinus	species	kronocius	Salvelinus	genus
Salvelinus drjagini	Salvelinus	species	drjagini	Salvelinus	genus
Salvelinus taranetzi	Salvelinus	species	taranetzi	Salvelinus	genus
Salvelinus schmidti	Salvelinus	species	schmidti	Salvelinus	genus

Figure 3.12: VTO in CSV file format for comparison of corpus edge lists with VTO contents and ranking of nodes

Automatic precision analysis: frequency and salience thresholds As was explained in Section 3.4.1, salience and frequency were used as filters in the method for this research project because of the perceived value of using these parameters to emphasise different features of the relations between nomenclature pairs identified in the corpora. A preliminary exploration of the effect of these thresholds was performed in Chapter 4, which identified an emphasis on frequent references in the corpora through the frequency threshold and strong, specific collocations through the salience threshold. Salience focuses on very strong collocations, which are measured by the relative frequency of a pair of words, in comparison with the relative frequency of said words appearing with other words. Salience is designed to calculate collocational strength, so by design has a lexical element and is subject to Zipf's law. The test data in this thesis, the nomenclature, is also highly context specific. As a result not only do many of the relations occur only once, but each element of the pair only occurs once in the whole corpus. This means that salience tends to be very high because the terms on their own only occur in this context and needs to be considered when using salience as a filtering parameter.

Frequency and salience was used throughout the validation and evaluation of the method to try to gain stability in the results being extracted. The validation of the method itself firstly applied the technique of precision to measure how well the method identified and extracted accurate nomenclature pairs from the test corpora. This general precision analysis (overview analysis) studied each corpus as a whole, in the various annotation/Word Sketch schema scenarios described in Table 3.5.

In the overview analysis, these threshold filters were used to see if there was an "ideal" balance which is denoted by stability or clarity of concept identification (an optimum balance of correct relation identification to granularity of relation network). A preliminary exploration of the effects of these filters was performed in Phase 0/1 of the research (see Chapter 4). The comparison also intended to identify differences between the corpus data and the ontology, and categorising these differences to better evaluate the precision of the method, and highlight potential differences between corpus and ontological data. The purpose of this stage was to validate the method as being fit for purpose and also evaluate the qualities of the frequency and salience filters quantitatively.

Coverage (recall) was not applied in this research because the corpora were not supposed to be representative of the VTO, nor is there any gold standard in the domain.

Main evaluation process: second stage

After the overview precision analysis, which looked at the patterns of behaviour revealed by the application of the two filter parameters. The breakdown served to better understand the data, adjust the precision scores to account for any out-of-scope, valid nomenclature pairs identified by the method that were not included in the VTO and also qualitatively examine the differences between the corpus data results and the ontology, looking empirically at the examples extracted in context.

Difference breakdown and analysis The detailed analysis included a breakdown of differences between the corpus data extracted and the ontology. This was performed on the JEFF and WEB corpora in Scenario 3 (JEFF (JEFF, WS subsection) and WEB (JEFF, WS subsection)).

The differences were defined according to the following broad criteria:

1. Misspellings of real nomenclature variant
2. Recognised nomenclature variant missing from the ontology (scope or other)
3. Recognised nomenclature variant classed as synonym
4. Unknown (unrecognised nomenclature variant)
5. Incorrect (incorrect labelling of the direction of a relation, partial matches, one or both terms not from scientific nomenclature)

A full description of the breakdown and decisions made as regards these criteria are given in Chapter 5. The differences were evaluated through a mixture of quantitative and qualitative measures. Quantitative analysis was based around the idea of precision, with percentages of matches versus non-matches. After the breakdown of the above criteria, the false negatives, that is the nomenclature pairs correctly identified by the method that were not in the VTO, were included in the correct matches to produce adjusted precision scores. The qualitative measures analyse each difference in detail, evaluating the characteristics of the difference itself, then also examining the occurrences of said difference in the test corpora. Here the qualitative aspect of lexicographic corpus linguistic study is particularly useful, because of the ease in which it is possible to refer to the empirical evidence of specific examples to understand better understand the numbers [132], means that conclusions can be drawn as to the reasons behind trends or anomalies.

Dual threshold - frequency and salience The detailed analysis was performed both on the JEFF (JEFF, WS subsection) and WEB (JEFF, WS subsection) filtered with a single frequency threshold of five or more hits, but also with what I have termed the dual threshold, which consisted of filtering for frequency four or more hits and then salience of nine, ten and eleven, respectively (see Section 3.5.3 for the tabular breakdown). This was to provide an opportunity for a detailed analysis of the salience filter, considering the way that it maintained many, very infrequent relations which made a manual analysis without some other layer of filtering across the whole corpus impossible because of time and resource constraints.

By choosing the frequency of four, the salience filters could be used to compare against the other detailed analysis of frequency five or more hits. This meant that the detailed analysis could compare precision of the data when filtered using these thresholds as well as any divergence or convergence of the results to draw conclusions as to the respective characteristics of the different filtering parameters.

This was followed by a detailed breakdown of the differences between the data extracted from the JEFF and WEB corpora in just one of the annotation/Word Sketch schema scenarios (detailed analysis). Two scenarios in each corpus were studied in the detailed analysis, a frequency only filter and a dual threshold (frequency and salience filter). The frequency only filter breakdown was performed looking at the data from each corpus that had been filtered to only include nomenclature pairs identified as having 5 or more hits in each corpus. This was to be able to focus on relations with patterns and also make the manual analysis possible. The second scenario aimed to study the effect of salience on the results. To do this, a dual threshold was applied. The dual threshold involved mixed frequency and salience filtering, and is described in Section 5.3.3. This aimed to test the hypothesis that frequency and salience emphasise nomenclature with different characteristics in the test data. This dual threshold was used to filter out the most infrequent nomenclature pairs, which was thought to provide some stability by removing a large majority of the outliers. The salience filter was then applied on the filtered data set to test the hypothesis that the nomenclature pairs identified would diverge with the frequency filter-only scenario results as salience increased. A full explanation of these two scenarios is given in Section 5.3.3 and see Chapter 4 for information about the preliminary results. The next section goes into the details of the results of the whole validation and evaluation process.

3.5.4 Phase 3: Nomenclature profiling studies

This phase of the research looks at applying the methods developed to specific cases, in order to produce profiles of nomenclature behaviour. The corpus processing follows the same basic workflow as shown in Figure 3.4. However, the inputs are different and will be explained here. There are also other aspects to the analysis which must be mentioned. This phase is split into two parts: a comparison between different domain knowledge representation resources and subsequently the nomenclature profiling studies.

Comparison between different taxonomic resources

As has been made clear throughout the thesis, there is no one accepted representation of either the biological taxonomy or their nomenclature lists. This means that each taxonomic resource may present the information in a different way. The comparison between different resources serves to provide a perspective as to the variation of existing ontological and other knowledge representation resources, all supposedly authoritative figures in the organisation and recording of scientific nomenclature. Three different taxonomic resources were chosen to provide an overview of available resources by comparing and contrasting the following features:

- An analysis of the organisations involved in the production of the resource
- Stated purposes of the resource
- Resource structure and format
- Choices made by the resources as to the breakdown and classification of information

This comparison highlights the importance of framing of a subject. The different contexts in which these resources operate result in a need to present and categorise information differently for it to be fit for purpose. It also provides a backdrop for the profiling studies. This information is used to guide the analyses of the difference between the representation of each taxonomic entity chosen for study.

Nomenclature profiling studies

These studies look at three different accepted scientific nomenclature names and how they are profiled in both the JEFF and WEB corpora. For all three profiles the work flow follows that of the second stage of research, but with the following adaptations:

- Scientific nomenclature and common name tagging involves all variants identified across the three ontological resources identified, plus the ranking up through the taxonomic hierarchy for the chosen taxonomic entity based on the VTO (see Appendix A.6).
- In all cases the unified corpus pre-processing step was included because of the need to consider multi-word names as single units in this application of the method
- Both corpora were normalised to lower-case because it was recognised that some taxonomic mentions were being missed when analysing the corpus without having normalised for this factor. This was because the script used to join the multi-word units would miss some mentions if they did not follow the capitalisation rules for scientific nomenclature, for instance.

The profiling studies comprise the following:

1. Analysis and breakdown of different variants for each knowledge resource
2. Analysis of coverage of each corpus in comparison with each of the knowledge resources
3. Use of traditional corpus analysis techniques, including frequency, normalised frequencies, dispersion and statistical significance analyses to profile usage of the different variants across each corpus
4. Empirical analysis by means of concordances to check specific results
5. Relation network graphs used to visually demonstrate and identify key patterns of relations is performed, always with a focus on three different characteristics between the representation resources and the two corpora: consistency, gaps and disagreements or ambiguities

Frequency and dispersion as profiling The profiles include raw frequency, frequency per million and ranking, to be able to compare the usage of different terminological variants of the same taxonomic concept in the test corpora.

Frequency dispersion is then used (in percentage and graph form) to examine the dispersion of these terms, co-occurrence or not of specific variants of interest to draw conclusions or make judgements. Although it is often described as an inexact measure, it was chosen in this case because the range$_2$ percentage (see3.2.2) because it is sufficiently descriptive for my purposes, which are to:

- ensure that mentions are not concentrated in few documents despite having very high frequencies.

- calculate the comparative spread across documents to see if variants are more localised or generalised in relation to their overall frequency
- look at co-occurrence at document level to see if where there are occurrences they coincide or if there appear to be particular exclusive terminology usage patterns.
- use graphical representation of variant dispersion and co-occurrence provides a visualisation of the dispersion which includes the frequency of mentions per document, although without any calculation to account for document length.

The focus on term co-occurrence within the same document mean that it is important to maintain document boundaries because each text is an item that needs to be able to be considered in isolation for co-occurrence metrics.

Empirical analysis Qualitative empirical analysis was used to evaluate specific instances by means of concordances. Concordances can be used quantitatively, to count the number of times in which specific words or phrases appear and to count the patterns of the context in which they are found. However, they are also a window into the empirical basis of the analysis and a way to check results. This was employed to examine the basis of results, to find out more details about the context of the usage of specific terms that would not be apparent through frequency and dispersion alone.

Relation network graphs The information extracted from Word Sketches was used to produce these graphs, to visualise the network of relations produced through the syntactic collocational relationships between different variants of the nomenclature profile terms. These profiles were used to compare against the representations presented by each existing knowledge resource, as well as making a comparison between the representations provided by each test corpus, JEFF and WEB.

3.5.5 Phase 4: Expert evaluation and outreach

The external evaluation of my work was centred around a focus group method. The focus group and accompanying interview focused on general topics relating to the use of scientific nomenclature and variant usage as well as specific outputs from the nomenclature profiling studies. The discussion was data-focused, focusing on the group's perspective on the realities of nomenclature usage and variance in their working lives and their evaluation of the data extracted using my method to evaluate the validity of my assertions and the relevance of my method for future applications. This was important for me to be able to access feedback from the scientific community as to the usefulness of my approach to the analysis of scientific nomenclature and variant usage in the biodiversity literature. This method was chosen because of its suitability in gaining insight from participants and also allowing participants the freedom to interact and spark further guided discussion on the themes of interest according to terms that make sense for them [123].

The stated aims for the focus group and outreach day were:

Aims:

1. Provide an external evaluation of method and nomenclature profiling studies for both validity and applicability purposes
2. Contribute to the debate surrounding knowledge representation in the literature

Research questions:

1. What are biodiversity professionals' interpretations of the data analysis and visualisation performed in my wider research project?
2. What are biodiversity professionals' opinions as regards possible applications of my research?
3. How can my research contribute to the debate in the field and to tackling the problems identified?

The participants were chosen from suitable candidates at the Natural History Museum and the University of Brighton. The participation criteria were professionals with expertise in biology, biodiversity, specifically freshwater fish if possible, in either research or collections (archives), or informatics. These criteria were chosen because an understanding of scientific nomenclature usage and its complications was required. It was hoped to get a variety of people from different backgrounds because personal perception is likely to vary and could generate interesting discussion in the focus group. It was considered that the unifying aspect of an understanding of scientific nomenclature was sufficient to create a productive focus group [123].

The final list of participants consisted of two researchers from the University of Brighton and one software engineer from the Natural History Museum. A further participant, another researcher at the University of Brighton was unable to attend the focus group because of the Covid-19 crisis. The focus group also had to take place remotely as a result of the crisis. I held a more informal chat with the final participant. This was agreed with the participant before the focus group took place and deemed necessary to clarify some species specific questions: the participant unable to participate in the focus group was the only one with a specialisation in fish species. The group of participants was considered to represent a suitable mix of experience as all use scientific nomenclature and vernacular variants in their daily lives but each would have a slightly different perspective as to the importance and relevance of certain aspects of their usage and application because of the different focuses of their particular specialisations.

The study was split into three stages: a pre-focus group questionnaire (see Appendix F.1), the focus group itself (see Appendix F.2) and a post-focus group evaluation form (see Appendix F.4).

Pre-focus group questionnaire The pre-focus group questionnaire consisted of 29 questions split into five different sections: participant's role and area of expertise, their usage of knowledge representation resources relating to scientific nomenclature, scientific nomenclature usage, misspellings and variants, vernacular variant usage. The purpose of the pre-focus group questionnaire was to collate principally quantitative data relating to each participant's background and their opinions over these general areas to inform the design and questions to use in the focus group.

The data was analysed using descriptive quantitative analysis techniques. There were very few responses and it was not necessary to generalise to the wider population given the purpose of the questionnaire.

Focus group outline The focus group itself took place over the course of one afternoon (3 hours). As described, it combined a mix of outreach and focus group style questions to elicit responses and discussion from the participants relating to the issues of scientific nomenclature and vernacular variant usage in the scientific literature, as well as opinions relating to knowledge representation resources and my approach to the problem.

This session was audio and video recorded for later transcription and thematic analysis. In principle the focus group was going to be recorded only with audio, but due to the change from an in-person session to a remote session, this was switched to audio and video due to the technical issues related with recording only audio. Consent was gained before the recording and then confirmed in writing after the focus group. The slides included in the focus group are included in the appendices (see Appendix F.3).

The informal chat followed a similar outline but focused primarily on specific questions relating to the nomenclature variants relating to fish species.

Post-focus group evaluation questionnaire The post-focus group evaluation questionnaire consisted of two sections: one which focused on the findings of the research, how they were presented and if the participants' could envisage any relevant application of the method to their work. The second focused more on the outreach day itself. The evaluation questionnaire was sent via link to participants at the end of the focus group for completion, as with the pre-focus group questionnaire. It had been planned to ask participants to complete at the end of the day was because the focus group would still be fresh in their mind and also because it reduced the possibility of not receiving the input. This was not, however, possible due to the current circumstances surrounding the pandemic.

Analysis of focus group data

For the focus group analysis, I followed the steps as set out in [24]. NVivo was used for the task because it was a suitable thematic analysis tool for the task, particularly to help order the themes arising from the discussion in a coherent hierarchy. The program also has useful visualisation features. The small number of participants meant that some of the visualisations were not as useful as when working with larger datasets.

I was interested in the results of the participants not only to see if they concurred as to my analysis, but also to see if they agreed that the method could be applied in any way to their areas of work. Therefore part of the analysis was based on how the comments made by participants concurred or disagreed with my assertions in the results.

Part of the afternoon was more exploratory, in which I was exploring the participants opinions on certain areas. This was being used to explore how they organised issues relating to ambiguity and clarity in nomenclature usage. Therefore in the thematic analysis of those sections I was more inductive in my approach, to see if the same themes and organisation of themes came out as in my research, or not.

The other aspect of the research was more like outreach and evaluation of my method from their perspective, and so was a more straight-forward analysis to feed into the results and also to serve as validation for the research and my approach in itself.

The informal chat analysis was also included and analysed in conjunction with the analysis of the focus group in Chapter 7.

Chapter 4

Phases 0 and 1: Work flow design and preliminary results

Phases 0 and 1 constitute the main design cycle part of the thesis and aimed to better understand the test data to develop a method to respond to Objective 1: "Model the hierarchy of relations between units of nomenclature as used in a specific corpus (by extracting the relevant information)". In the context of the design science model, these phases refer principally to the design cycle. There was a pilot stage (Phase 0), as described in Section 3.5, which focused on the extraction of trophic interactions and nomenclature references. Five classes of words were identified in the pilot phase, which were then used to adapt Word Sketches to look at both noun hierarchies between different types of nomenclature reference and the interaction verbs that join them. During this stage a number of complexities involved in extracting and profiling nomenclature usage were identified, relating to shifting and multiple meaning of the different classes and the interaction between taxonomies and scientific nomenclature. Therefore the subsequent phases (Phase 1 and beyond) of the thesis focused specifically on nomenclature profiling. Trophic interactions would constitute further work.

The first section of the chapter relates to Phase 0. It outlines the identification of classes of words within the concept of nomenclature and the study and categorisation of linguistic patterns which denoted semantic relationships between these classes. This formed the basis of later stages of the research in building the relation network graphs. Preliminary work relating to trophic interactions took place in Phase 0 and is also described here. While not pursued in this thesis, there were some important findings that are relevant for future work.

The subsequent sections of the chapter relate to Phase 1, where relation network graphs, the graph visualisations developed for profiling in this research, are introduced. There are two sections relating to relation network graphs: one which introduces the filtering mechanisms used to adjust their characteristics, and another which looks at how specific characteristics of the graphs can be used for disambiguation purposes. Finally, there is a section which considers how framing the data in different ways affects the result of the profiling. These sections involve the preliminary findings of the research which were used to define the method used.

4.1 Phase 0: Extraction and representation of nomenclature references and trophic interactions

The initial phase of the research focused on identifying patterns of behaviour between different classes of words in the test corpus, including the relations between nomenclature-related nouns, as well as the patterns of usage as regards specific verb forms that indicated trophic interactions. The pilot stage used the Zenodo corpus (see Section 3.5.2) in its explorations and annotation in the lempos column of the Sketch Engine vertical WPL file meant that Word Sketches could be produced on the classes defined.

Five classes were defined as relevant to the identification of nomenclature classes:

- SCI1, which was used to define first or one word scientific nomenclature references (genus or higher)
- SCI2, for second word scientific nomenclature references (species level references)
- NCOM for vernacular, or common variant terms
- NGENCOLL for general collective terms such as species, fish, etc.
- NGENPRT lifestage terms such as egg (see the annotation schema in Section 3.5.2)

On top of the five noun classes, there was one (TI) which was used to denote a trophic interaction, represented by “eat”, “feed”, “consume”.

The process is summarised here (a full description can be found in Section 3.5.2):

- Identify suitable corpus (Zenodo) and process (OCR)
- Upload corpus to Sketch Engine and download processed WPL file (with POS tagging, etc.)
- Perform NER tagging using the classes identified (with classes in the lempos column)
- Upload tagged corpus to Sketch Engine
- Study Word Sketches manually to identify hierarchies and patterns in trophic interactions

4.1.1 Class hierarchy

Phase 0 was used as the preliminary step in the research to see if hierarchical relations between different levels of nomenclature references (ranking of scientific nomenclature, or from more general to more specific terms) could be extracted using adapted Word Sketches. This line of investigation was followed because of previous work relating to the principle of Hearst Patterns [90] and also the Head-Modifier principle [95], explained in more detail in Chapters 2 and 3.

Word Sketches for the five classes were downloaded in .CSV format. An example can be seen in Table 4.1, in which the collocations of SCI1 broken down into relation patterns can be seen. On the basis of these Word Sketches, a breakdown of the patterns of behaviour for the different classes in relation to other classes was developed.

Table 4.1: Example of SCI1 Word Sketch with breakdown of relations and collocations

keyword	**pos**	**name**	**hits**	**score**	**coll**	**hits**	**score**
SCI1	-n	modifiers of X	1376	17.68	sci1	490	11.45
SCI1	-n	modifiers of X	1376	17.68	genus	22	8.98
SCI1	-n	modifiers of X	1376	17.68	large	31	8.93
SCI1	-n	modifiers of X	1376	17.68	hydrophilidae	20	8.87
SCI1	-n	modifiers of X	1376	17.68	family	19	8.76
SCI1	-n	modifiers of X	1376	17.68	larval	19	8.71
SCI1	-n	modifiers of X	1376	17.68	spp.	10	7.85
SCI1	-n	modifiers of X	1376	17.68	sci2	11	7.67
SCI1	-n	modifiers of X	1376	17.68	pooled	6	7.12
SCI1	-n	modifiers of X	1376	17.68	mm.	6	7.12
SCI1	-n	nouns modified by X	4332	55.66	sci2	926	12.4
SCI1	-n	nouns modified by X	4332	55.66	sci1	490	11.45
SCI1	-n	nouns modified by X	4332	55.66	sp	390	11.36
SCI1	-n	nouns modified by X	4332	55.66	ngenprt	199	10.27
SCI1	-n	verbs with X as object	228	2.93	ti	32	10.43

Table 4.2 provides the different relations of interest identified between the different classes, according to the Word Sketch grammatical and semantic relations.

Relations of interest were those identified as being hierarchical in nature and were assigned as parent-child relations accordingly. In the case of this research, the decision was made to conflate all the relations as parent-child because of the main focus on nomenclature usage in

Table 4.2: Hierarchical relations identified by class in Word Sketches

Annotation	Relationship described	Related annotation
NGENCOLL	"the generic of"	SCI1, NCOM, NGENCOLL, NGENPRT, SCI2
NGENCOLL	"has part"	SCI1, SCI2
NGENCOLL	"as modifier of"	NGENCOLL, NGENPRT
NGENCOLL	prepositional phrases	SCI1 (e.g. species of – "the generic of" type relation)
SCI1	"modifiers of"	SCI1 (often misidentified lists)
SCI1	"modifiers of"	SCI2, sp., NGENPRT, L., spp. (species to genus relations – "generic of" type relation)
SCI2	"modified by", "is a type of"	SCI1 (genus to species relations); NCOM (latter show explicative relations where the common name is given followed by the scientific name)
NCOM	"is a"	NGENCOLL
NCOM	"modifies"	NGENPRT (explicative)

and between families. Future work would look at separating out the relations to provide a more granular view. For visualisation purposes, these results were converted into simple graphs manually (Figure 9.1) which show the hierarchies that can be identified in the data. The arrows show the relation from parent to child (arrow indicating the child of the relation). The yellow nodes indicate the classes as tagged and show the different paths of the hierarchy in according with the relations identified. To clarify the following hierarchies can be identified here as a result of the Word Sketches extracted:

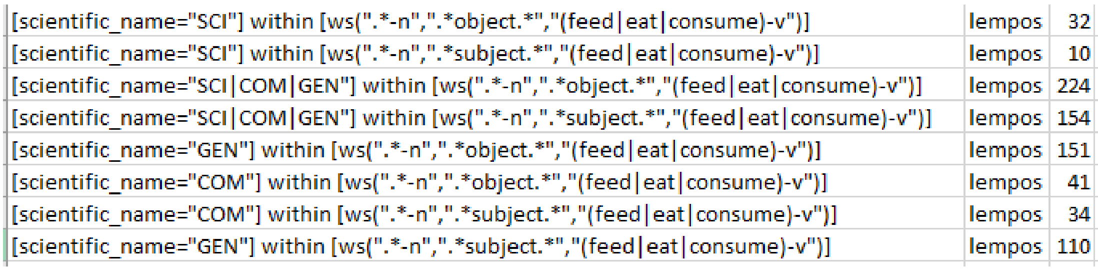
- NGENCOLL-SCI1-SCI2
- NGENCOLL-SCI1-NGENPRT
- SCI1-SCI1 (SCI1 is broad (it can relate to any term from genus level up the taxonomic hierarchy)

It is worthy of note that the relations identified with NGENPRT passed through SCI1 rather than SCI2. NGENPRT and NGENCOLL only formed part of the tagging for this preliminary stage of the research. The nodes not highlighted in yellow were those that came out because of some abbreviations relating to nomenclature usage but were not included here.

These graphs formed the basis of the visualisation throughout the rest of the investigation, the difference being that these classes were used to control the Word Sketch output and the individual nomenclature references appeared in each of the nodes, as can be seen in Section 4.2. The next section describes the preliminary investigation into trophic interactions.

4.1.2 Trophic interactions

Trophic interactions were explored in Phase 0, in the same way as the noun class hierarchies were explored in the previous section. The literature review (see Chapter 2) highlighted many complications in the reality of biological taxonomies and nomenclature usage, and the preliminary investigations in Phase 0 confirmed that profiling nomenclature would be sufficiently complicated to warrant a whole thesis, so the interaction work ceased after the pilot phase of the research. Despite this, the preliminary exploration revealed some interesting results as regards interactions. Vernacular variants and general terms were clearly preferred terms compared to scientific nomenclature when looking at direct subjects and objects in relation to interaction words (verbs such as “eat”, “consume” and “feed”). The actual numbers identified were calculated using CQL queries, as shown in Figure 4.1. This clearly shows the heavy weighting towards common (vernacular) and general terms in finding direct links to trophic interactions identified in the corpus.



[scientific_name="SCI"] within [ws(".*-n",".*object.*","(feed|eat|consume)-v")]	lempos	32
[scientific_name="SCI"] within [ws(".*-n",".*subject.*","(feed|eat|consume)-v")]	lempos	10
[scientific_name="SCI|COM|GEN"] within [ws(".*-n",".*object.*","(feed|eat|consume)-v")]	lempos	224
[scientific_name="SCI|COM|GEN"] within [ws(".*-n",".*subject.*","(feed|eat|consume)-v")]	lempos	154
[scientific_name="GEN"] within [ws(".*-n",".*object.*","(feed|eat|consume)-v")]	lempos	151
[scientific_name="COM"] within [ws(".*-n",".*object.*","(feed|eat|consume)-v")]	lempos	41
[scientific_name="COM"] within [ws(".*-n",".*subject.*","(feed|eat|consume)-v")]	lempos	34
[scientific_name="GEN"] within [ws(".*-n",".*subject.*","(feed|eat|consume)-v")]	lempos	110

Figure 4.1: Zenodo corpus: CQL query calculation of classes identified in trophic interaction

Figure 9.2 provides a graph representation using the Cytoscape tool employed later in the research to demonstrate how in future research further links could be identified. This representation was produced solely to aid visualisation of possible future avenues of investigation (links between trophic interaction words, through common and general terms, to scientific terms). The nodes now represent examples of nomenclature terms from the corpus instead of classes and the node “consume” represents all trophic interaction words tagged. No arrows were included in this graph because of the way trophic interactions have been identified through pairs in the Word Sketches. The different colours and sizes of the nodes are related to neighbourhood connectivity and closeness centrality measures that are used later in the thesis but are not relevant here. The relevance of this graph is to see the links, even at this nascent stage, between the interaction word, general and common terms through to scientific terms. These considerations were used to inform the analysis of the differences between scientific name and vernacular variant usage throughout the rest of this thesis and are relevant to some of the comments made in the expert evaluation in Chapter 7.

4.2 Phase 1

Phase 1 consisted of working to develop further insight into the representation of a simple hierarchy of relations between nomenclature references for later nomenclature profiling. Nomenclature profiling is the profiling of a specific nomenclature term using its relation with other terms in a specific corpus. Phase 0 demonstrated how the hierarchy of relations between classes could be modelled. These classes and relations between the classes identified in Phase 0 were used as parameters by which to restrict Word Sketch output to these entities. The switch from class nodes to individual nodes resulted in the production of more complex graphs, and as a consequence, the potential for more complex hierarchies. Phase 1 focused on the production of these more complex graphs, to produce what have been called relation network graphs because of the way they create a network of relations in a graph form. The process to do this is described in detail in the Research Design (see Section 3.5). It focused on automating the Word Sketch capture through an API because of the increased number of Word Sketches and the compiling of these as edge lists which represented the collocated terms and the relations between them.

The sections that follow focus on four different variables that have been manipulated to highlight different qualities of the data in the production of these graphs. The first two are related to filtering: frequency and salience. The second two are related to how nomenclature usage is analysed: scientific nomenclature terms often comprise multi-word units. The data has been analysed from two different perspectives: one in which all the words from a term are considered as separate units (original), then another in which the multi-word terms are joined by an underscore so they are considered one (unified). A comparison was then performed to see how this affects the results.

4.2.1 Filtering using relation network graphs

Word Sketches, as described in Sections 3.2.3 and 3.4.1, provide a summary of a word's behaviour in context using syntactic patterns and collocations. The statistical information given as part of the Word Sketch is the salience (statistical association) of the relation (syntactic collocation) plus the frequency at which the two words appear in that same grammatical construct across the corpus. Salience is calculated by using logDice, as described in Section 3.4.1. Importantly, the salience measure is defined by the specific Word Sketch because it refers specifically to the relation between the two words in question within the test corpus, while the frequency is simply the number of times that relation appears in the corpus. The combination of these different features therefore provides a multi-faceted summary of a word's behaviour for analysis.

These identify different aspects of a word's behaviour in a particular corpus plus the weighting of these features, much like word embeddings identify features of a word's behaviour in context in the calculation of its vector. Section 4.1 outlined how I identified semantic relations between classes of nomenclature according to the syntactic collocations. This section considers how the two weighting measures, salience and frequency, can be used to filter the relations extracted from all the Word Sketches extracted from the corpus in an attempt to find a suitable way to present the information and how these two measures affect said representation. Full details of the Word Sketches extracted in Phase 1 can be found in Section 3.5.2. The characteristics of these measures as described in Section 3.4.1 led to the supposition that frequency would be a useful filtering tool to highlight frequent terms and relation patterns between them, whereas salience might be able to highlight relations which indicate a very specific usage.

The analyses below were performed on two versions of the JEFF corpus: the original JEFF corpus (which treated each nomenclature reference word separately) and the unified JEFF corpus (which treated multi-word nomenclature terms as one). The graphs were created as described in Section 3.5. Each node represents a nomenclature reference, with the relations again being indicated as parent-child with the arrow pointing down the hierarchy (to the child).

Frequency

Both the original and the unified JEFF corpora analysis scenarios produced a pronounced long tail of relations. Analysis of the original JEFF corpus identifies 1613 relations, 932 of which appear only once in the corpus (see Table 4.3 for full breakdown with rising frequency threshold). Analysis of the unified JEFF corpus identifies 1384 relations, 1126 of which appear only once (see Table 4.4 for full breakdown).

Table 4.3: Original JEFF corpus breakdown of relations ID'd (frequency threshold)

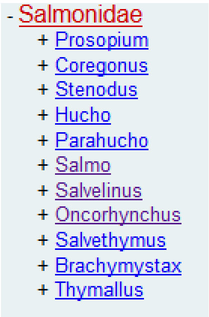
Frequency threshold	**No of rels ID'd**
No filter	1613
5 hits or over	283
10 hits or over	170
15 hits or over	115
20 hits or over	83
25 hits or over	63
30 hits or over	53
35 hits or over	47
40 hits or over	41
45 hits or over	36
50 hits or over	34

Table 4.4: Unified JEFF corpus breakdown of relations ID'd (frequency threshold)

Frequency threshold	No of rels ID'd
No filter	1384
5 hits or over	55
10 hits or over	30
15 hits or over	20
20 hits or over	10
25 hits or over	7
30 hits or over	6
35 hits or over	5
40 hits or over	5
45 hits or over	4
50 hits or over	4

This long-tail in lexicon is known as a Zipfian distribution. The drop in relations was even more marked in the unified corpus. Single data points cannot be generalised, so frequency was considered suitable to filter nodes in order to untangle the web to a certain degree. Sufficient relations were necessary to identify trends in the data for later profiling.

Increasing filter thresholds were applied. The higher filters significantly reduced the number of hits, which when creating graphs resulted in an easier to read graph. The analysis in this stage focused on edges (relations) and nodes (nomenclature terms) relating to salmon and trout mentions in the original JEFF corpus. This focus was chosen because of the high frequency of terms related to this family (Salmonidae), making them a suitable example to explore profiling possibilities.



- Salmonidae
  + Prosopium
  + Coregonus
  + Stenodus
  + Hucho
  + Parahucho
  + Salmo
  + Salvelinus
  + Oncorhynchus
  + Salvethymus
  + Brachymystax
  + Thymallus

Figure 4.2: Extract from VTO [151]

Figures 9.3, 9.4 and 9.5 can be compared see the difference in the readability of the graphs produced. The graphs show the progression of reduced relations and nodes as the frequency filter increases. This results in easier to read graphs as the relations and nodes become more visible. Figure 9.3 produces a graph that is difficult to analyse visually because of the sheer number of nodes and how intertwined all the nodes are with multiple relations. However, Figures 9.4 and 9.5 both produce graphs which the nodes and relations between them can be visually analysed, because there are clear parent to child relation links between specific nodes that fit with the nomenclature. Here, the relations between references are recognisable as real nomenclature terms relating to species or relations between linked common and scientific terminology, showing how these thresholds could be used to start to profile the usage of particular terms. In the original corpus, filtered for relations which have at least 20 hits, a number of interesting links were identified between different scientific nomenclature terms and the common names salmon and trout (see Figure 9.5). The profile here correctly demonstrates the links between Salmo, Oncorhynchus and Salvelinus genera, which are all part of the Salmonidae family, as Figure 4.2, an extract from the Vertebrate Taxonomy Ontology (VTO) shows [150]. This profile also shows us that trout is used for all three genera, and salmon in this case one genus. A search on various taxonomic resources confirmed that salmon and trout are both used for different species within the same genera. These distinctions were clearly profiled in the graph according to the evidence in the test corpus. The frequency filter is shown here to present a profile which highlights frequent nomenclature, also particularly emphasising in that case vernacular variants, which would suggests that frequency would be suitable for nomenclature that is widespread in a corpus and to profile scientific nomenclature against vernacular variants for disambiguation purposes.

The hits over 10 graph revealed further links between salmon and Oncorhynchus (see Figure 9.4), and an increase in the number of relations. Visually it was harder to discern the interesting links but it should also be recognised that a number of valid relations were identified that were missing from the over 20 hits graph, including all the Coregonus profile relations (in the top right of the Figure 9.5). This indicates that different filter thresholds would be useful depending on the focus of the analysis and the test corpus size.

As regards the lower filters, it would be possible to zoom into the graph in Cytoscape to identify identify correct relations. However, there was also a higher frequency of grouping relations identified, where terms would be identified as being related because they appear in a similar habitat, for example in Figure 9.6.

The analysis here suggested that filtering for frequency suggests that the higher the filter, the higher precision or accuracy of the results, but to the cost of other relations which are valid but less frequent, so finding a suitable filter threshold would be a question of balance and probably would depend on the purpose of the investigation. This is examined further in the methodology evaluation chapter (see Chapter 5). Salience was then explored to contrast the filter qualities of this measure.

Salience

The initial thought was that salience would emphasise less frequent, stronger relations in the corpus, in contrast with frequency which seems to emphasise widely-used terms. The preliminary findings, which are explored in this section, indicated that this was the case. In the case of salience filtering, high numbers of relations between terms were identified and the number decreased very slowly as the filter threshold increased. The number of nodes in the graph only reduced significantly when filtering for a very high salience threshold (no filter: 1613 relations, salience 10 filter: 974 relations, salience 13.5: 172 relations), by which point many other important links had been lost, as salience 13.5 is nearly as high as salience filtering can go, with the upper limit being 14. This made it difficult to perform an analysis across the whole corpus. To overcome this issue, salience was used to analyse subsections of the corpus. The subsection was defined by studying solely relations linked to the Salmonidae family. This was again because of the frequency of relations within this family identified, with the thought that this would reveal interesting profiles to analyse. Figure 9.7 focuses solely on the Salmo part of the original JEFF corpus and is filtered for salience of 9 (about mid-range for the salience of the relations identified). The number of edges and nodes are still high, but it is becoming easier to discern the interesting links: a mix of common and general terms (fish and species) as well as coherent links relating to Salmo, Oncorhynchus and how they are interconnected particularly through the vernacular variants salmon and trout.

Figure 9.8 shows the same focus on Salmo-linked mentions, filtered for salience 10. The picture here is clear: the more general and the vernacular variants (such as salmon, trout, species) serve as linking points through different parts of the graphs. The graph formation and shape will be discussed further at the end of this chapter.

These preliminary findings indicate that while frequency gives more weight to frequent, widespread relations in the corpus, salience tends towards maintaining highly specific relations, as you would expect from the description of the measure in Chapter 3 (see Section 3.4.1). As nomenclature itself represents very specific collocations, in these cases the different filters may be used to identify profiles of nomenclature usage that are a central theme in the corpus versus outliers as regards nomenclature used in a corpus. The results also indicated that salience could be used to focus on specific groups of nomenclature which would reduce the original number of nodes and edges in the graph. These aspects, as well as combining frequency and salience filters, are explored further in the Chapter 5.

4.2.2 Nomenclature profiling using relation network graphs

This research identified that nomenclature usage could be profiled using graph visualisations and that differences and characteristics of the nomenclature could be discerned in this way. Section

3.4.3 outlined research and applications of collocation networks to better visualise relations between different words in corpus linguistics. Here I applied this technique to visualise the profiles of nomenclature terms by the relations between them.

Cytoscape, the program used to create the network graphs in this thesis, has features which apply specific network analysis measures to the data in question. Exploration of the data through these graphs revealed a number of patterns in the behaviour as regards the nodes and relations between them. The measures that were studied in this thesis are specifically neighbourhood connectivity, closeness centrality and edge counts, because of the patterns identified in the data. These aspects of network analysis were used to identify features of nodes and edges within the graphs produced to see if they could be used to identify the taxonomic ranking of particular nodes, or other identifying features.

Firstly, for information, a brief definition of the measures being explored here:

- Neighbourhood connectivity measures the connectivity of a node in accordance with the number of neighbours it has. It is defined as the "average connectivity of all neighbours of n". [140]

- Closeness centrality is a measure that reflects the speed by which information spreads from one node to other reachable nodes in the network [140]. The calculation takes the average of the shortest path length from the node in question to all the other nodes in the network.

- Edge count is the number of edges that come into or go out of a node.

The study of the graphs revealed that certain nodes had contrasting and consistent characteristics as regards these measures. The following sections evaluate the capacity and reliability of this technique in the profiling and disambiguation of nomenclature terms (ranking, links from one to another). The following sections will go into detail about the specifics of the characteristics.

Hubs

Clusters of nodes were identified in the graphs. The term hubs was chosen to describe these groups of nodes all connected by one central node. These hubs occur both in isolation and connected to the main graph (often through the central node). It was also discovered that the nodes surrounding the central node and the central node have contrasting qualities which aid in their identification.

These nodes were identified as having specific characteristics:

Table 4.5: Filtering requirements for hub (central or surrounding) node definition

Node type	Edge count	NC	CC
Central node	3 or more	Bottom third of range	0.4-1
Surrounding nodes	Under 3	Top half of range	0-0.4

- Central nodes (node in the centre of a hub): These nodes have been found to have classification qualities, in that they occupy a position higher up the hierarchy than the surrounding nodes. They have also been found to sometimes have disambiguation qualities, when a word can either represent a species- or a genus-level term.

- Surrounding nodes (nodes surrounding a central node): These surrounding nodes tend to be a rank lower than the central hub node.

Figures 9.9 and 9.10 demonstrate the contrasting neighbourhood connectivity and closeness centrality of central (relatively high closeness centrality (CC)) and cluster (relatively high neighbourhood connectivity (NC)) nodes in the original corpus. Arrows indicate source to target (parent to child) relations. This tells us that a node that, in the graph, is higher up the hierarchy (as a hub node - a node which is the source of a number of relations) will have a high CC and relatively low NC, whereas a node lower down the hierarchy (a surrounding node - a node which is the target of one or few relations) will have a high NC and low CC.

One other type of node has also been identified, but its characteristics are less defined.

- Linking nodes: These nodes group different parts of the graph through common terms, often linking different hubs, as can be seen in Figure 9.11, where disambiguation of the links between trout, salmon and different genera of the Salmonidae family are shown.

Having identified these nodes manually, it was then necessary to see if there were more concrete ways of identifying these nodes and if these patterns could be seen in other parts of the graph, not just as regards the Salmonidae family which involved the relations with the highest frequencies in this corpus.

Cytoscape filtering options were used to filter the graphs to identify nodes that matched the central and surrounding hub node characteristics identified (see Figure 4.5). The central hub nodes also included a filter for the node to have at least 3 edges linked to it, as this was seen to be an important feature (to be classified as a hub it would have to be connected to at least 3 other nodes). These filters were defined through the manual identification of hub and surrounding nodes as described above, plus the calculation of the values for these measures in the cases identified.

Two different ways were identified to select hubs in Cytoscape: the first was to follow the filtering options in Figure 4.5 above: select the central hub nodes according to the central hub filter plus select the nodes according to the surrounding hub filter and looking at all these together. The second option in Cytoscape was to identify the central hub nodes using the requirements in Figure 4.5 and then look at what is connected to them through the Cytoscape feature to select related nodes. This can be explored in full using the Cytoscape files in Appendix C.

Having identified an automatic filter to select hub and surrounding nodes, I then developed a theory as to how to use this to identify specific semantic characteristics of a node, depending on whether it was a hub or a surrounding node. The definitions are set out in Table 4.6. The testing of this hypothesis is described in the subsequent sections of the chapter, which considers how data can be presented, or framed, in different ways, depending on whether a nomenclature term is considered on a per word basis (original) or unified as one (unified).

Table 4.6: Identification of node characteristics through hubs and graph positioning

Node identification	**Meaning**	**Node description**	**NC**	**CC**	**Edge count**
species (except parr)	Classification (as species); disambiguation (many species level names also genus level in other context)	Hub - outer node	High (top fifth or top half of range)	Low (0 - 0.04)	Under 3
genus; common; general	Identification of central nodes of hubs - classification level.	Hub - central node	Low (bottom third of range)	High (0.5-1)	3 or over
common; general (collective and life-stage); family; order	Linking node between different parts of the graph (classification)	Link node (not very well defined)	High (top fifth)	High (0.5-1)	3 or over

4.2.3 Original corpus: general profile characteristics

The previous sections in this chapter have outlined the preliminary findings are regards profiling of nomenclature through their visualisation in relation network graphs. The shapes of the graphs serve to identify hierarchies between terms and groupings of linked terms, as can be seen in Figure 9.12. Here the term “species” acts to group other terms together, linking out to a number of genus level terms, such as “Salmo” and “Salvelinus” and “Rutilus”, which in turn link down to species level terms such as “trutta”, “alpinus” and “rutilus”. This demonstrates how hierarchies are extracted from the general term species at the top, through the genus-level and species-level terms. Figure 9.12 also clearly demonstrates how species level terms tend to huddle around a genus central hub node, as described in the previous section. Finally, Figure 9.12 also shows how common names such as “salmon” and “trout” link different species together, through the species-level terms in this graph. These are the characteristics that can be used to profile the meaning of a nomenclature term in relation to the other terms in the corpus.

4.2.4 Data framing: original versus unified JEFF corpus

As described previously, data framing refers to the fact that data can be presented, or framed, in different ways depending on the purpose of an analysis. In the case of this thesis the data in question is the nomenclature. Nomenclature terms are often made up of multiple words, but each word is also a building block that refers to the taxonomic hierarchy. This means that the terms can be treated in various ways.

For this reason, I decided to study the profiles resulting from the extraction of nomenclature terms and the links between them in narrative text in two different ways, by treating the multi-word terms that make up scientific nomenclature as either single or multiple units. The previous sections only considered the JEFF corpus in its original form, treating each word within the nomenclature as a separate entity. In the original form a binomial nomenclature term comprises two separate units, the genus unit and the species unit. This is useful in its own right because each term has a meaning on its own. However, really it is the combination of the two which represents the taxonomic entity itself as a physical concept. For this reason, the corpus was also processed according to this logic to see how it affected the profiles extracted.

The following sections look at the profile characteristics of the data extracted from the unified corpus, and makes a comparison with the original corpus profile characteristics.

Unified corpus: general profile characteristics

Figure 9.13, extracted from the unified corpus, displays some different characteristics. We still identify a hierarchy but the characteristics of the graph and hierarchy changes. Hubs are mainly centred around common name variants (such as trout), or general-type (such as species, as before) and also general life-stage terms such as larva. By joining the multi-word terms to make them single units a layer of the hierarchy has been removed (in that you do not see a path from genus to species as this now represents one node, not two). However, this also means that the nomenclature term, which refers to a physical taxonomic entity, is considered as a unit. This changes the way each node relates to each other. In the original corpus, vernacular variants link through the graph to the species level nodes, which makes sense from the point of view that the common name refers to the species-level segment of the term, and can be used to highlight the way that a mix of species from different genera share common names in the nomenclature, and also highlights the link between common names and the species-level term itself, not the genus. However, the unified term corpus highlights the grouping role of common and general names and emphasises the physical reality of taxonomic entities, or species. These different characteristics are described in the next section.

Original and unified JEFF corpora: comparison of profile characteristics using hub selection criteria Sections 4.2.3 and 4.2.4 have provided general descriptions of the differences between the profiles extracted from the original and the unified JEFF corpora. To formalise the observations and also test the criteria developed in Section 4.2.2, analyses using these criteria on both the original and unified JEFF corpus were performed. The results are based on graphs which include relations with frequencies of 10 or more, or frequencies of 20 or more respectively for the original corpus, and frequencies of 5 or more for the unified corpus (the unified corpus had fewer numbers of higher frequency relations). They were filtered for the criteria set for identifying hubs and surrounding nodes, and the node contents were compared with the VTO to obtain information as to their scientific nomenclature ranking. In the case of the original corpus, the central hub nodes are nearly exclusively genus-level nodes, surrounded by species-level nodes. For example, in Figures 9.10 and 9.9, Coregonus is a genus-level word for a type of Salmonidae. The surrounding nodes are all species-level words of this genus. Further examples can be seen by referring back to Figure 9.12, with the nodes Salvelinus and Oncorhynchus. In some cases there are other nodes connected to some of these surrounding nodes. For example, Coregonus lavaretus is connected to whitefish, which is a common name for this species. Both Coregonus nasus and Chondostroma nasus are examples of species, but the former is a species within the family Salmonidae, the latter within the family Cyprinidae.

In the unified corpus, the central nodes are normally common names or general-type names, surrounded by genus_species names. Figures 9.14 and 9.15 provide examples (these are zoomed in perspectives from Figure 9.13). Species of various genera maybe called either salmon or trout, respectively. There is no clear distinction that matches a genus name to a specific common name. It is important to note that the profiles are correct but not complete in comparison with the VTO, as it is a reflection of only what is in the corpus. To fully explore the data for both corpora, please follow the link in Appendix C. It is not possible to reproduce all the graphs in their entirety on A4 paper.

The results of the filtering according to the definitions for central and surrounding hub nodes produced the outcomes in Table 4.7.

Table 4.7: Table demonstrating percentage compliance with defined hub characteristics

			% of nodes in graph which fit these criteria and which comply with meaning criteria given		
			original_corpus		**unified_corpus**
Node identification	**Meaning**	**Node description**	**Frequency 10**	**Frequency 20**	**Frequency 5**
species (except parr)	Classification (as species); disambiguation (many species level names also genus level in other context)	Hub - outer node	87.87% correct species to genus, 97.77% species, others correct links but skip from species to common etc. Other multiple.	94.7% species, 5.3% general. All correct links, with genera. Some have multiple links which link to common names too, also correct.	100% species (genus_species)
genus; common; general	Identification of central nodes of hubs - classification level.	Hub - central node	85% genus to species, 15% common name to species or genus	88.88% general, 11.11% common (this last has highest NC in group)	80% common names, 20% general collective name
common; general (collective and life-stage); family; order	Linking node between different parts of the graph (classification)	Link node (not very well defined)		Doesn't seem to have any - the trout in the previous one could count - but NC on limit of bottom third	

These analyses show that the higher frequencies are more stable and reliable. The links themselves tend to be correct but the semantic grouping (genus, species, general) of the node itself varies more in lower frequencies. The surrounding nodes seem to have more stable characteristics than the central nodes.

4.3 Preliminary findings from Phases 0 and 1

The first stages of the research set out to explore the data and identify the linguistic patterns and classes to be used in the subsequent stages of the research design. This stage identified the relations of interest and assigned them as parent-child relations (decided which was the parent, which the child). All were grouped in this way as every relation of interest had a hierarchical element. This grouping was chosen simplicity in developing the method and because of the qualities of the references being investigated.

Subsequent aspects of the research in these stages aimed to identify the effect of different filter parameters on the data output and the effects of framing the nomenclature according to separate word units (original) or unifying multi-word nomenclature terms as one (unified). The results of the filtering research indicated that frequency is a good general filter for the data at hand because of the Zipf's Law long-tail distribution of lexical items. Frequency removed outliers and quickly resulted in a network graph in which different relations can be identified. Salience maintained higher numbers of relations for longer which highlights its focus on strong, unique collocations. This made it difficult to use across the whole corpora. In studies using a subsection of the graph, however, salience was shown to identify relations accurately. The differences between salience and frequency are explored in a more systematic way in Chapter 5.

Hubs were identified as a characteristic of the graph representations that aided interpretation of the information in the graphs. Hubs consist of nodes which are surrounded by many other nodes, with the hub node having a grouping function, in that its meaning is more general than the nodes surrounding it. Patterns in neighbourhood connectivity and closeness centrality were found to be useful to automatically filter and identify such nodes.

Finally, an exploration of the impact of framing the data in different ways helped to highlight differences in the output of the graphs, depending on whether nomenclature terms are treated as single units (unified) or multiple ones (original). The hierarchies extracted from each demonstrated that the original corpus, which treats multiple word nomenclature terms as various separate units, highlights the links between genus-level and species-level terms, and links between species-level terms and common names. In the unified corpus, which treats each term and a single unit (as relating to a physical entity) highlights the links between common and general terms as hubs which collect species' names (binomial nomenclature terms) around them. These findings were applied in the subsequent phases of the research.

Chapter 5

Phase 2: Method validation and technical evaluation

This chapter presents the results of the method validation and technical evaluation developed for this thesis as described in Section 3.5.3. It can be considered to form part of the relevance cycle in that this is an evaluation of the method developed in the design cycle. While it is a technical evaluation, it places the evaluation in the domain as the process consisted of a comparison between the Vertebrate Taxonomy Ontology (VTO) and the nomenclature pairs extracted from the two test corpora, JEFF and WEB. The method developed for this thesis constitutes part of the research contribution, so a validation of the method was necessary to ensure it did actually extract the relations as intended. The method was also evaluated to measure the precision to which it extracted nomenclature pair relations, and also to evaluate qualitative differences between the different filter parameters in a systematic way, to build on the preliminary conclusions drawn in Chapter 4. The method was later applied to real-life scenarios, described in Chapter 6. This validation and evaluation chapter therefore sets out the work which responds to Objective 3, "to produce an evaluation method to compare the nomenclature pairs identified for precision, recall (quantitative measures) and differences (quantitative and qualitative measures) between the different expressions of knowledge (RQ2)".

5.1 Chapter overview

The validation and evaluation process was split into two stages of analysis: an overview analysis and a detailed analysis. The systematic evaluation of the filtering parameters spanned both analyses and will be included as such in this chapter. The overview analysis consisted of an automatic precision analysis, which provided simple quantitative results relating to the number of nomenclature pairs identified and the relative precision against the VTO for the respective corpora according to three different annotation and Word Sketch scenarios (as set out in Table 5.1).

Table 5.1: Corpus tagging and names

Scenario	**Analysis**	**Corpus**	**GNRD list**	**Word Sketch list**	**Corpus label**
1	Overview	JEFF	full JEFF	JEFF full	JEFF (JEFF, WS full)
1	Overview	WEB	full JEFF	JEFF full	WEB (JEFF, WS full)
2	Overview	JEFF	full JEFF	JEFF full	JEFF (JEFF, WS full)
2	Overview	WEB	full WEB	WEB full	WEB (WEB, full)
3	Overview	JEFF	full JEFF	JEFF subsection	JEFF (JEFF, WS subsection)
3	Overview	WEB	full JEFF	JEFF subsection	WEB (JEFF, WS subsection)
3	Detailed	JEFF	full JEFF	JEFF subsection	JEFF (JEFF, WS subsection)
3	Detailed	WEB	full JEFF	JEFF subsection	WEB (JEFF, WS subsection)

This overview analysis was performed on all three scenarios identified to draw conclusions as to the validity and reliability of the methodology and was used as a basis from which to choose the most suitable scenario to which to apply the detailed analysis. The overview analysis also provided the initial exploration into general trends in the qualitative differences between frequency and salience as filter parameters. The detailed analysis was performed manually and provided a detailed breakdown of the differences between the VTO and the corpora as annotated and Word Sketches extracted in Scenario 3. This detailed analysis meant that the automatic precision scores calculated in the overview analysis could be critically analysed, providing an adjusted precision score. This adjusted precision score was calculated taking into account any nomenclature pairs correctly identified by the method which were not identified by the authoritative resources used in the validation. This used the same calculation as the normal precision score, but was adjusted as to the number of relations considered as “real or correct” relations (true positives). The adjusted precision and detailed analysis not only served as further evidence as to the reliability of the method but also served to draw some general evaluative comments in relation to the authoritative resources used. Finally, the detailed analysis was used to provide systematic evidence as to the convergence or divergence of filter selectivity between the frequency and salience parameters, in order to draw more specific conclusions as to their value as filters.

The chapter is therefore set out as follows:

- Overview analysis (Scenarios 1, 2, 3)
  - Nomenclature pair relations identified across all scenarios

  - Effect of frequency and salience thresholds on the total number of nomenclature pairs identified across all scenarios
  - Precision score: percentage of nomenclature pairs extracted from each corpus which match VTO (per corpus and annotation schema)
  - Effect of frequency and salience thresholds on the precision calculations (per corpus and annotation schema)
- Detailed analysis (Scenario 3)
  - Detailed analysis of differences between JEFF and WEB corpora versus VTO (for the JEFF and WEB corpora tagged as in the Scenario 3 description in Table 3.5), frequency filter 5
  - Detailed analysis of differences between JEFF and WEB corpora versus VTO (for the JEFF and WEB corpora tagged as in the Scenario 3 description in Table 3.5), dual threshold
- Discussion section which outlines the general findings and conclusions of the validation and technical evaluation

5.2 Overview analysis

The overview analysis was used to provide a general validation of the method design itself, evaluate to what overall precision the method extracted nomenclature pairs and looked for stability in the results between corpora. As regards frequency and salience, general analyses relating to the patterns of behaviour of these filters and relations identified versus precision were considered.

5.2.1 JEFF and WEB corpora: nomenclature pair relations identified across all scenarios

Table 5.2 shows a breakdown of the total number of nomenclature pairs identified in each corpus. The tagging and Word Sketches called for the JEFF corpus are the same for both Scenario 1 and 2. The WEB corpus is tagged with the full JEFF name list and includes Word Sketches for the full JEFF name list in Scenario 1, while in Scenario 2 the tagging and Word Sketches are based on the full WEB name list. Scenario 3 again considers both the WEB and JEFF corpora tagged with the full JEFF name list. The Word Sketches analysed are only based on a subsection of the JEFF name list (JEFF (JEFF, WS subsection) and WEB (JEFF, WS subsection)). The Word Sketches identified through this method equate to approximately 75%

of the nomenclature pairs identified by each corpus in Scenario 1 (which involved the analysis of the Word Sketches from the full JEFF name list). The WEB corpus in Scenario 2, tagged on the basis of the WEB name list, identifies more than double the number of nomenclature pairs in comparison with those identified when it was tagged with the JEFF name list (WEB (JEFF, WS full or subsection)).

Table 5.2: Breakdown of nomenclature pairs identified in each corpus, depending on the annotation schema followed

Scenario no	Analysis	Corpus as annotated	Total relations found
Scenario 1	Overview	JEFF (JEFF, WS full)	1715
Scenario 1	Overview	WEB (JEFF, WS full)	1581
Scenario 2	Overview	JEFF (JEFF, WS full)	1715
Scenario 2	Overview	WEB (WEB, WS full)	4014
Scenario 3	Detailed	JEFF (JEFF, WS subsection)	1218
Scenario 3	Detailed	WEB (JEFF, WS subsection)	1351

Effect of frequency and salience on nomenclature pair relations across all scenarios

The potentially divergent focuses of salience and frequency was first explored in Chapter 4. This chapter systematises the analysis of these filter parameters to identify differences in behaviour relating to the number of nomenclature pairs identified and the precision against an accepted ontological resource when applying these filters. The validation and evaluation described here use graphs to track the pattern of change across both JEFF and WEB corpora annotated according to different criteria, which are described in the previous section and in more detailed in Section 3.5.3 of Chapter 3. Figure 9.16 shows how although the WEB corpus (WEB, WS full) (Scenario 2), identified over double the number of nomenclature pairs in this scenario than in the WEB (JEFF, WS full/subsection) scenarios (Scenarios 1 and 3), the majority of these hits were very infrequent, bringing the total relations identified down to near the other scenarios with very low filter thresholds.

Figure 9.17 clearly shows the long-tail pattern of nomenclature pairs identified in terms of frequency, whereas Figure 9.18 shows how the number of nomenclature pairs identified when filtering for salience remains high for a long period, forming what could be described as an inverse Zipf curve. The graphs also show how this pattern applies to both corpora and across all scenarios, demonstrating that these trends are not a one-off phenomenon. Please see Appendix D, Tables D.1 and D.2 for a breakdown of the numbers.

5.2.2 Precision score: percentage of nomenclature pairs extracted from each corpus which match VTO (per corpus and annotation schema)

The previous section discussed the relative impact of frequency and salience thresholds on the number of nomenclature pairs identified in the corpora. This section analyses the impact that frequency and salience had on the percentage precision of the results when comparing the nomenclature pairs identified in the corpora with the VTO, from an overview perspective.

Figures 9.19 and 9.20 show the precision trends across all corpora and all annotation/Word Sketch scenarios. As the thresholds in each case rise, so does precision. The JEFF corpus obtained a higher precision score against the VTO than the WEB corpus in all annotation/Word Sketch scenarios. Without a full analysis the reasons for the lower precision in WEB corpus scenarios cannot be ascertained (it could be due to methodological issues, out-of-scope references, the corpus content among other reasons), but this would constitute further work. The same inverse Zipf-curve pattern is identified in all scenarios, which supports the decision to look at the reduced section of data as a valid sample.

Precision scores plus number of nomenclature pairs identified

While precision is a useful measure, it is not sufficient to consider precision alone, overall frequency of nomenclature pairs identified must be considered together with this score. As was mentioned in the previous section, the long-tail distribution of nomenclature pairs means that the frequency of the relations identified drops very quickly, which could have meant that a high precision score was simply linked to very low frequencies of relations. The following graphs compare frequency of relations identified and the precision for each corpus, for both types of filtering, across different scenarios. The Scenario 2 frequency graph, Figure 9.21 shows the nomenclature pairs versus precision for five or more hits and above as it produces an easier to read graph. The large drop in relations identified between no filter and the 5-hit filter in the WEB (WEB, WS full) corpus (87% of nomenclature pairs) meant that otherwise it was difficult to read the comparisons between between the WEB (WEB, WS full) and JEFF (JEFF, WS full), corpora. This graph, based on the Scenario 2 analysis, confirms what we have seen in the previous slides about the similar curves both as regards nomenclature pairs and precision. It confirms that the WEB corpus appears to have lower precision and higher number of nomenclature pairs, although most of these are infrequent. Figure 9.22 shows the salience filter results for Scenario 2 across both the WEB and JEFF corpora. Here there is a greater difference in the precision, but also a much greater disparity between the number of relations identified. The same trend is followed in both cases, however, with rising precision as salience rises.

Figures 9.23 and 9.24 and 9.25 and 9.26 show the comparative trends of nomenclature pair relations identified for the JEFF and WEB corpora when subjected to both Scenarios 1 and 3, and then filtered for frequency or salience thresholds. These graphs show that both corpora follow similar trajectories in the numbers of relations identified in the different scenarios. The decision to perform the detailed analysis on Scenario 3, using a subsection of the JEFF name list Word Sketches extracted from each corpus, was taken because the comparison here demonstrated the similar trends followed by each corpus results between Scenario 1 and 3. Scenario 3, which used the subsection of the Word Sketch list, placed further emphasis on more frequent nomenclature pair relations and was therefore thought to be suitable for such an analysis. The detailed analysis specifically looks at the relations identified in the Scenario 3-tagged corpora when filtered for 5 or more hits. Figures 9.23 and 9.24 also show how the filter 5 provided a balance between a relatively high precision score and also sufficient relations to identify some patterns on which it was feasible to perform a manual analysis.

The previous graphs have shown that the precision score was generally lower with the salience filter than the frequency filter, as can be seen by comparing Figures 9.27 and 9.29, with Figures 9.28 and 9.30. Precision only rose significantly at the top salience thresholds. When considering the salience filter, the number of nomenclature pairs identified remained fairly stable and high throughout most of the trajectory. This characteristic can be linked to the purpose of a salience score in emphasising strong relations (the salience calculation works on the basis of highlighting the number of times two words co-occur in comparison with the times each word occurs separately with another word). These peculiarities do suggest that the types of relations identified by the respective filters might diverge. While the salience score avoids the over-emphasis given to infrequent terms that calculations such as the Mutual Information score assign, it does still tend towards more infrequent, strong relations. In the case of scientific nomenclature it is likely that many of these terms only appear once in the corpus. The limitation of the scope of the Vertebrate Taxonomy OntologyVTO meant that I could not be sure whether the precision score without any filtering was accurate - the lower precision may only indicate that many of the hits were infrequent invertebrate nomenclature pairs. For this reason, the lower salience for the WEB corpus in Scenario 2 (see Figure 9.22 could not be attributed with certainty to a less accurate or consistent usage of the nomenclature. It may just indicate a higher variety or increased focus on invertebrate species. Future work would aim to develop a method for analysing these differences or try to identify an ontology that would fit all the needs to facilitate an automatic evaluation of this aspect of the research.

This section has detailed the overall analysis of the relations between the JEFF and WEB corpora in different annotation scenarios, to demonstrate the validity of the methodology through the stability of patterns across the corpora and scenarios. It has gone further to identify that precision increases as each salience and frequency filter thresholds rise and also used the patterns of relations identified to select Scenario 3 for the final detailed analysis on the basis of the heightened emphasis on more frequent relations and the validity provided by all different scenarios following the same pattern. Finally, the inverse trends observed when looking at either frequency or salience thresholds indicates that the two different filters may highlight distinct data because of the focus on different properties within the data. This is all explored further in the subsequent detailed analysis, which has been used to provide a full technical evaluation of the method and of the characteristics of the frequency filter versus the salience filter.

5.3 Detailed analysis: breakdown of comparison between JEFF and WEB corpora (Scenario 3)

This section deals with the breakdown of differences between each Scenario 3 corpus and the VTO, and then performs a comparative analysis of these differences between the two corpora. This section constitutes the detailed analysis. This relates to Scenario 3, which means that all analyses from this point refer to the JEFF (JEFF, WS subsection) and WEB (JEFF, WS subsection) corpora. The breakdown of differences permitted an adjustment to the precision scores where necessary to account for any false negatives in the overview analysis - true relations identified in the corpora which had no match in the VTO for whatever reason. The next sections look at these breakdowns.

5.3.1 Detailed analysis frequency and salience filter decisions

The breakdown analysis can be broken down into two parts. The first part of the evaluation provides a breakdown of the differences between the JEFF and WEB corpora and the VTO as identified in the nomenclature pair relations extracted with a 5-hit threshold. A 5-hit threshold was chosen because it excludes the high number of nomenclature pairs which have only one hit while maintaining decent numbers of relations.

The second part of the analysis aimed to explore the hypothesis that salience would result in a divergence of the relations identified in comparison with the relations identified using the frequency filter. To make it possible to do this, it was necessary to first exclude some of the most infrequent hits, and then apply the salience filter. This meant that it was possible to compare the convergence or divergence of results between the dual threshold filter (when frequency and salience were used together) with the single frequency threshold filter described in the first part of this section.

Differences criteria

Table 5.3 sets out the differences criteria and their definitions as classified for the final breakdown analysis. These criteria were set up to distinguish between true and false incorrect matches. This was necessary because the automatic overview analysis only took into consideration matches that were included as entries in the VTO. The VTO is not a gold-standard resource as there is no gold-standard resource for the nomenclature of species. Also VTO only includes vertebrate species so any non-vertebrate species identified by the method would not be matched as a true relation in the overview analysis, thereby constituting a false negative. These criteria were used to identify which were true incorrect matches (relations identified by the method which do not exist in real life) and false incorrect matches (true relations identified by the method but which were counted as false by the overview analysis because they did not exist in the VTO). Synonyms and scope-related differences are considered false incorrect matches (true relations) because they represent real examples of scientific nomenclature, but that do not form part of the VTO ontology being used in the evaluation. Any relations categorised as one of these criteria were identified in the breakdown analysis and included with the correct matches to provide an adjusted precision score for each corpus. True incorrect matches (false relations) are those that do not in fact represent a real nomenclature term or relation in the hierarchy. The criteria of skipped ranking and misspelling, while not being incorrect matches as such, have not been included in the adapted precision scores because of any potential ambiguity as to the interpretation of these terms.

Originally the breakdown analysis criteria included a criterion of "Synonym not in the VTO". This was originally classified as "synonym not identified in the VTO but identified in CoL or ITIS as a synonym for a vertebrate included in the VTO". However, there were contradictions between the definitions in the CoL and ITIS for a number of the examples identified. All the examples consisted of alternative spelling variants and were therefore reclassified as misspellings. A breakdown of the lists of names and classifications can be found in the appendices (see Appendix D).

Table 5.3: Differences criteria and definitions

Criteria	**Definition**
Synonym in the VTO	Synonym as identified in the VTO (in evaluation only main accepted name considered)
Correct taxonomic reference not in the VTO	Correct nomenclature but not found in VTO (found in CoL or ITIS) (in subsequent table this grouped in scope (other)
Plant	Correct nomenclature but outside scope of VTO for being plant (found in CoL or ITIS)
Invertebrate	Correct nomenclature but outside scope of VTO for being invertebrate (found in CoL or ITIS)
Misspelling	identifiable nomenclature pairs but with misspelling of the name - both through OCR issues and misspelling of other kinds.
Partial name	Where two parts of a multi-part scientific nomenclature are identified. This could be two consecutive parts, such as a species level plus subspecies or author depiction, or could jump, such as genus to the author depiction.
Unknown	Unidentified name or one that cannot be evaluated without an expert
Incorrect matching - parent/child reversal	The matched pair has identified the parent-child relation in the inverse (for example: for the term Salmo salar, the genus would be identified as the child, the species level as the parent) (all incorrect matching grouped in subsequent tables - breakdown can be found in later sections)
Incorrect matching - group same rank	The two parts of the name are from the same lineage but same rank (for example: Salmo and Salvelinus of the Salmonidae family)
Incorrect matching - group various rank	The two parts of the name are from different lineage and different ranks (in case of ambiguity expert involvement needed for evaluation)
Ranks skipped	The two parts of the name are from the same lineage in the nomenclature but they skip a rank, for example: family to species
Incorrect matching - not scientific name	One or both parts of the name is not part of the scientific nomenclature
Match with common name	Common name
Same	Exact match with accepted name in the VTO (main accepted name)

Table 5.4: JEFF corpus precision and differences breakdown

Filter applied	Precision	Differences								
		Total	Misspellings	Accepted tax. ref. not in ontology		Synonyms	Unknown	Incorrect	Partial name	Other (skipped ranking)
				Scope	Other					
No filter	37.77%									
Frequency 5	81.5%	42	5	10	1	14		4	7	1
Frequency 10	87.02%	17		3	1	9			4	
Frequency 15	89.89%	9		1		6			2	
Frequency 20	94.03%	4		1		3				
Frequency 25	92.45%	4		1		3				
Frequency 30	93.33%	3				3				
Frequency 35	92.68%	3				3				
Frequency 40	91.67%	3				3				
Frequency 45	93.75%	2				2				
Frequency 50	96.67%	1				1				
Frequency 55	100%	0								

Table 5.5: WEB corpus: precision and differences breakdown

Filter applied	Precision	Differences Total	Misspellings	Accepted tax. ref. not in ontology		Synonyms	Unknown	Incorrect	Partial name	Other (skipped ranking)
				Scope	Other					
No filter	30.27%	942								
Frequency 5	70.07%	85	5	30	1	18	1	18	12	
Frequency 10	78.61%	37	3	14	1	10		8	1	
Frequency 15	79.55%	27	2	6	1	10		7	1	
Frequency 20	84.91%	16		3		8		5		
Frequency 25	83.70%	15		3		7		5		
Frequency 30	84.81%	12		2		5		5		
Frequency 35	85.71%	10		1		4		5		
Frequency 40	85.25%	9		1		3		5		
Frequency 45	94.44%	3		1		2				
Frequency 50	95.92%	2		1		1				
Frequency 55	95.35%	2		1		1				
Frequency 100	95.24%	1				1				

Differences breakdown of nomenclature pair relations identified in JEFF and WEB corpora: frequency filter

Tables 5.4 and 5.5 show the breakdown of the comparative overview precision and differences according to the set criteria across both JEFF and WEB corpora. The graphs show that in the case of the JEFF corpus, the highest frequency difference is consistently that of synonyms (a definition of synonyms and how it is being used here can be found in the Ontology Validation chapter). The WEB corpus starts with much higher out-of-scope (species that would not be expected to be found in the VTO) relations, probably due to the nature of the corpus, it being an automatically scraped corpus from the internet according to key terms defined from the JEFF corpus. For a visual representation of the differences and how they evolve as the frequency filter rises, please see Figures 9.31 and 9.32.

As the frequency threshold filter rises, in both cases it is the synonym-difference relations that persist longest. In the original overview precision analyses for each corpus, the JEFF corpus reaches 100% precision at frequency filter of 55 hits, whereas the WEB corpus even has one synonym at 100 hits. If the precision is adjusted to account for the various differences in which the method has correctly identified nomenclature relations, but that were not picked up as such in the overview analysis of this chapter (synonyms, out-of-scope relations, valid nomenclature not included in the VTO), the precision scores improve for both corpora improve. Figures 9.33 and 9.34 show that when you exclude both the out-of-scope (e.g. invertebrates) nomenclature pairs and those of synonyms (nomenclature variants of specific species) from the differences between the corpus and the VTO (thereby adjusting the precision accordingly), the JEFF corpus reaches 100% precision as early at the 20 hit threshold, and the WEB corpus at the 45 hit threshold. The incorrect matches can be seen to reduce as filtering rises, which supports the validity of the method. It demonstrates that the method correctly identifies the data it is looking for when frequency rises, and the consistency of the trajectories taken by each corpus demonstrates cross-corpora stability of the results.

The differences in the results between the two corpora does indicate that the JEFF corpus is more consistent in its presentation of the nomenclature than the WEB corpus. Also, to ensure these precision markers are not simply an indication that there are no data points left, we can refer back to the graphs provided earlier. At the 20 hit filter in the JEFF corpus there are a total of 67 relations identified, and at the 45 hit filter in the WEB corpus this related to 91 relations. The indications of the WEB corpus being less consistent may indeed just be a reflection of its size. The breakdown of differences also demonstrates the number of relations identified which are actually correct hits, to see an immediate improvement in the adjusted precision scores.

5.3.2 Analysis of differences between JEFF and WEB corpora (Scenario 3): frequency only filter

This section performs an empirical analysis of the relations identified, separated by difference criteria, to analyse the reason for these relations being identified, any particular qualities in the corpora of these word pairs in context and also consider any weaknesses in the method developed that could be improved in future work. As regards the latter, most incorrect matches relate to incorrect groupings of some sort. To resolve or mitigate this problem, in further work, a more fine-grained system of relation identification could be implemented, instead of grouping all as parent-child relations.

Inverted parent-child relations Inverted-parent child relations represented one of the incorrect matches identified in the analysis, and refers to incorrectly identified relations in terms of the incorrect identification of the direction of a relation.

In the JEFF corpus filtered for 5 hits or more, there are no relations which have been identified which were inversely identified according to their parent/child relation. In the WEB corpus two examples were identified, as seen in table 5.6. This served as a demonstration that the method is correctly identifying the empirical evidence in the literature as regards nomenclature reference, because the numbers of inverted parent-child relations are very small. For the JEFF corpus the inversely matched parent-child relations when looking at 5 or more hits was 0% whereas for the WEB corpus it represented 0.03% of the different relations identified. This dropped to 0% for both corpora when considering relations identified with 45 or more hits in the WEB corpus.

Table 5.6: Inversely matched parent-child (source-target) pair in WEB corpus, filter 5

Source	Target	Difference description
ESOX	CYPRINIDAE	Wrong way round, also different families
THYMALLUS	SALMONIDAE	Wrong way round (Thymallus is a genus of the family Salmonidae)

As regards the Esox, Cyprinidae pairing, when looking at the Word Sketches themselves to identify the contexts, it was found that all the instances (42) were actually from the same text, but on different URLs as they had been captured through the Linguee translation website. This demonstrates a useful quality of the lexicographic approach, which uses the mix of quantitative data to look for patterns in language usage, with the possibility of going into the source documents to check the context for any interesting or surprising results and either keep or discard according to the findings. The fact that there were so many instances of this same sentence does highlight one of the drawbacks with automatically scraped corpora, but it is something to be aware of, and as mentioned before then with the capability of going into the data to evaluate the validity of results.

While a weakness of the method, very few relations were wrongly identified in this way, supporting the reliability of the method from this perspective.

Incorrect matching - group same rank Incorrect matching - group same rank refers to cases in which groups are incorrectly matched as hierarchical relations. This is because of the lack of granularity in the method which only permits parent-child pairs. Tables 5.7 and 5.8 show that, when filtered for five or more hits, the corpora have three and eight instances of this difference, respectively. This particular issue disappears in the data from frequency 10 in the JEFF corpus, with the WEB corpus having one match at frequency 10 filter, a link between Ephemeroptera and Plecoptera (mayflies and stoneflies), with 19 hits. The application of further semantic constraints to develop a more fine-grained approach to the tagging and to identify potential sibling-sibling relationships may be used mitigate this issue in future work. This has been identified in the literature as a way of overcoming ambiguity problems in linguistic pattern-based methods [125]. In the case of matching of groups of the same rank, every single example is the result of incorrectly identifying the word pairs as nouns with modifiers when they are parts of lists. In any case, while there could be some way of making an assumption about terms in a list, this would never be 100% infallible, nor if there any discernible way of separating out modifiers which present hierarchical relations from those which present sibling relations through this method.

Table 5.7: JEFF corpus filter 5 - incorrect matching (group same rank)

Source	Target	Difference description
SALMO	ONCORHYNCHUS	two genera - incorrect joining
EPHEMEROPTERA	PLECOPTERA	two order level terms - incorrect matching
LUCIUS	FLUVIATILIS	two species level terms - incorrect matching

Incorrect matching - group various rank Group various rank described where terms from different ranks in the taxonomic hierarchy are identified in pairs, but instead of identifying two terms from the same family, the terms originate from different families. In this case there is one example of this type of error in the JEFF corpus, and four in the case of the WEB corpus when filtering for five or more hits (see Tables 5.9 and 5.10).

In the JEFF corpus, as before, the wrongly identified relation was a result of wrongly identifying a list as a noun and modifier. This is the most prevalent problem with the method

Table 5.8: WEB corpus filter 5 - incorrect matching (group same rank)

Source	Target	Difference description
PLECOPTERA	TRICHOPTERA	Group of different insect orders
SALAR	TRUTTA	Grouping various species level terms for same family together
SALMO	SALVELINUS	Grouping of like ranks within same family
EPHEMEROPTERA	PLECOPTERA	Mayflies and stoneflies (grouped of same rank)
TRICHOPTERA	EPHEMEROPTERA	Groups of different orders of flies
BOSMINA	CHYDORUS	Both genera
CERNUUS	CERNUA	Grouping of species level terms (looks like name variants for same species)
MOLLUSCA	NEMATODA	Nematoda (phylum) and mollusca (phylum) - also group of same-level ranking

Table 5.9: JEFF corpus (filter 5) - group various ranks

Source	Target	Difference description
CLADOCERA	COPEPODA	two different groups of organisms - copopoda (subclass) versus cladocera (suborder)

identified and work to adapt the linguistic patterns in future work could be used to mitigate this issue.

Looking into the actual data in the WEB corpus, the first four relations actually arise from the same sentence. This sentence is repeated 43 times in the corpus. The sentence reads: “Cyprinid waters shall mean waters which support or become capable of supporting fish belonging to the cyprinids (Cyprinidae), or other species such as pike (Esox lucius), perch (Perca fluviatilis) and eel (Anguilla anguilla).” The sentence has been scraped from translation websites, which goes to explain how it can appear so many times.

In the case of Cyprinidae plus fluvialitis, at first glance a possible interpretation could have been that fluvialitis was a species-level term for something within the Cyprinidae family because it is a genus species pair. However, looking at the data, it is clearly linked to the species Perca fluvialitis and the sentence is not claiming that Perca fluvialitis is a cyprinid at all.

The last two examples are, as in the previous section, a list of names. While the appearance of Cladocera and Copepoda seems to be more widespread throughout the corpus, Gastropoda and Nematoda instances all arise from one document which is a repeated list of names. This example, along with the translation website example highlights some of the issues with web-scraping corpora, and how this can skew quantitative results. It would also suggest that to

Table 5.10: WEB corpus (filter 5) - grouping various ranks

Source	Target	Difference description
CYPRINIDAE	ESOX	Esox (genus) of the Esocidae family and cyprinidae (family) - latter are carps or minnows
CYPRINIDAE	PERCA	Family to genus (Cyprinidae) to genus (Perca) but should be Percidae
CYPRINIDAE	FLUVIATILIS	Skipping rank across different lineages
CYPRINIDAE	LUCIUS	Linking Cyprinidae to lucius from Esox lucius (pike) (same as Cyprinidae Esox links)
GASTROPODA	NEMATODA	Gastropoda class of mollusca phyllum
CLADOCERA	COPEPODA	two different groups of organisms - copopoda (subclass) versus cladocera (suborder)

improve the method of compiling a corpus, that translation websites should be blacklisted. Other websites to avoid could also be blacklisted. The flexibility of the lexicographic approach here, which allows the researcher to look at the specific evidence should be recognised: you see a pattern, you look at the various contexts in which it appears and the origin of these contexts to be able to make a judgement call as to the validity of the assertion which arises from the pattern. This is one of the great strengths of the lexicographic approach and why it was chosen for the research.

Incorrect matching - not scientific name Matches categorised as “not scientific name” were classified as incorrect matches because there is at least one in the pair of words identified in each relation which is not a scientific name. However, looking at the actual examples, the instances in which one of the pair was a number. When this was investigated looking at the actual data, in the concordances it was not clear what number the Word Sketches were referring to - it would seem that there might be some anomaly in the Word Sketches and how it links to the concordances.

Again, this type of incorrect matching only appears in the WEB corpus when filtering for five hits or more. The increased frequency of the incorrect matches in the WEB corpus in contrast with the JEFF corpus could be related to its size, or the greater variability of the data. To be able to discern the specific reason behind this it would be necessary to create corpora with more specific boundaries as regards size and representativeness.

Misspellings One of the difference criteria in the method evaluation is that of misspellings. I chose to include this criterion because there are potentials for misspellings both as a result of OCR issues (a methodological issue) and author typographical errors (something recognised as a common problem in the area of biodiversity [172, 207, 212]). Chapter 6 looks in more detail at this issue, because not all knowledge representation resources categorise “misspellings” and “synonyms” in the same way.

Table 5.11: Misspellings in JEFF and WEB corpus, plus frequency and distribution

Corpus	Source	Target	Difference description	No. of docs	Frequency of hits
JEFF	MICROPTERUS	DOLOMIEUI	Micropterus dolomieu	4	8
JEFF	ONCHORHYNCHUS	MYKISS	Oncorhynchus mykiss	5	5
JEFF	PROTOTROCTES	MAREANA	Australian grayling - misspelling (Prototroctes maraena)	1	12
JEFF	SULMO	GUIRDNERI	misspelling or OCR error - Salmo gairdneri	2	9
JEFF	SULMO	TRUTTA	misspelling or OCR error - Salmo trutta	8	15
WEB	ONCHORHYNCHUS	MYKISS	Oncorhynchus mykiss	13	20
WEB	MICROPTERUS	DOLOMIEUI	Micropterus dolomieui Lacepède, 1802 (synonym) of Micropterus dolomieu according to CoL.	22	26
WEB	SALVELINUS	WILLUGHBII	Windermere charr, is a cold-water fish in the family Salmonidae. In the VTO only Salvelinus willoughbii, but CoL says this form is a synonym for the latter.	3	7
WEB	MYOXOCEPHALUS	THOMPSONI	Deepwater sculpin. Misspelling seen in articles but not databases.	8	11
WEB	ONCORHYNCUS	MYKISS	Oncorhynchus mykiss misspelling. Seen in articles but not databases.	6	6

The distinction between misspellings and synonyms is actually not 100% clear. For the purposes of the precision, as this chapter is looking at a comparison between the VTO and the corpus, any entities included as related synonyms within the VTO ontology are considered and classified as synonyms, whereas anything not classified as a synonym in the VTO, but that can be considered to represent an alternative spelling of a recognisable term in the scientific nomenclature, is classified as a misspelling. A more in-depth discussion about synonyms and misspellings is provided in Chapter 6 (nomenclature profiling studies). However, the recognition of these misspellings, whether they be a result of authorship or OCR issues, can also be considered a strength of the method and the use of the GNRD which picks up on these variations, in order to gather a broader picture of all the nomenclature pairs included in a given corpus, and also to see if there are specific patterns relating to said variants.

Table 5.11 shows that, of the five examples of misspellings in the JEFF corpus, two were the result of OCR issues. There were no examples of OCR issues in the WEB corpus as it had been scraped from the internet. This shows that while OCR could be an issue it does not seem to have been in the case of this corpus (as they represent a very small number of the errors). This would change should you be working with older documents that have been scanned, for example. However, there are other ways to deal with OCR issues that are not essential to the method design. This would be able to be dealt with on a case by case basis.

The other misspellings are common misspellings found within the literature. The identification of misspellings is important because they still represent references to nomenclature in the corpus, whether they are correct or not. In fact it could be taken into account in the precision because it is not an incorrect identification of a name, but the correct identification of an incorrectly spelled name which can be useful in profiling nomenclature usage (if there are patterns across authors or time, for example). The fact that GNRD picks up common misspellings can be seen as an advantage of the method against, for example, a dictionary-based approach which would exclude anything not included in the dictionary, although it may be just as valid to use a dictionary to identify terms of interest. It depends on the purpose of the analysis. It is valuable that the method can identify these for analysis as they represent real usage in the literature. The method developed here can investigate patterns relating to how different spelling variants are used, which could be used to identify context-, author-, time-specific trends among others.

While OCR is often considered something that greatly impairs work, in this case it does not seem to have greatly impacted the effectiveness of the method. However, should there be an OCR'd corpus which contains older material, which is likely to cause more OCR problems then to ensure that the method maintains efficacy it would be important to either manually correct OCR issues or opt for a more thorough OCR process through which these issues can be picked up and corrected to ensure reliability.

Partial names This incorrect matching highlights another weakness in the method. Scientific nomenclature is usually binomial, but because of authorship and ranks such as subspecies, sometimes the names span over more than two words. This has resulted in partial matches of names in the results of the investigation. When analysing the breakdown of differences in comparison to the VTO, it was found that for the JEFF corpus, 100% of these partial names included Linnaeus in the pair (as child) (for hits over 5), in comparison with 92% for the extended corpus, with the remaining example being that of a pair with the abbreviation spp., which stands for "several species". From this we can see that it is a common occurrence for this sort of anomaly. This is not technically incorrect, but it highlights issues with my method which only picks up two-word pairs as a nomenclature reference relation, on the basis of Sketch Engine's Word Sketches which work by identifying word pairs in specific relations with each other. One of the ways around this would be to join certain scientific nomenclature or to adapt the Sketch Grammars in a specific way.

This aspect of the method impacts on the accuracy of pulling out scientific nomenclature that refer to species with the authorship included (which would be four part when split by word), or subspecies which can be even longer. Most of the nomenclature in the test corpora are binomial. However, this is not always the case and if looking at texts, such as taxonomic texts, in which the full authorship is expected to be given or a domain in which subspecies are very common, this would be an important factor to bear in mind.

Synonyms The synonyms aspect of the differences will be described in more detail in Chapter 6, as the definition of synonym varies from knowledge resource to knowledge resource. In the case of the method validation and evaluation all these different perceptions have been grouped because to test whether the method is doing what it says it is doing this is sufficient. However, to validate and discuss ontological knowledge representations against the empirical data this must be discussed and considered.

In the breakdown of differences between the two test corpora and the VTO, a significant proportion of these can be accounted for by what have been classified as synonyms. Please refer to Appendix D, Table D.3 for a full breakdown of the synonyms identified. The following was identified about these synonyms:

- Twenty-three unique synonyms were identified

- Ten of which appeared in both corpora (43%)

- Of the remaining thirteen, six were identified in the JEFF corpus and eight in the WEB corpus

This indicates that synonyms are used across the board and that there is some stability in which synonyms are being used.

Here it is sufficient to note that synonyms, in this case recognised as nomenclature pair relations correctly identified by the method, are the differences that persist the longest at the frequency threshold is raised, which supports the argument that the method is reliable.

The way that synonyms are identified in the method is important as it allows for analysis as regards the usage of these synonyms in contrast with accepted names. It could be used to identify patterns as to where in a document specific terms are used, and whether there are time, domain or authorial factors that govern their use (among others). Further metadata would be needed to gather information relating to some of these variables, but this issue is explored further in Chapter 6 .

Rank skipping Rank skipping refers to where the relations identified are within the same lineage but they have skipped ranks (for example from family to species level rank). Therefore the match is not necessarily incorrect, but it does not appear as the direct hierarchy expected from the VTO.

Only the JEFF corpus identified a relation at frequency five filter, which was Cyprinidae and Teleostei. This fitted the correct identification through lineage but skipping one or various rank levels. A number of the partial names also fitted the description but were not classified here because they were part of the genus, species, authorship trio).

5.3.3 Analysis of differences between JEFF and WEB corpora (Scenario 3): dual threshold filter

As was mentioned in the introduction to this chapter, the salience filter maintains most of the relations identified until very high levels of salience, because of the Zipf's law curve of instances (in this case relations) with a large proportion of the relations identified being identified once, with accompanying high salience scores because of the technical nature of the terminology being analysed. This resulted in a graph that, unfiltered or filtered with a low threshold, was too interconnected and failed to identify and useful patterns. Conversely when filtered for high salience, because of this characteristic, many of the frequent terms were excluded because the salience threshold was so high. Therefore, in order to evaluate this aspect of the method it was decided to perform a comparison between the frequency filter of 5 evaluated above and a mixed filter of frequency 4 plus salience 9 to 11 to compare the output. These particular filters were chosen for a number of reasons. The main evaluation was based on a frequency filter of 5 because of the perceived balance between filtering excessive relations and maintaining interesting ones at this filter threshold. Therefore, to compare the salience filter with the frequency filter outputs it made sense to compare a variety of salience filters slightly below that of frequency filter 5 to look for convergence or divergence in the relations identified. A range of filters were studied to be able to identify any possible patterns of convergence or divergence. See Table 5.12 for the comparative numbers of relations plus precision for filters between frequency 4 and 5, with varying salience for the JEFF corpus.

Table 5.12: JEFF corpus: table comparing precision and relations identified with combinations of frequency and salience filtering

Filter	**Total found**	**Not in VTO**	**In VTO**	**% only in JEFF**	**% precision**
JEFF freq 4	274	60	214	22%	78%
JEFF freq 4, sal 9	268	56	212	21%	79%
JEFF freq 4, sal 10	246	46	200	19%	81%
JEFF freq 4, sal 11	224	39	185	17%	83%
JEFF freq 4, sal 12	196	34	162	17%	83%
JEFF freq 5	227	42	185	19%	82%

The WEB corpus was also analysed in the same way. Table 5.13 shows that in general precision was lower than that of the JEFF corpus, following the same pattern as was found earlier in the chapter that studied relations identified filtering solely for frequency. The rest of this section will look at the precision as automatically calculated. It will then go on to look at the adjusted precision taking into account false negative differences as in the detailed analysis in the previous section and any convergent or divergent properties of the relations identified by the frequency and salience filters, as well as any differences between the two corpora.

Table 5.13: WEB corpus: table comparing precision and relations identified with combinations of frequency and salience filtering

Filter	**WEB total**	**Not in VTO**	**In VTO**	**% only in WEB**	**% Precision**
WEB freq 4, sal 9	307	95	211	31%	69%
WEB freq 4, sal 10	288	87	200	30%	69%
WEB freq 4, sal 11	263	76	187	29%	70%
WEB freq 5	284	85	199	30%	70%

The rest of this section will look at a breakdown of the relations identified in both the JEFF and WEB corpora, filtered for frequency 4, salience 9 to 11, and compare these with the respective corpus filtered for frequency 5 as outlined in the frequency filter analysis earlier in the chapter.

Similarity comparison between the two representations

The first analysis performed was based on precision modelling. This time the precision modelling was used to compare the similarity between the frequency-only filter outputs and the dual threshold frequency and salience outputs. To perform these comparisons, the scripts used to compare the corpora to the VTO were adapted to compare the relations identified in each of the filter scenarios.

Table 5.14: Total relations and comparative breakdown JEFF filtered for frequency 4, various salience and frequency 5

			Relations		
filter	**JEFF freqsal**	**JEFF freq5**	**JEFFfreqsal only**	**JEFFfreq only**	**In both**
JEFF rel 4 sal 9	268	227	45	4	223
JEFF rel 4 sal 10	246	227	39	20	207
JEFF rel 4 sal 11	224	227	33	36	191

Table 5.15: Total relations and comparative breakdown WEB filtered for frequency 4, various salience and frequency 5

Filter	**WEB freqsal**	**WEB freq 5**	**WEBfreqsal only**	**WEBfreq only**	**In both**
WEB freq 4, sal 9	307	284	36	14	271
WEB freq 4, sal 10	288	284	32	29	256
WEB freq 4, sal 11	264	284	28	50	235

Table 5.16: Percentage similarity and divergence in the JEFF corpus between filter frequency 5 and frequency 4 with salience 9-11

filter	**% JEFFfreqsal only**	**% JEFFfreq only**	**% in both (JEFFfreq)**
JEFF freq 4 sal 9	17%	2%	98%
JEFF freq 4 sal 10	16%	9%	91%
JEFF freq 4 sal 11	15%	16%	84%

Tables 5.14 and 5.15 compare the respective corpora filtered for both frequency and salience (freqsal) and filtered only for frequency (freq). These tables indicate that the relations identified by the salience filter diverge from those identified by the frequency filter, by the fact that the number of the relations in both outputs (frequency only and frequency plus salience) as a proportion of the dual threshold filter relations identified remains stable throughout the different options analysed. However, the percentage of relations which appear in both representations

Table 5.17: Percentage similarity and divergence in the WEB corpus between filter frequency 5 and frequency 4 with salience 9-11

Filter	**% WEBfreqsal only**	**% WEBfreq only**	**% both (WEBFreq)**
WEB freq 4, sal 9	12%	5%	95%
WEB freq 4, sal 10	11%	10%	90%
WEB freq 4, sal 11	11%	18%	83%

as a proportion of the number of relations identified in the frequency only output increases dramatically as the salience filter increases. So, while the percentage difference/similarity remains stable across the filters analysed for the frequency and salience aspect, the percentage of relations uniquely appearing in the frequency 5 filter corpus increases as the salience increases (from 1.8% for frequency 4, salience 9, to 15.9% for frequency 4, salience 11, in the case of the JEFF corpus). The WEB corpus follows as similar trajectory. A breakdown of the percentage similarity/divergence can be seen in Tables 5.16 and 5.17. This indicates that the salience filter focuses on or selects different relations than the frequency filter, as the commonality of the results diverge.

Breakdown of similarities and differences between single and dual threshold

The previous section set out how the lower the salience filter threshold, the greater the similarity of relations identified between the frequency and salience. However, as the salience filter increases, these differences diverge. The increase in the percentage of relations only in the frequency-only filter representation, while the percentage only in the dual threshold representation remains fairly stable suggests that salience emphasises different relations to that of frequency, while also identifying the same relations as frequency at lower levels. The increasing percentage with increasing salience show these similarities diminish. The breakdown shows that the differences between the VTO and each representation are still heavily weighted towards synonyms and out-of-scope relations.

Graphs 9.35, 9.36, 9.37, 9.38, 9.39, 9.40 show that both synonyms and out-of-scope terms are still highly significant as regards the differences identified between the test corpora and the VTO. Where the salience and frequency filter seem to diverge is particularly as regards misspellings and partial names. The frequency filter tends to favour the identification of these relations. Salience consistently has higher numbers of incorrect matches, however it does not always have the lowest precision when weighted for synonyms and out-of-scope, which are prominent in the data. None of the differences are large enough to be able to calculate scientific significance from this data. A wider study would be needed to draw more concrete conclusions.

The misspellings and partial names which remain in the frequency filter, but disappear in the higher salience filters, seem to be related to the connectivity of the specific nodes, or the generality of the nodes. More connected nodes, such as Linneaus, which are not specific to one nomenclature term, but can be linked to many, are favoured by the frequency filter but eliminated by the salience filter. As regards misspellings, the ones related to the Salmonidae family, which is the most highly linked taxonomic family in both of the test corpora, are maintained through the frequency filter but also lost in the higher salience filters.

The presence of a large, connected section of the graph at the top of the frequency 5 graph image is notable by its absence in the dual threshold filter graph, when comparing the overview of the JEFF corpus between filter for frequency 5 (see Figure 9.41) and frequency 4, salience 11 (see Figure 9.42). The large, connected section in the frequency 5 filter graph represents the relations identified relating to the Salmonidae family. This is much less interconnected in the dual threshold filter graph. Figures 9.43 and 9.44 show a similar visualisation in the WEB corpus. This is interesting because it means that a brief glance can be informative about various trends within the representation.

Figures 9.45, 9.46 and 9.47 show close ups of the highly connected areas of the respective relation network graphs. These figures show that the nodes that are emphasised in the frequency filtering tend to be the same nodes that are removed through the salience filtering.

These three different relation network graphs (see Figures 9.45, 9.46, 9.47 show that the frequency-only filtered graph has the highest level of connectedness (in which the connections are made through the genus to genus level or rank skipping usually). The graph for frequency 4, salience 9 has been included to show the gradual change as salience filtering increases. The frequency only filter often links different genus nodes through the species-level nodes or through common terms such as the authorship of Linneaus. It also includes many more variants of different taxonomic entities. As the salience filter rises, the graphs gain clarity because they remove many of the more ambiguous links which are not true parent-child links and also the more general terms. However as salience rises further it removes many interesting links relating to the variety of terms linked to a specific genus term, so for the completeness of the image of variant names used for this family of fish then it does not provide as rich a picture in the higher salience filter graphs.

The WEB corpus provides a similar picture (see Figures 9.48, 9.49 and 9.50), although the connectedness of the graphs showing the salience filters remains more stable. The WEB corpus also extracts more of a hierarchy from the data, with Salmonidae and Salmoniformes appearing in the graphs and remaining through the different filter thresholds.

As was mentioned before, the lower salience is very similar to the frequency only filter. It is only with higher salience that there is greater divergence. As salience rises, it would seem that the salience filter seems to favour a less connected graph. It was found that the salience filter will remove those nodes which are more connected: such as the term Linneaus that can be linked to many nomenclature references. This can be seen particularly in Figure 9.45 in comparison with Figure 9.47. Also the different variants for many species of Salmonidae seem to drop off in the higher salience filter examples. This would make sense because the association measure, by looking at the strength of the relation between two words, will lower if it is found in many other contexts. Therefore, the Salmonidae family in this case is dropped as the genera are mentioned in the cases of multiple species, lowering the salience score. It is something to bear in mind when thinking about the purpose of the analysis. It does seem to be true that salience maintains many correct, single references to specific species or different nomenclature references. Therefore the choice of filter depends on the purpose of the analysis. If this purpose is to get a broad overview of the main references to species or families within a particularly corpus, the data would indicate that the frequency filter would be the appropriate choice, whereas salience may be more relevant if outlying species were topics of interest. Further work must be done in this area to be able to draw more concrete conclusions.

Dual threshold: adjusted precision

The detailed analysis earlier in the chapter provided a detailed breakdown of the differences between the VTO automated precision score and a manual precision score to identify any relations accurately identified by the method not picked up through the automatic analysis. This also had to be performed with the dual threshold filter to be able to compare the relative precision of these filtering options against the filter 5 baseline. Figures 9.51 and 9.52 show that the automatic precision score would indicate that the frequency filter is generally better at least in comparison with the lower salience measures. However, when accounting for out-of-scope and synonym usage the image starts to shift.

Here the impact of the salience filter on less frequent, less connected relations in the corpora is particularly clear. When accounting for synonyms and out-of-scope matches, Figure 9.51 demonstrates that in fact the JEFF frequency 4, salience 11 has slightly higher precision throughout. As highlighted in the previous section, this does not necessarily mean that it is the better choice, it depends on the purpose of the analysis - the exclusion of a number of interesting terms relating to the Salmonidae family which would be lost through this form of filtering. However, it does indicate a wider variety of real different nomenclature terms and relations are identified and maintained in the dual threshold representation.

The WEB corpus provides a similar but less consistent representation. There is a heavy weighting to out-of-scope terms in the salience numbers, but when accounting for synonyms alone then the frequency only filter still maintains higher precision. However when looking at any measure accounting for scope the difference is quite clear that the precision of the salience filter variable is equal to or greater that of the frequency filter.

5.4 Discussion

This chapter has outlined results which demonstrate that the research method developed for this thesis produces reliable results across both test corpora as regards precision against a reputable knowledge source. In the absence of a gold standard against which to measure, the validation/evaluation has demonstrated how my method can be used to analyse the differences between the knowledge resource being used as a guide stick and the information extracted from the test corpora, quantitatively and qualitatively. The fact that the results show stability across both corpora, with descending numbers of incorrect matches identified as filtering levels increase, is another indicator of the reliability of the method. Finally the analyses performed indicate that frequency and salience could be measures to be used either on their own or simultaneously. The results of these analyses indicate that each parameter emphasises different characteristics of the data. This could be a focus of future work to better explore these issues.

The evaluation here is a technical one, and is supported by expert input in Chapter 7 in a more formal external evaluation of the data extracted and analysed in this thesis. This was performed to be able to provide a different perspective to the findings, and was particularly useful in cases of ambiguity.

The fact that the results are relatively stable across the two test corpora provides some evidence of reliability and the possibilities of generalising the results found here, further tests on corpora of different types and from different domains within biodiversity would be necessary to be able to strongly argue that this was the case. This is another focus of future work to test the corpora on a wider variety of data.

The granularity of the relations identified require more work to try to identify sibling-sibling relations and reduce the number of relations identified in which like items are grouped. That said, the evidence would point to the reliability of the method. The evaluation here shows that the method design used does indeed produce the data analyses that it sets out to do, and in a sufficiently consistent and reliable way. This could be used by taxonomy experts interested in looking at patterns of usage of scientific nomenclature across different domains, times or authorship. It could also be used when intending to perform integration processes of varied datasets to check that the integration will be successful. It could be used to look at patterns of common name usage across countries and the links between these common names and scientific nomenclature. All of which would aid in the assurance that data is not integrated erroneously. In other areas, these techniques could be adapted to look at terminology change and conceptual meaning behind terms to help to gather information about the stability of such terms intra- or inter-domain. Comparison of conceptual stability across languages would also be a possibility.

Chapter 6

Phase 3: Nomenclature profiling studies

This chapter aims to respond to Objectives 2 and 4 of the thesis. Objective 2 is “to create a graph/tree hierarchy image of this model to compare to the ontological structure for validation and evaluation purposes”; Objective 4 is “to perform comparisons between the hierarchies extracted between different corpora and ontologies of choice to evaluate the conceptual stability of nomenclature references”. This phase in relation to the design science research structure could be considered to be in part related to the design cycle because it was in this phase the actual structure of the nomenclature profile studies was defined. However it would also in part form part of the relevance cycle in its link to the evaluation stage because the results constitute the application of the method on which experts are asked to feed back.

The chapter begins with an analysis of three different taxonomic resources to demonstrate the variability within different representations of the scientific nomenclature described in the literature review. The following section of the chapter consists of three nomenclature profiling studies based on the evidence from the chosen taxonomic resources, which was then cross-examined using the test corpora data. The three taxonomic entries identified for study were chosen due to specific aspects of nomenclature variant-usage (existence of known synonyms in the corpora) or frequency in the test corpora. The idea was to present three different profiles from which to extrapolate guidelines as to behaviour patterns for future applications.

6.1 Evaluation and analysis of different taxonomic resources

The following sections will consider three different taxonomic resources: Vertebrate Ontology Taxonomy (VTO), Integrated Taxonomic Information System (ITIS) and the Catalogue of Life (CoL). The VTO was chosen because of the focus on vertebrate, specifically fish, species, therefore suitable for the test corpora content. The format of the resource comes in what is called the obo format, a biology-oriented language for building ontologies, based on the principles of Web Ontology Language (OWL). This format was an advantage, because it was easy to convert into a usable format for the research. It is included in various knowledge resource platforms of official standing such as the European Bioinformatics Institute [51]. ITIS and CoL were chosen because both ITIS and the CoL came up as respected sources in the background research. Both ITIS and CoL also share many collaborators and are part of the same initiatives for taxonomy integration, so I felt that this was an interesting choice to be able to compare the different representations across the sources. The following sections will provide a breakdown of the people involved, the purpose of the ontology/knowledge representation resource and any other relevant information which will help the reader to understand the position the resource takes as regards scientific nomenclature, its variants (and vernacular variants, if applicable) and the resulting differences in the resources as regards approaches and content.

6.1.1 Vertebrate Taxonomy Ontology (VTO)

The VTO was created "to fill the need for a single taxonomic ontology including both modern and ancient vertebrate taxa" [150]. The ontology was created by integrating a number of different existing resources and also used a ranking system that has been kept separately to promote integration and merging of other systems used.

The existing resources used to produce this resource is the NCBI taxonomy and the Paleobiology Database, with further information incorporated from Teleost Taxonomy Ontology (TTO) (based on the Catalog of Fishes),and AmphibiaWeb to provide more detailed information about specific taxonomic groups.

The data base identifiers are:

TTO: Term defined in the TTO
NCBITaxon: ID from NCBI taxonomy database
CASSPC: ID from the Catalog of Fishes species table (CAS = California Academy of Sciences)
CASGEN: ID from the Catalog of Fishes genus table
TAXRANK: Term defined in the Taxonomic Rank Vocabulary
FISHBASE: Term defined in Fishbase (usually common names) AWeb: Term defined in Am-

phibia Web [150]

Unlike the CoL and ITIS, VTO is, strictly speaking, an ontological knowledge base. Based on a taxonomy it would not be described as a full ontology, but it does link out through codes to further information about species in the NCBI and CASSPC. Synonyms are described as related (for alternative scientific nomenclature) or related common (vernacular names for said species). Information about vernacular names is sparse and not a focus of the ontology, focusing much more strongly on the scientific nomenclature, the hierarchy and links between the different species. The taxonomic ranking is marked through the agreed codes using the taxonomic rank vocabulary. At the time of writing it consisted of 107,137 classes, which are the accepted name entries within the ontology [89]. The annotation properties related to synonyms include the following:

- synonym_type_property (which can be common name, misspelling or name with (author year)
- has_related_synonym
- has_exact_synonym
- has_narrow_synonym [53]

While "exact synonym" is theoretically a possible annotation property, it only appears three times, all of which reference PaleoDBtaxon (Paleobiology database), rather than providing an actual synonym name in the ontology. There are 20 examples of an "exact misspelling" which is not officially included in the annotation properties as given on the website. Unlike the ITIS and the CoL, in most cases the VTO does not include the authorship (author and date) in the accepted name entry. Midford [150] describes the changes that take place in the nomenclature within the context of taxonomy as a result of changing ideas to do with categorisation of a taxon. He highlights the importance of being able to group synonyms no matter what the categorisation of said name variant, which appears to be a motive for having a more generic name in the accepted name entry. This is more similar to the ethos of the CoL than the ITIS.

6.1.2 Integrated Taxonomic Information System (ITIS)

The ITIS is a resource that aims to provide "authoritative taxonomic information on plants, animals, fungi, and microbes of North America and the world" [103]. The US Department of Commerce, National Oceanic and Atmospheric Administration (NOAA), Department of Interior (DOI, Geological Survey (USGS), Environmental Protection Agency (EPA), Department of

Agriculture (USDA), Agriculture Research Service (ARS), Natural Resources Conservation Service (NRCS), Smithsonian Institution and the National Museum of Natural History (NMNH) were the original partners involved in working towards producing a resource of "scientifically credible taxonomic information" [104]. This has grown to include other agencies across Canada and Mexico (ITIS-North America) and other organisations and taxonomic specialists. ITIS is also partnered with Species 2000 [225] and the Global Biodiversity Information Facility (GBIF) [71]. The ITIS and Species 2000 Catalogue of Life (CoL) are used as the basis for the taxonomic backbone of Encyclopedia of Life (EOL) [55]. The Encyclopedia of Life is an initiative to provide encyclopaedic information about species to both the general public and professionals.

Data sources The original data source was based upon taxonomic data maintained by the NODC and the NOAA, called the taxonomic data code. This original dataset included approximately 210,000 scientific names. These entries were of varying quality, with errors including missing taxonomic groups, misspellings and typographical errors. Species' names without the proper authorship citation (author and date) and incorrect allocation of species within groups were mentioned as other problems with some of the entries [106]. The ITIS work has focused on two main things: adding highly credible new names or checklists and also reviewing the original NODC data, verifying and classifying it to ensure a higher standard of data quality. As of August 2019, "ITIS has grown to more than 804,000 scientific names, more than 89% of which have been verified in the literature, leaving about 91,000 names as unverified legacy data" [106].

The database has been produced manually, with experts working to link recorded names to one or more credible references such as print publications, recognised experts or databases. The use of experts to verify entries according to reliable sources is central to the initiative.

Description of data structure/definitions When searching the database for species' names, the database provides a page that presents information as the following: kingdom, taxonomic rank, synonym(s), common name(s), taxonomic status and record credibility rating. Synonyms have hyperlinks which link out to their own related page. The taxonomic status categorisations for Animalia, Archaea, Bacteria and Protozoa are as follows:

- homonym & junior synonym
- junior homonym
- junior synonym
- misapplied

- nomen dubium
- nomen oblitum
- original name/combination
- other, see comments
- pro parte
- subsequent name/combination
- unavailable, database artifact
- unavailable, incorrect
- orig. spelling
- unavailable, literature misspelling
- unavailable, nomen nudum
- unavailable, other
- unavailable, suppressed by ruling
- unjustified emendation
- unnecessary replacement
- unspecified in provided data [107]

The number of types of variants itemised here may come as a surprise to the layperson, who may not understand the taxonomic process or the process of naming and renaming that occurs as part of this perpetual reevaluation. As was identified in the literature review, laypeople and specialists of areas outside taxonomy understand a taxon as something fixed, on which a label is placed. However, taxonomists understand a taxon as a hypothesis, something that is constantly being reviewed and that is subject to reassignment or re-circumscription [16]. The nomenclature should be updated and/or changed, provided with extra markers such as the author or the date of the circumscription and linked to the specific specimen used to create that circumscription and used accordingly. However, because of extensive usage of scientific nomenclature in the vast array of domains, this is not always followed and the interpretation of the meaning of one term or another may not always be the same. This is similar to what is explored in the introduction and the literature about the meaning of words in general, in that meaning is assigned according to the criteria given for a specific purpose, words do not inherently possess specific meanings, despite how we might think of them doing so.

ITIS also considers such variants as valid or invalid. To date, from the research carried out in the nomenclature profiling studies, the assessment as valid as only been seen in relation to accepted scientific names, whereas all other variants have been classified invalid although no exhaustive search of the database was performed.

Below the first part of the ITIS entry there is a taxonomic hierarchy which lays out where the species fits within the taxonomic hierarchy of species according to ITIS and subspecies defined as direct children. Below that are the references linked to the validation of this name. Any invalid synonyms are not included in the aforementioned hierarchy, with only references from experts, publications and other sources where this name/taxonomic definition can be found. Common names have no hyperlinks out but are included in various languages where available. The taxonomic status, as can be seen in the list above, gives information as to the validity or reasons for the synonym occurring. Finally, the record credibility rating provides information as to whether the record has been verified by experts and if such standards have been met. They are then given a score accordingly.

Data formats Entries into the ITIS database can be downloaded in TWB and DwC-A formats [105]. TWB is a taxonomic workbench format, which consists of a .csv file in which the different aspects of the entry are included, whereas the DwC-A format includes entries only which are included in the Darwin Core vocabulary in a text file with accompanying .XML files. See appendix E for links to the files as an example of each (Oncorhynchus mykiss entry). It is too large to put in the main text of the thesis.

6.1.3 Catalogue of Life

The Catalogue of Life consists of a compilation of checklists from 172 taxonomic databases. The aim is to produce a “comprehensive catalogue of all known species of organisms on Earth” [43]. The 2019 edition contains 1,837,565 living and 63,418 extinct species. The data has been compiled on the basis of work carried out by large networks of specialists. As with ITIS data, databases are peer-reviewed, and different data is then selected and integrated to form a “single coherent hierarchical classification”, in this case by the Species 2000 and ITIS teams [43]. It is funded primarily by the European Commission but also many other institutions across the world. The Catalogue of Life is the product of a collaboration between the Species 2000 and ITIS organisations that started in 2001, with the aim of joining forces and being able to make better, more efficient use of resources. The combined Annual checklist is used as the main taxonomic index in the GBIF and EoL data portals. It is also recognised by the CBD [43]. Much of the focus of the Catalogue of Life it to encourage integration of biodiversity information and interoperability between different existing sources. It aims to list every different species from each group of organisms, bringing these together in such a way as to have one reference which links to all known synonyms and vernacular terms for each taxon in one place. There should be references for all recorded uses and information by which alternative taxonomies can be followed. Integration and the creation of a common understanding and usage of terminology is a key aspect to this initiative. Over 200 expert taxonomic databases from around the world have been integrated so far, with over 3,000 taxonomic specialists collaborating on the project [43]. As with the ITIS, the original information is not always complete. Incomplete information is included through semi-automatic processing and is only put through the same expert scrutiny as complete checklists when a fully-validated species database is available for a particular group.

Each species in the Catalogue of Life is listed with an accepted scientific name, a cited reference and its position in the hierarchical classification. In addition, common names, synonyms, and distribution and ecological data are provided, but these data may not be complete. The list of field groups (known as the "Catalogue of Life Standard Dataset") is given below:
(1) Accepted scientific name with references
(2) Synonyms with references
(3) Common names with references
(4) Classification above genus
(5) Distribution
(6) Life Zone/Environment
(7) Current and Past Existence
(8) Additional data (optional)
(9) Latest taxonomic scrutiny (specialist name and date)
(10) Source database name and version
(11) Link to online resource [42]

Catalogue of Life is intended for use by research scientists, policy and decision makers and citizen scientists. One of its primary objectives is to facilitate searching across the globe for species under their many different guises (synonyms or alternative names, alternative spellings or common or vernacular names). In comparison with the ITIS database, the focus in the CoL seems to be to facilitate the easy searching of all possibilities which might link to a specific taxon, and be able to validate the name they are using or find the accepted name. There is also a strong focus on the multilingualism and different vernacular names for specific species, to be able to link these to have a common understanding and be able to search and arrive at the same page through what the CoL call synonymic indexing. Part of the underlying data structure in the Catalogue of Life is to link all these names when performing a search in what they call “synonymic amplification”, which means that when someone searches for something it will search for all known synonyms of this taxonomic entity. The Catalogue of Life does not put the same emphasis as ITIS on the validity or invalidity of synonyms. It simply recognises them as existing (on the webpage entry synonyms which are classified as “invalid” in ITIS are classified simply with the term “synonym”. An example of this would be Salmo gairdneri Richardson 1836 [35, 155]. Catalogue of Life seems to take a position as regards synonyms that leans more towards collating all terms that might be found in the literature (every entry links to the reference information of at least one publication in which the taxonomic name is apparently linked to the accepted name). The concept of accepted scientific name is consistent with what ITIS classifies as a “valid” entry in the taxonomic status section of the entry (refer back to previous entry).

6.1.4 Definition of accepted names and synonyms across the different taxonomic resources

The preceding sections have provided an overview of the knowledge representation sources that will be used in the subsequent nomenclature profiling studies. The analysis has made reference to the aims and objectives of each initiative, the data sources and involved parties, and the types of information and how it is presented in each resource. This section will provide a deeper analysis as to the differences between the three sources, to investigate how differently the same information can be presented.

This provides evidence as to the need to understand empirically the use of nomenclature in a particular context because it may be more appropriate to use one or other resource. Should this be done without an analysis of the conceptual representation between nomenclature references in a corpus in comparison with knowledge resources, there is a risk of imposing an incorrect conceptual model onto the data or ignoring large parts of it.

Looking at the comparison between the different resources, both the VTO and CoL present the nomenclature terms which are the official, accepted names as “accepted names”. They use the same term to describe it. ITIS, however, describes this term as having “valid” taxonomic status, stressing the validity or invalidity of certain name variants. Moving to the actual content of this part of the entry, the different sources present accepted names in different ways. The inclusion of authorship is something which is mentioned time and again in the literature as the way in which different circumscriptions of the same nomenclature can be identified (described as a taxonomic concept), as one way of eliminating such ambiguity [64, 171], but also that this standard is not followed across the board when the names are used in the literature. In fact, the Zoological Nomenclature code, although it advises the use of authorship, does not require it [102]. Both ITIS and the CoL include the authorship, indicating that the conceptual basis of what constitutes an accepted name is the same for both of these resources as regards the parts that comprise an accepted name, despite the different emphasis that the ITIS gives to the validity of a term. The VTO does not include authorship, except in some cases in which the version with the authorship is included as a synonym in the class entry. This, besides the claim as highlighted in the VTO section to ensure all related terms can be kept under the same umbrella, might also reflect the the purpose of the resource if the aim is to use it to identify mentions in real texts. As mentioned previously, the VTO is the only true taxonomy ontology of the three resources, whereas the other resources are more for taxonomic information retrieval through searches. This distinction is important to note and highlights important differences in possible applications and therefore requirements when developing such resources. It also links to potential issues in integration (narrower and broader meaning) or where there are ambiguities because of subsequent or differing circumscriptions. To help with this, the VTO includes references which provide a reference code that links to where the information is from (cross-referencing to other ontologies or the unique identifiers involved in the Global Name Architecture initiative). This should provide further information as to the circumscriptions to which the specific entry refers. Unfortunately, when some of these links were tested many were broken, which highlights a common issue within digital media of data maintenance [96] and was mentioned in the expert evaluation in Chapter 7.

Moving onto synonyms, each resource conceptualises them differently:

VTO: exact, related, specifically in terms of whether the synonym conceptually refers to the same thing

ITIS: the term synonym is used on the page where linked, but on its own page defined as valid/invalid (taxonomic status)

CoL: validity or exactness not considered, simply the fact that there is a link with an associated reference and it is described as the same thing. Synonyms defined as synonyms, ambiguous synonyms or misapplied names.

The ITIS classifies names as having valid or invalid taxonomic status to impose an authoritative structure and in the aim to homogenise and standardise usage of the accepted terms. This also shows a definition that focuses on the nomenclature usage rather than a focus on the biological taxa that the name represents. The VTO, through its categorisation of synonyms as exact or related, emphasises the conceptual element of taxonomic definitions and the fact that different descriptions may vary in the exactitude of the concepts that they represent. The fact that there are next to no exact synonyms present in the ontology provides evidence as to the perspective of those curating the ontology as to the possibility of exact synonyms behind the multiple names given to different taxa. Finally, the CoL and the lack of descriptions as regards these synonyms reveals its perspective of being more interested in linking all possible names and descriptions together to enable a more global perspective of the situation. It places more emphasis on collating the widest scope of literature written about specific taxa together and not missing it because of name variation, than the exactitude of the synonymity or validity of the names used.

Finally, the inclusion or exclusion of common name variants is of interest. The VTO places the least emphasis on common names, with entries often being void of any vernacular variant. Where they do exist they are described as "related common name", which alludes to the fact that a common name cannot be an exact synonym of the accepted scientific nomenclature and what that represents. The ITIS includes common names as a standard in the entry. They are defined as "common name(s)" and from the limited search performed tend to appear primarily in English, sometimes with Spanish and French common names variants included too. No more than one or two variants for each language are usually present. The CoL entries place the most emphasis on common name variants. The examples linked to each entry are extensive and specify both language and a country of origin for each variant, as well as a reference link to where the example was found. To summarise the properties within these resources relevant to the subsequent analyses, please refer to Table 6.1.

Table 6.1: Comparison of relevant properties across VTO, ITIS and CoL

Resource	**Properties**			
	Accepted name	**Authorship usage**	**Synonym**	**Common names**
VTO	Accepted name	Not usually	Exact or related	Related common name
ITIS	Valid taxonomic status	Yes	Invalid taxonomic status	Common name
CoL	Accepted name	Yes	Synonym; ambiguous synonym; misapplied name	Common name

This comparison supports the argument made in the Introduction and Background (Chapters 1 and 2) as to the differing perspectives taken by different resources as to the qualities and features of nomenclature variants.

For the purpose of the following nomenclature profiling studies, all related synonyms, as well as synonyms with invalid and valid taxonomic status were considered under the more general umbrella of synonyms for the reason that they are terms used in a broadly synonymous fashion in the literature. The nature of these synonyms in context were analysed to see if this revealed anything about differences in their usage and therefore the conceptual nature of each term.

Having performed an analysis of the approaches taken by the knowledge representation resources chosen, the next section presents the format of the nomenclature profiling studies which were used to apply the techniques developed throughout the rest of this thesis. The nomenclature profiling studies take the following format:

- Comparison of nomenclature terms included in the relevant entry for each resource. The comparison is used to evaluate the scope of each resource, to see the convergence and divergence of the actual terms used and if resources are consistent in their representations or if there are important differences between the different representations.
- Frequency and dispersion comparison of mentions from the JEFF and WEB corpora (identify general characteristics of accepted names, synonyms and vernacular variants). These analyses are used to identify patterns of behaviour within the test corpora for different nomenclature variants, whether patterns of behaviour of accepted names versus other variants can be identified, if combined dispersion patterns can be used to draw conclusions about meaning between different variants and also to look at the stability or instability of variant usage across different corpora.
- Analysis of the representation extracted from the JEFF and WEB corpora (profile usage of specific terms). This analysis is an extension of the previous, continuing to compare behaviour patterns and specific usage examples through lexicographic techniques to draw conclusions about the usage of specific variants within the test corpora.
- Compare data found in the JEFF and WEB corpora to identify (in)consistencies and gaps in the resources and between corpora. This final stage draws together the different representations profiled to be able to draw conclusions about usage across the test corpora, the stability or lack thereof of the representations across the test corpora and identify any gaps in the resources analysed.
- Links to Word Sketches and supporting concordances can be found in Appendix E for further information should they be required

6.2 Nomenclature profiling study 1: Oncorhynchus mykiss

Oncorhynchus mykiss was chosen as the first study as a result of the analysis performed on the JEFF and WEB corpora in the previous phases. In the analysis two alternative scientific nomenclature terms were found for Oncorhynchus mykiss (Parasalmo mykiss and Salmo gairdnerii, as well as a number of literature misspellings/synonyms (resources classified these in various ways), such as Onchorhynchus mykiss. This, along with the fact that these terms appeared frequently in the corpus, made the species seem like a suitable subject for the first nomenclature profiling study.

6.2.1 Oncorhynchus mykiss in the taxonomic resources

Comparison of the accepted names In the comparison of the VTO, the CoL and the ITIS, both the CoL and the ITIS record exactly the same accepted name: Oncorhynchus mykiss (Walbaum, 1792). The VTO includes the same nomenclature but excludes the authorship.

Comparison of the synonyms The three resources differed to a much greater degree as regards synonyms. There was only one exact match between two knowledge resources: Onchorhynchus mykiss (Walbaum, 1792) between the ITIS and CoL. In the ITIS this synonym was classified as "invalid - unavailable, literature misspelling", whereas in the CoL this was simply classified as a synonym. The ITIS counted a total of four synonyms: Salmo mykiss (Walbaum, 1792), classified as "invalid - original name/combination", the aforementioned Onchorhynchus mykiss, then Salmo gibsii (Suckley, 1859) and Oncorhynchus mykiss gibbsi (Suckley, 1859), both considered "invalid - junior synonym".

In the ITIS representation, a number of direct children were identified, these being classified as subspecies. The other knowledge resources do not seem to split these variants in the same way and lump them together with any synonyms. A further discussion of lumping and splitting of species can be found in the expert evaluation in Chapter 7. The ITIS entry also includes 3 common names for the English language (rainbow trout, steelhead, and redband trout).

The VTO includes 27 different scientific names as synonyms, and one common name (rainbow trout). The CoL includes 34 different scientific names as synonyms, and 51 common name variants for the English language. This shows the emphasis the CoL places on having an extensive scope, as it seems to promote inclusiveness in classification.

While the CoL and VTO have similar numbers of scientific nomenclature synonyms, there were no exact matches across the different synonyms. There were 15 partial matches between the two knowledge resources, all of which differed in the inclusion or exclusion of the authorship details (CoL included, VTO excluded), as described in the preliminary analysis. This still left 31 entries across all knowledge resources that were unique to the knowledge resource they were found in (19 in CoL, 11 in VTO and one in ITIS). This equates to 56% of the CoL synonyms uniquely appearing there, 41% uniquely appearing in the VTO and 25% uniquely appearing in the ITIS. This first study seemed to demonstrate the vast numbers of possibilities in existence in the nomenclature and also the varying perspectives and purposes of said knowledge resources and how this affects what is it included or not. These are all considered authoritative figures in scientific nomenclature, but the need for conciseness and validity in the ITIS could clearly be seen above the CoL, whose purpose is primarily to collate as much of the available data, including as many of the variations of names as possible. It is also interesting because the CoL and ITIS work in close collaboration with each other [43].

6.2.2 Oncorhynchus mykiss in the JEFF and WEB corpora

Frequency and incidence in comparison with ITIS, VTO and CoL

Both raw frequencies and frequencies per million (normalised, or relative frequencies) were analysed. Both were valid measures for analysis due to the focus on trends of comparative frequency of the terms within the same corpus, and to see if these trends remained stable across corpora. However, normalising allowed for the relative prevalence of each nomenclature term to be compared across the two corpora. The name variant preference ranking was used across corpora to get a feel for the relative preference of certain terms.

Tables 6.2 and 6.3 and Figure 9.53 show that for the first six variants (including trout and brown trout) both corpora follow the same preference of variants. Figure 9.53 is based on a logarithmic scale to reveal the the scale of differences between the very frequent variants right down to the very infrequent. After the first six, the ranking varies somewhat, but the numbers are very small so no specific conclusions can be drawn. Brown trout and trout were not considered in this analysis. Trout is an umbrella, much more general term. Also, concordances and other analyses later in the chapter clearly linked brown trout to Salmo trutta, another species of the Salmonidae family. Further information as to this decision can be found in the Subsection relating to Relation Network Graphs later in this profiling study. Looking more generally at frequencies, the table and graph show that vernacular variants appeared most frequently, specifically the variants of rainbow trout and steelhead trout. The accepted name for this species, Oncorhynchus mykiss, was the third most frequent variant, which shows usage reflects the categorisation of all three ontologies. These three variants, plus the Salmo gairdneri variant to a lesser extent, far exceed the frequencies of other variants. To demonstrate more clearly the strong preference for these four variants (excluding trout and brown trout), Figure 9.54 uses a linear scale of normalised frequencies which compare the two corpora. This also shows that the emphasis on these species is much higher in the WEB corpus than the JEFF corpus, which could be a result of the choice of seed words in its compilation (see Appendix A.4 for seed word list), which were based on keyword from the JEFF corpus. Therefore the increased prevalence of occurrences of this nomenclature references cannot be used to draw any conclusions.

Neither corpus had examples of any name variant with the full authorship as detailed in the ITIS and CoL taxonomic resources. This is interesting to bear in mind when thinking about usage and supports the choice of the VTO to present the information in this form, although taxonomically it is less exact. The chart above only shows the details for name variants included in the respective taxonomic resource entries that also appeared in the corpora. Appendix E contains the breakdown of all the name variants for Oncorhynchus mykiss included in each

Table 6.2: Frequency comparison and ranking of name variant frequencies

Resource	Name variant	Frequency in JEFF corpus	JEFF corpus ranking	Frequency in WEB corpus	WEB corpus ranking
CoL	trout	6767	1	8108	1
CoL	brown trout	3274	2	3159	2
ITIS/CoL/VTO	rainbow trout	949	3	2533	3
ITIS/CoL	steelhead	515	4	1806	4
VTO	Oncorhynchus mykiss	240	5	809	5
CoL	rainbow	207	6	407	6
VTO	Salmo gairdneri	57	9	195	7
CoL	steelhead trout	104	7	158	8
CoL	redband	2	15	42	9
CoL	Kamloops	1	17	33	10
ITIS	redband trout	10	10	29	11
CoL	salmon trout	7	11	21	12
CoL	bow	71	8	19	13
CoL	Kamloops trout	0	22	16	14
CoL	hardhead	0	22	14	15
CoL	silver trout	0	22	10	16
VTO	Oncorhynchus mykiss gairdneri	1	17	7	17
VTO	Oncorhynchus mykiss aguabonita	1	17	6	18
VTO	Salmo mykiss	3	14	5	19
VTO	Oncorhynchus mykiss irideus	1	17	5	19
VTO	Salmo gairdnerii	1	17	5	19
CoL	Kamchatka steelhead	2	15	3	22
VTO	Parasalmo mykiss	5	12	2	23
CoL	baiser	5	12	1	24
CoL	coast rainbow trout	0	22	1	24
CoL	coast angel trout	0	22	1	24
CoL	coast range trout	0	22	1	24
CoL	summer salmon	0	22	1	24
CoL	Kamchatka salmon	0	22	1	24
VTO	Salmo whitei	0	22	1	24
VTO	Oncorhynchus kamloops	0	22	1	24
VTO	Salmo masoni	0	22	1	24
VTO	Salmo nelsoni	0	22	1	24
VTO	Salmo purpuratus	0	22	1	24

Table 6.3: Frequency per million comparison and ranking of name variant frequencies

Resource	Name variant	JEFF corpus (freq per mill)	Ranking	WEB corpus (freq per mill)	Ranking
CoL	trout	1325.5	1	2302.23	1
CoL	brown trout	641.3	2	514.83	2
ITIS/CoL	rainbow trout	185.89	3	412.81	3
ITIS/CoL	steelhead	100.88	4	294.33	4
VTO	Oncorhynchus mykiss	47.01	5	131.84	5
CoL	rainbow	40.55	6	66.33	6
VTO	Salmo gairdneri	11.16	9	31.78	7
CoL	steelhead trout	20.37	7	25.75	8
CoL	redband	0.39	15	6.84	9
CoL	Kamloops	0.2	17	5.38	10
ITIS	redband trout	1.96	10	4.73	11
CoL	salmon trout	1.37	11	3.42	12
CoL	bow	13.91	8	3.1	13
CoL	Kamloops trout	0	22	2.61	14
CoL	hardhead	0	22	2.28	15
CoL	silver trout	0	22	1.63	16
VTO	Oncorhynchus mykiss gairdneri	0.2	17	1.14	17
VTO	Oncorhynchus mykiss aguabonita	0.2	17	0.98	18
VTO	Salmo mykiss	0.59	14	0.81	19
VTO	Oncorhynchus mykiss irideus	0.2	17	0.81	19
VTO	Salmo gairdnerii	0.2	17	0.81	19
CoL	Kamchatka steelhead	0.39	15	0.49	22
VTO	Parasalmo mykiss	0.98	12	0.33	23
CoL	baiser	0.98	12	0.16	24
CoL	coast rainbow trout	0	22	0.16	24
CoL	coast angel trout	0	22	0.16	24
CoL	coast range trout	0	22	0.16	24
CoL	summer salmon	0	22	0.16	24
CoL	Kamchatka salmon	0	22	0.16	24
VTO	Salmo whitei	0	22	0.16	24
VTO	Oncorhynchus kamloops	0	22	0.16	24
VTO	Salmo masoni	0	22	0.16	24
VTO	Salmo nelsoni	0	22	0.16	24
VTO	Salmo purpuratus	0	22	0.16	24

resource, to compare.

The relative frequency differences between popular scientific and common name variants within each corpus is worthy of comment. At various points in the literature review, remarks were made as to the lack of vernacular name usage in academic work [209], and to the inclusion of vernacular names for non-expert users [150] (implying lack of relevance in academic work). One could therefore predict that vernacular name usage would be comparatively more frequent than scientific nomenclature usage in a non-academic corpus. However, the analysis shows that the proportions of instances of common variant usage versus scientific nomenclature usage between both corpora are very similar (over four fifths common name usage versus less than one fifth scientific nomenclature) (see Table 6.4). These ratios suggest that in both corpora the usage of scientific nomenclature versus vernacular names is, on average, similar. This potentially suggests that there is no difference between narrative texts in their usage of common versus scientific nomenclature between academic and other texts, in the domain of fish. This is an indication for potential future work, in which corpora distinctly defined to respond to such a research question could be used.

Table 6.4: Proportion of scientific nomenclature (SCI) to vernacular variants (COM)

	Total SCI	**Total COM**	**% sci**	**% com**
JEFF	309	1873	14%	86%
WEB	1039	5097	17%	83%

Variant coverage

Table 6.5: Comparison of variants recorded in VTO, ITIS and CoL versus coverage in the respective corpora

	Total	% coverage
No of variants: VTO, CoL, ITIS	88	
Number of variants (SCI)	64	
Number of variants (COM)	24	
JEFF total	21	24%
WEB total	34	39%
JEFF number SCI	8	13%
WEB number SCI	13	20%
JEFF number COM	13	54%
WEB number COM	21	88%

The next analysis consisted of analysing coverage, or the percentage of variants which appeared in each corpus in comparison with the analysed taxonomic resources. Coverage is described in more detail in the methodology chapter and here consisted of comparing the taxonomic resources being studied and the data extracted from the respective corpora. Table 6.5 gives the tabular breakdown. Coverage analysis could be used to identify a "best match" to which to map the corpus, as well as provide an analysis of any gaps in the taxonomic databases. In the JEFF corpus just under a quarter of the total options appear, whereas the WEB corpus covered 39% of the names included across the sources chosen, with 21 of the 34 (approximately three-fifths) being vernacular variants. As regards the JEFF corpus, eight of the mentions are scientific nomenclature, with the remaining 13 being vernacular variants, meaning that approximately two-fifths of the scope is attributed to scientific names, with the other three-fifths being vernacular terms. The coverage of the WEB corpus is broader than the JEFF one consisting purely of scientific articles, although as was mentioned earlier no conclusions can be drawn from this. However, the high number of vernacular variants that appear in both corpora indicates prevalence of common variants in both cases, which seems to partially undermine the assertion that common names are not important in the scientific literature and also highlights variation in both corpora, despite the intuition that the academic corpus would demonstrate more informed and stable usage than the web-scraped corpus, which is a mix of many different sorts of texts.

Table 6.6: Corpora coverage of VTO

	Total	% coverage
No of variants: VTO	27	
Number of variants (SCI)	26	
Number of variants (COM)	1	
JEFF SCI match with VTO	8	32.5%
JEFF COM match with VTO	1	100%
WEB SCI match with VTO	13	50%
WEB COM match with VTO	1	100%

Tables 6.6, 6.7 and 6.8 show that only the VTO had exact matches with either corpus as regards scientific nomenclature because of the authorship. Had the authorship not been included as a necessary criterion to match the name, CoL would have had 3 matches, but ITIS would just have the accepted name as a match. Depending on the reason for mapping corpora in future applications, it could either be good to maintain the authorship as a requirement for a match (the lack of authorship leaving the term usage as too ambiguous to be useful for mapping) or could signify that information would be needlessly excluded. None of the synonyms listed in ITIS appeared in either corpus.

Table 6.7: Corpora coverage of CoL

	Total	% coverage
No of variants: CoL	58	
Number of variants (SCI)	36	
Number of variants (COM)	23	
JEFF SCI match with CoL	0	0
JEFF COM match with CoL	13	57%
WEB SCI match with CoL	0	0%
WEB COM match with CoL	21	91%

Table 6.8: Corpora coverage of ITIS

	Total	% coverage
No of variants: ITIS	7	
Number of variants (SCI)	4	
Number of variants (COM)	3	
JEFF SCI match with ITIS	0	0
JEFF COM match with ITIS	3	100%
WEB SCI match with ITIS	0	0%
WEB COM match with ITIS	3	100%

Coverage for common variants was fairly high, ranging from 57% as regards the JEFF corpus, to 100% for both the JEFF and WEB corpora as regards the ITIS database. In this case, the ITIS would exclude many common and scientific variants that would be identified by the other resources. It is interesting to see that the WEB corpus covers 91% of the variants provided in the CoL. This can be considered a demonstration of the pervasiveness of multiple common variant usage, and indicates that being able to map the contextual patterns in which these variants are used would be a useful method for data mapping. This was confirmed in the discussions in Chapter 7.

The analyses here cannot draw any final conclusions about the content, but serve as a guide to differences in the different knowledge bases. It also highlights the variety of common variant usage within both corpora, which if considering the mapping of more complex information within a corpus (such as trophic interactions) would be essential. In the pilot phase of the research, it was found that common variants were more frequently found in direct links to trophic interaction mentions (see Chapter 4). These numbers provide indications as to how to begin to evaluate how resources should be applied in automatic processing and to demonstrate how they can be compared against real bodies of text to compare their coverage.

Scientific nomenclature variant usage

The frequency and ranking of taxonomic terms in both corpora support the consistent use of Oncorhynchus mykiss as the accepted name in both academic and mixed articles. Both JEFF and WEB corpora indicate that authorship does not tend to be used widely in these contexts, which was confirmed in the discussion in the expert evaluation (see Chapter 7). In fact, the only two scientific nomenclature terms that occur more than 10 times in the either corpus are Oncorhynchus mykiss and Salmo gairdneri (240 and 57 (47.01 and 11.16 hits per million), and 809 and 195 (131.84 and 31.78 hits per million), respectively). The ratio of usage across both corpora are 80:20. The prevalence of Salmo gairdneri, despite it being, according to the ITIS resource, the incorrect spelling of an invalid synonym (Salmo gairdnerii), is much more frequent. The frequencies of Salmo gairdneri and Salmo gairdnerii in each corpus were 57 compared with 17 (11.16 compared with 0.2 hits per million) in the JEFF corpus, and 195 compared with 19 (31.78 compared with 0.81 hits per million) in the WEB corpus, respectively. This is a ratio, in both cases, of about 98:2, in favour of the incorrect spelling variant. Spelling variations, acceptability of terms and actual practice are also discussed in more detail in Chapter 7 in the external evaluation with domain experts.

Straight frequency information provides some insight into the profile of the terms used, the focus of the corpora and general usage, but to compare usage of these terms across the length of the corpora, dispersion analysis was performed, looking at the range of mentions across the different documents of the corpora. These are all described in further detail in the methodology (Chapter 3). The following graphs consider the comparative usage of specific scientific nomenclature variants, to see if their usage is spread across the corpora, or if links between specific terms can be identified using this analysis technique. The analysis focuses first on the JEFF corpus, before proceeding to analyse the WEB corpus.

Figure 9.55 shows the dispersion (range$_2$) of the two most common scientific variants for Oncorhynchus mykiss in the JEFF corpus. The graphs show that both appear scattered throughout the corpus. Oncorhynchus mykiss is considerably more extensive in usage (Oncorhynchus mykiss occurs in 43% of documents in the JEFF corpus, and Salmo gairdneri occurs in 7%). The scale of Figure 9.55 is such that it is not possible to see if these overlaps mean that the terms co-occur in same document, but the numbers show that 23 documents in the corpora have mention of both variants (53% of the occasions in which Salmo gairdneri is used). This implies that the terms are not used exclusively (solely in one context or another). As the dispersion measure applied only works down to document level granularity, these conclusions cannot be used to gain more specific insight into contextual co-occurrence, but concordances showed that the majority of times it appears in the JEFF corpus is in the references section. Where the names appear in the same document, usually the author uses Oncorhynchus mykiss in the main body of the text, with Salmo gairdneri appearing in the reference section.

The spread in the scientific nomenclature distribution is broadly similar in the WEB corpus (see Figure 9.56). Oncorhynchus mykiss appears in 246 or 23% of the documents, with Salmo gairdneri appearing in 79 or 7% of the documents. In this case they co-occur 63% of the time. Looking at concordances, while still used overwhelmingly in the references section, in the WEB corpus, it was found being referred to as the former term for Oncorhynchus mykiss twice in the main body of the text. This spelling does not even appear in the ITIS, whereas Salmo gairdnerii does appear as an invalid term (original name/combination) for the subspecies of Oncorhynchus mykiss, Oncorhynchus mykiss gairdnerii (Richardson, 1836). In the CoL it does exist as a synonym for Oncorhynchus mykiss, albeit including the authorship and in the VTO it is included as a related synonym as it is shown here. Spelling issues and reasons as to why an incorrect spelling of a former accepted variant may be more common than the correct spelling was further explored with domain experts in Chapter 7.

The consistency of reference usage provides further support for one, the consistency of usage in the scientific corpus, and two, the time-related nature of usage for these terms. The time-related nature of usage refers to how accepted variants change over time, as do preferred terms, and how this can be seen through the usage analysis performed here. To further explore this sort of usage profile, a time-delimited test corpus would be needed and these sorts of analyses would be possible in a semi-automated way. The same applies for looking, in an automatic way, at the place in the document that the terms appear. The corpora would need to include metadata as to the different sections of documents.

Common variant usage: frequent terms

In the analysis of common variant usage, there is a clear preference towards rainbow trout and steelhead trout name variants throughout both the JEFF (academic) and WEB (mixed) corpora (see Figures 9.57 and 9.58). Trout and brown trout were excluded from the analysis given their anomalous status as described in the previous section.

Figure 9.57 shows a preference in the JEFF corpus for steelhead over steelhead trout, whereas rainbow trout is preferred over rainbow. Overall rainbow trout was by far the preferred and most extensively used term. At this level of analysis the ambiguity remained as to whether the occurrences of rainbow were related to its meaning as a type of trout, on its own or because of factors such as word order in grouping terms such as “steelhead and rainbow trout”. While hits are spread across the corpus, it is important to note that there are certain spikes at specific points. Where there are spikes, the terms often co-occur in a document. However, the frequency is usually heavily weighted towards one or the other, which indicated that the terms, while being closely related in meaning, are not completely interchangeable as synonyms and that they are used in slightly different contexts. It also indicated that the documents in this corpus tend to have favour one or the other rather than using both terms equally. Concordances and relation network graphs are used later in this profiling study to provide a deeper analysis into these patterns of behaviour. It should be noted that what is important is the proportionality of co-occurrence of one or the other. The spikes themselves could just indicate longer documents, but considering the way the corpora were built, this was difficult to discern. This should be considered in future work.

Figure 9.57 and Table 6.9 show that all four terms are distributed across the JEFF corpus, without being focused on one specific area.

Table 6.9: JEFF corpus: comparative distribution of steelhead and rainbow trout variants

Distribution		%
Rainbow trout number of docs	160	27%
Rainbow no of docs	48	8%
Steelhead number of docs	71	12%
Steelhead trout number of docs	47	8%

Table 6.10: JEFF corpus: co-occurrence of different variants of rainbow and steelhead trout. No. of poss. Docs refers to the number of documents in which the terms could possibly co-occur.

Cooccurrence	Co-occur docs	No. of poss. Docs	% co-occurrence docs
All four in same document	4	47	9%
Steelhead trout and rainbow trout	31	47	66%
Steelhead and rainbow	8	48	17%
Steelhead trout and rainbow	7	48	15%
Rainbow trout and steelhead	35	71	49%
Rainbow trout and rainbow	23	48	48%
Steelhead and steelhead trout	24	47	51%

Table 6.10 provides information on co-occurrence between these variants. These suggest a linguistic reason for the variants, such as saying a combination of "steelhead and rainbow trout" and indicates a preference for this order (steelhead and rainbow trout co-occur in 49% of cases but rainbow and steelhead trout co-occur in only 15%). This can be used to extrapolate that there is some difference in meaning between the two, that they are not mutually exclusive and just different choices of the same term. They are also used about half the time alone in documents. Concordance lines were used to explore this further.

Dispersion in the WEB corpus differed somewhat, and you can start to see more differences that are likely to be related to the different content of the corpus (not purely academic) (See Figure 9.58).

Rainbow trout was also by far the preferred vernacular variant (appeared in 34% of the documents) in the WEB corpus. Rainbow was the second-most frequent (13% of the documents). The concordances again had to be checked (included in the Relation Network Graph section) to identify if these hits all referred to trout or if they were examples of rainbow in another sense of the word. Steelhead was also very frequent, appearing in 11% of the documents, more or less on a par with the JEFF corpus. Steelhead trout was again the least dispersed of these variants, appearing in 7% of the documents. As regards their comparative co-occurrence, the same pattern is seen in the WEB corpus to the JEFF corpus (see Table 6.11), although there are higher levels of co-occurrence of these terms. This could be to do with the sorts of documents in one corpus or another. Further empirical study was necessary to evaluate the synonymy of specificity of usage of these terms, which was performed through the relation network graphs and discussion with domain experts.

Table 6.11: WEB corpus: co-occurrence percentages of steelhead and rainbow trout variants

Cooccurrence	Co-occur docs	No. of poss. Docs	% co-occurrence docs
All four in same document	15	72	21%
Steelhead trout and rainbow trout	57	72	79%
Steelhead and rainbow	41	120	34%
Steelhead trout and rainbow	27	120	23%
Rainbow trout and steelhead	101	120	84%
Rainbow trout and rainbow	96	139	69%
Steelhead and steelhead trout	48	72	67%

Common variant usage: infrequent terms

Having looked at frequent scientific nomenclature and common name variants separately, dispersion was then used to see if any particular links or differences in the usage of infrequent common variants versus the nomenclature could be identified.

Figure 9.59 shows co-occurrence different infrequent vernacular and scientific variants in the JEFF corpus. A number of variants do not co-occur, but some do. The variants redband, redband trout co-occur a number of times with Oncorhynchus mykiss gairdneri in the JEFF corpus. This indicates that “redband” is a possible common name variant for this subspecies.

A quick google search would seem to support this, although this cannot be considered definitive evidence. The co-occurrence of Parasalmo mykiss and Kamchatka trout and Salmo mykiss was revealed here, and was explored further again in the relation network graph section. Otherwise there were no overlaps that would imply specific meanings. It should be noted that in these cases the frequencies are too low to draw any conclusions, although these indications are useful to indicate areas of interest for future exploration. The general topic of vernacular names and their links to specific scientific variants was discussed in Chapter 7.

In the WEB corpus, the dispersion of both vernacular and scientific infrequent variant types seems to reveal that various terms were only used in specific circumstances. As with the JEFF corpus, these terms are very infrequent so they can only be used to highlight areas for potential future research and methods by which to do this. The relation network graphs are used later in the chapter to explore hierarchy representation and specificity of contextual usage in these case of infrequent mentions.

Figure 9.60 shows co-occurrence of infrequent scientific and vernacular variants in the WEB corpus. This figure reveals that redband trout co-occurred in documents with various different taxonomic scientific terms, contrasting with the collocations seen in the JEFF corpus. It co-occurred on three occasions with Oncorhynchus mykiss irideus. This is interesting as it has wider dispersion than the other collocations. The other scientific nomenclature with which it co-occurred was Salmo whitei, Salmo gairdnerii and Oncorhynchus mykiss aguabonita. The relation network graph section explores whether this document-level co-occurrence was an indication of anything more interesting. To investigate any further information about this usage, it would be necessary to create corpora more specifically focused on these variants, perform google searches or speak with domain experts (or a combination).

The WEB corpus also revealed a unique link between Kamchatka steelhead and Parasalmo mykiss, which was also considered in more depth in the relation network graph section.

Summary of frequency and dispersion The fundamental points to take away from the frequency and dispersion analysis part of the profiling is that there is broadly consistent terminology usage across both corpora for this taxonomic entity.

The high numbers of common names across both corpora showed that even in scientific articles, scientists, at least in this context, use the common names of species to communicate, and that there is more variety in the usage of the common names than there is with the scientific nomenclature. It may be possible to use unaltered Word Sketches (as used for normal lexicographic purposes) of the contexts in which common names appear in contrast with their scientific nomenclature counterparts to reveal patterns in usage and if each tend to be used in different contexts or different parts of the article or web page. This provides some evidence as to the consistency of the use of scientific nomenclature, in both the JEFF and the WEB corpus.

The consistent ranking of the most common terms provide evidence as to the consistency of usage of the different variants across different corpora, despite one having been compiled from a web-scrape. Where there are differences with the knowledge resources or between each corpus is where much of the interest lies. The seemingly more prevalent usage of Salmo gairdneri was further explored with P4 in the outreach discussion (see Chapter 7). As regards specific terms, these analyses can be used to provide an overview by which to identify possible semantic/collocational preferences of some common names with other scientific nomenclature, which is useful when moving forward with analyses.

Relation network graphs

This section of the chapter applied the methods developed in Phase 0 and 1 of the research (see Chapter 4) to the practical application of nomenclature profiling. The analysis was used to identify where the data matched that in the various knowledge representation sources compared in the first section of this profile study, where there were gaps and where there were any ambiguities or disagreements,to use for further discussion.

Consistency/agreement In looking for consistency and agreement with the knowledge representation sources being studied, it is important to look at the wider picture and compare the representations with that of the resources. For this, a number of different filters were applied. In the following graphs, the size of the nodes is relative to their closeness centrality, as described in Chapter 4, and the arrows indicate the direction of the relation from source to target. The size of the arrows are relative to the frequency each relation appears in the corpus, so is therefore a demonstration of the strength of the relation in the given corpus.

When looking for similarities and coherence with the ontologies, it is important to look at the hierarchies forming in the graphs and how different terms link together. The unfiltered graph shows the links between Oncorhynchus mykiss, Salmo gairdneri and Salmo gairdnerii and rainbow trout and steelhead (trout). Only steelhead trout is linked to Salmo gairdnerii. Figure 9.61 show links between the aforementioned variants, having filtered out the anomalous brown trout and trout links to provide a clearer picture. The reasons for this are detailed in the Disagreements and Ambiguities section. This provides the fullest picture of the different links between taxonomic mentions relating to Oncorhynchus mykiss in this corpus. Salmonidae, the family to which Oncorhynchus mykiss belongs, is not linked to the rest of the genus nodes in the JEFF corpus graph. However, the graph does show the highly linked character of the two main common variants and the different scientific nomenclature variants.

The WEB corpus unfiltered image was less clear (see Figure 9.62). The Sketch Grammar used in the research does not filter for some of the "part of" relations because of issues with compiling. This had not caused issues until now, but in the WEB corpus looking specifically at this data unfiltered it produces quite a lot of noise. For this reason, the subsequent graphs for this corpus in the rest of the section have been filtered to remove relations that have only one hit. While this does remove some correct matches, it leaves a clearer picture which can be read and analysed.

Figure 9.62, clearly shows how Euteleostei links down though Salmonidae to rainbow trout to Oncorhynchus mykiss. The way that the common names usually sit above as the source part of the source-target pair in relations with the scientific nomenclature indicates their broader definition. The further hierarchy revealed in the WEB corpus could be because the corpus is made up of a wider range of documents. In scientific research articles a lot of information will be considered known, whereas in a web-scraped corpus it is more likely to have encyclopaedic or paedagogic material that will explicitly mention the taxonomic lineage of species. The ClearEarth project [205] mentioned this in their paper which talked about their training data in training the NLP algorithm.

In correcting the corpus for lower case, the only hierarchy that can be seen is through Salmonidae, and Euteleostei. When analysing the WEB corpus without having corrected for lower casing, the links through Salmonidae, salmoniformes and Euteleostei and Protacthopterygii are clearly seen (see Figure 9.63). As mentioned before, the lower-casing was performed to ensure that as many mentions as possible were correctly identified and counted. However, looking in more detail at the reason for the differences in the Word Sketches, it appears that lower-casing the whole corpus resulted in no terms being identified as Proper Nouns (PN). This resulted in changes to the profiling of some terms in the Word Sketches according to the patterns, and should be considered when processing corpora for analysis.

Chapter 4 described the concept of hubs identified within the graphs as a term disambiguation tool. In the JEFF corpus, according to the selection criteria of closeness centrality above 0.4, neighbourhood connectivity in the bottom third of the range and 3 or more edge count (see Chapter 4), Salmonidae, trout and steelhead classify as hubs (see Figure 9.64). In the WEB corpus, according to the same selection criteria, trout, steelhead, rainbow and Salmonidae all classify as hubs (see Figure 9.65). These are all terms which sit above other terms in the hierarchy and include others under them, which is a technique that can be used to work out meaning within these graphs. More will be said about trout in the subsequent sections.

Going back to the WEB corpus image filtered for 2 or more hits Figure 9.62, as with the JEFF corpus, links between rainbow trout and steelhead trout with both Salmo gairdneri and Oncorhynchus mykiss are revealed. Steelhead, actually more common than steelhead trout, is only linked to Oncorhynchus mykiss according to the graph when filtered at this level. In contrast with the JEFF corpus, Kamloops trout is also seen linked here. Salmo gairdnerii does not appear in the graph at all, given the lower frequency of occurrences of this name. An internet search of a number of different databases provides varied feedback, but the general consensus is that double “i” at the end of the term is the correct spelling for the former name, whereas the single “i” is an incorrect spelling for the former name. For this reason, identifying that the supposedly incorrect spelling is much more common in the literature is interesting. When looking at incidence of this term, it was identified that in both corpora the majority of instances appeared in the references section, which supports the fact that this name was previously an accepted name. The increased incidence of a variation which does not appear in the ITIS, which only includes the double “i” ending in its entry, was further explored in Chapter 7. Unfortunately the corpora do not have time stamps on each article, as this would have been useful to see if articles in which the term appears in the main body of the corpus were older. In one case, the context in which the term is found is as an explanation, describing Salmo gairdneri as the former term for Oncorhynchus mykiss (see Figure 9.66).

Looking at other frequent terms, “rainbow” and “bow”, it was important to be aware that they can have different meanings in context. Concordance analysis revealed that in the WEB corpus, the term “rainbow” appeared nearly exclusively in reference to a species of fish. The only cases in which it did not were references to a specific paper in which one of the authors’ surnames was Rainbow, which counted for only five of 407 hits in total. Rainbow often collocated other fish-related terms such as smelt. However, with bow, 14 of the 19 occurrences referred to something other than a fish, such as a river or the bow of a boat. The remaining were fish references, either using bow as an abbreviation it would seem to refer to rainbow trout, or because the rainbow was hyphenated. In the JEFF corpus, 207 (41.28 per million) instances of “rainbow” were found, all of which referring to the fish. What can be seen across both these corpora is how the adjective in the term of the vernacular is used in conjunction with lifestage or other words to describe the fish instead of overtly using the full name of “rainbow trout”, for example.

“Bow” was more common in the JEFF corpus than the WEB one, with 71 instances (14.16 per million). In this case the use of bow is split mainly between the name of a river and the mistakenly annotated rainbow trout as a result of a hyphen. It was interesting to see that the term rainbow does not come up very frequently in the context of smelt or other terms for lifestages in the JEFF corpus, in contrast with the WEB corpus.

Gaps Potential gaps in the knowledge resources were identified through this aspect of the analysis. In the case of Oncorhynchus mykiss, the potential gaps were related to potential collocational usage of terms in specific contexts, or related to specific nomenclature pairing with vernacular equivalents. While not all the specificities of the terms were explored in the external evaluation because of specific domain focuses, Chapter 7 goes some way into shedding light on all these issues.

The scientific nomenclature Parasalmo mykiss, which only appeared in one document across the JEFF corpus, was only linked to the common term Kamchatka steelhead (Figure 9.67). On further inspection, this usage appeared in the corpus in a references section of an article. All of the articles using this scientific nomenclature variant, and on two occasions the vernacular too, had been translated from Russian, indicating possible geographical/cultural variation in terminology usage. This is an interesting, although isolated incident in this corpus. Kamchatka is a place in Russia, which provides further weight to this idea, as it being used as a geographical descriptor of the common name. This can be seen in the graph as these are just two nodes joined to each other but separated from the rest of the graph.

The discussions in Chapter 7 did indicate that geographical variation was a common phenomenon in nomenclature usage. According to P4, the fish specialist, this geographical variation is a common issue. P4 described obstacles to effective communication resulting from different preferences as to scientific nomenclature usage as travelling round the globe. In the focus group the discussion focused more on specific authors or domains potentially favouring one term over another.

In the WEB corpus, Parasalmo mykiss is also separated from the rest of the terms in the graph without a filter (see Figure 9.68). Part of the reason for this is the infrequency of the mentions in each corpus. However, we can also see consistency in its link to the Kamchatka steelhead, which is useful to see the link between the use of these two terms specifically, and also provides information that does not at first glance seem to be available in any of the studied knowledge representation resources. This is an example of how corpus analysis can indicate where there may be specific collocational differences or specifics which are not included in the knowledge representation resources.

The link identified in the WEB corpus between salmon and Salmo gairdnerii is included in Figure 9.68. The concordance in the Sketch Engine shows that this was a link correctly identified by the method (see Figure 9.69) but this is incorrect according to the taxonomic resources. The concordance identified the origin of the statement as the URL frammandearter.se, a Swedish website. When I tried to access the original it was no longer available. This highlights the way the method can be used to track the location of statements, as well as patterns of correct and incorrect usage.

Ambiguities and disagreements Both JEFF and WEB corpora highlight the same disagreement as regards the inclusion of the term brown trout as included in the CoL as a synonym for Oncorhynchus mykiss. When looking at the corpus data, brown trout principally appears in the context of Salmo trutta, not of Oncorhynchus mykiss (with 191 hits in comparison with one in the JEFF corpus, and 176 versus one in the WEB corpus). That indicates an error on the part of CoL. The referenced work was “Fishes, fishing implements and methods of Nepal, Shrestha, J." and when searching on the internet seems to have links to the UN Food and Agriculture Organization, however no copy of the actual text could be accessed to check the link. In my test data, when looking at the concordance data for the one link in each case, in the JEFF corpus the instance in which brown trout is seen to be linked to Oncorhynchus mykiss appears to be a mistake in the writing, shown in Figure 9.70. The order of the scientific names in the piece of work would indicate that the author of the paper may have written them the wrong way round, referring to the knowledge sources used as a baseline in this piece of research.

In the WEB corpus, there was also one example of a link being identified between Oncorhynchus mykiss and brown trout. In this case it is a weakness of the methodology, as it incorrectly makes the link because the original text would have been in a table. The outcome is it finds a meaningful link where there is none. Figure 9.71 shows the concordance line for this.

However, there is just one example in the corpus, which is one of the reasons why frequency filtering can be used to eliminate spurious results such as this one (see Figures 9.72 and 9.73).

The same occurs in the WEB corpus. Figure 9.74 shows a graph of the relations in the WEB corpus filters for two hits or more and arrow thickness that relates directly with the number of hits for a visual clue of the strength of the relation. Trout and relations to trout have been selectively removed from this graph to make it easier to visualise the relations of interest. This shows that while there is a link between Oncorhynchus mykiss and brown trout, the link between brown trout and Salmo trutta is hundreds of times stronger. In fact, as has already been mentioned, the data shows that the link is a spurious one and can be filtered out (either in a blanket way by excluding uncommon hits or specifically if the researcher is interested in specific phenomena).

As regards ambiguity, in the CoL trout is included as a common name for the species Oncorhynchus mykiss. However, when we look at the graph, we can see that while it is not incorrect, trout is an umbrella term for many Salmonidae species, rather than being a term specifically linked to Oncorhynchus mykiss. Below are some graphs which highlight the hub nature of trout in both corpora, and how we can see that trout is an umbrella term that links out to other species of Salmonidae. Here is where corpus analysis can be used to identify the inclusion, exclusion and broader and narrower definitions of specific terms.

As was mentioned in the Consistencies section, the trout node forms a hub which supersedes the other common names referring to Oncorhynchus mykiss and other species level names, for which reason the issue is further explored in the ambiguity section. The CoL had categorised it as a synonym, but we can see in Figures 9.75 and 9.76 that the term is connected to far more than just terms within the category of Oncorhynchus mykiss: it seems to be used across many parts of the Salmonidae family. So this it an important disambiguation technique to bear in mind with this method.

6.2.3 Summary of findings

Nomenclature profiling study 1 focused on an example of scientific nomenclature that was frequent in the test corpora, and showed signed of having some, although limited, instability in the scientific nomenclature usage. Through the use of frequency and dispersion analyses, an initial profile was drawn up to identify possible areas of interest and to gain an overall view of the behaviour of the different scientific and vernacular variants throughout the corpus. Some findings from this analyses were:

- Oncorhynchus mykiss is the current accepted name and usage supports this
- Although Salmo gairdnerii (with a double "i") is considered by ITIS as the only "valid" prior variant for Oncorhynchus mykiss, the single "i" variant appears much more frequently across both corpora
- Salmo gairdneri had wider-spread use in the past (given its appearance in many publications, particularly in the reference sections)
- Rainbow, rainbow trout, steelhead and steelhead trout are all accepted and widely used vernacular variants of Oncorhynchus mykiss
- Co-occurrence of the above vernacular variants in the same documents indicate possible nuances of meaning between these variants, particularly the indication that they often are used in pairs
- Some other, infrequent common names occur in the same document as other, infrequent scientific nomenclature variants, which require further exploration

The profile then proceeded to a deeper analysis, using the techniques developed in previous stages of the research to transform Word Sketches, which take into account both grammatical and collocational behaviour. These were viewed in Cytoscape again to discuss their qualities. In the case of this specific profile study, the following was discerned:

- Brown trout is linked to the scientific species Salmo trutta, not Oncorhynchus mykiss
- Links are seen between Oncorhynchus mykiss and rainbow trout, steelhead, rainbow and steelhead trout
- WEB corpus shows more links between the higher echelons of the taxonomic hierarchy than the JEFF corpus

- Vernacular variants such as rainbow trout can be identified as hubs which disambiguate meaning

The graphs also show the JEFF and WEB corpora to be broadly consistent. They differ more in relation to coverage than presenting a completely contradictory image. It should be noted that the image produced by the JEFF corpus unfiltered was a much clearer image to work with than the unfiltered WEB graph.

The resource comparison revealed that there was a lot of variation in the approaches taken, with the ITIS taking a much more limited approach than either the CoL or the VTO. The CoL has a strong focus on vernacular variants and the VTO on grouping many different scientific variants to ensure broad coverage.

6.3 Nomenclature profiling study 2: Sander lucioperca

Sander lucioperca was chosen for the second nomenclature profiling study because it was the second most common nomenclature term which was identified across both corpora (calculating the frequency of the synonym variant plus the accepted name across both corpora). The nomenclature profile is much more limited in this case, because frequency of term variants relating to this species is much lower than in the case of Oncorhynchus mykiss. For this reason the profile serves as an example of how to profile rarer terms within a corpus. The low frequency of hits means that no generalisations can be made, but gives an indication of potential points of interest should a corpus which focuses more on this species be built and studied.

6.3.1 Sander lucioperca in the taxonomic resources

Comparison of the accepted names The same differences and similarities between the chosen resources as in the previous study can be found. Both ITIS and CoL agree and present exactly the same accepted name, Sander lucioperca (Linnaeus, 1758). The VTO once again presents the accepted name without any authorship, Sander lucioperca. It includes Sander lucioperca (Linnaeus, 1758) as a related synonym entry.

Comparison of the synonyms Less variation and more consistency was found in comparison with the previous profiling study. There are 16 scientific nomenclature variants in total, with three vernacular variants. There are four exact matches (two scientific, two vernacular) and three no matches (two scientific, one vernacular). There are then 12 partial matches as regards scientific nomenclature, which follow the same basis as in the previous study, in which the CoL presented a name with the authorship, whereas the VTO includes the entry without this information.

The exact matches in the resources are between ITIS and CoL. One of the matches is the accepted name itself, and the other Stizostedion lucioperca (Linnaeus, 1758). ITIS categorises this synonym as “invalid – subsequent name/combination” and CoL as a straight synonym. The CoL includes two scientific variants which have no equivalent in the other resources, which supports the assertion that this resource is more inclusive than the other two. The vernacular variants include: pikeperch (CoL), pike-perch (CoL and VTO) and zander (CoL and ITIS). A full list of both the scientific and vernacular variants is included in Appendix E.

6.3.2 Sander lucioperca in the JEFF and WEB corpora

The JEFF corpus covers 19% of the scientific nomenclature variation of the three resources (3 of 16) and the WEB corpus includes 25% (4 of 16). All three different vernacular variants appear in each corpus. The comparative weighting of scientific variants to vernacular variants is approximately the same 20:80 ratio as in the previous profile.

Table 6.12: Frequency comparison and ranking of taxonomic mentions across JEFF and WEB corpora

Resource	Name	Frequency in JEFF	Rank in JEFF corpus	Frequency in WEB	Rank in Web corpus
CoL	pikeperch	290	1	282	1
CoL/ITIS	zander	12	5	189	2
VTO	Sander lucioperca	34	3	105	3
CoL/VTO	pike-perch	16	4	105	3
VTO	Stizostedion lucioperca	50	2	60	5
VTO	Lucioperca lucioperca	1	6	4	6
VTO	Lucioperca sandra	0	7	1	7
CoL/ITIS/VTO	Sander lucioperca (Linnaeus, 1758)	0	7	0	8

Table 6.12 shows that the ranking for the WEB corpus follows what might be expected given the findings in the previous profiling study: vernacular variants and a truncated version of the accepted name as used in the VTO, without the authorship, most frequent, followed by less preferred terms. There are, however, more raw hits of the synonym Stizostedion lucioperca in the JEFF corpus.

Table 6.13 and Figure 9.77 set out the relative frequency of Sander lucioperca variants normalised for frequency per million. These show that the frequency for Stizostedion lucioperca has the same weight in the JEFF corpus to the WEB corpus, whereas Sander lucioperca has about half the weight frequency-wise in the JEFF corpus in comparison with the WEB corpus.

Table 6.13: Sander lucioperca frequency comparison of nomenclature terms (frequency of hits per million)

Resource	Name variant	JEFF corpus (freq per mill)	Rank in JEFF corpus	WEB corpus (freq per mill)	Rank in Web corpus
CoL	pikeperch	56.71	1	45.87	1
CoL/ITIS	zander	2.35	5	30.74	2
VTO	Sander lucioperca	6.65	3	17.08	3
CoL/VTO	pike-perch	3.13	4	17.08	3
VTO	Stizostedion lucioperca	9.78	2	9.76	5
VTO	Lucioperca lucioperca	0.2	6	0.65	6
VTO	Lucioperca sandra	0	7	0.16	7
CoL/ITIS/VTO	Sander lucioperca (Linnaeus, 1758)	0	7	0	8

The ratio of one variant to another in the two corpora gives a ratio of 40:60 for Sander lucioperca to Stizostedion lucioperca in the JEFF corpus, and the WEB corpus demonstrates the reverse, approximately 60:40 in favour of the former variant. The ratio has been calculated using the raw frequencies. It is also interesting to note that the synonym, Stizostedion lucioperca ranks 5th in the WEB corpus according to the normalised frequency ranking but 2nd in the JEFF corpus.

The numbers were very small in this profiling study and therefore impeded generalisation, but one possibility for the disparity in frequency/ranking was that it was the result of hits appearing concentrated in one document. However, looking at the dispersion graphs for both corpora, this was not the case (see Figure 9.78 for the JEFF corpus dispersion and Figure 9.79 for the WEB corpus dispersion). These graphs show dispersion across the documents where one or other variant appears, to make the graph easier to read. Stizostedion lucioperca appears in 29 or 3% of the documents in the WEB corpus and 23 or 4% of the documents in the JEFF corpus, giving it a similar dispersion across both corpora - infrequent but dispersed across a number of documents. Not knowing the length of the documents, as mentioned before, this exercise serves to look at the comparative dispersion of the different variants in each corpus, and co-occurrence. A similar pattern to that found in the previous profile for synonyms was found: the majority of hits were in references sections (as verified by concordance searches). In this case it was noted that the term was found in the context of non-English language documents, such as Spanish (in the JEFF corpus) and German (in the WEB corpus). The numbers are too small to draw specific conclusions about these tendencies, but could indicate a geographical as well as a time element to name preference. To test this, corpora could be made which compare tendencies across languages with specific scientific nomenclature. The accepted name, Sander lucioperca is consistently found more in the main bodies of articles in both corpora.

Pikeperch was the preferred vernacular variant in the JEFF corpus, but the WEB corpus showed more variety of usage. The dispersion clearly show that the different common names in the WEB corpus are dispersed similarly in accordance with their frequency (pikeperch across 43 documents, or 4% of the corpus, zander across 27 or 3% of the corpus, and pike-perch across 22 or 2% of the corpus), whereas pikeperch is dispersed across 43 documents or 7% of the JEFF corpus, in comparison with 1 and 2% for pike-perch and zander, respectively.

Table 6.14: JEFF corpus: co-occurrence of different scientific and common variants for Sander lucioperca

Co-occurrence	**No of docs.**	**% co-occurrence**
Sander lucioperca and Stizostedion lucioperca	5	23%
Sander lucioperca and pikeperch	16	73%
Sander lucioperca and pike-perch	3	43%
Sander lucioperca and zander	4	33%
Stizostedion lucioperca and pikeperch	19	83%
Stizostedion lucioperca and pike-perch	4	57%
Stizostedion lucioperca and zander	9	75%

Table 6.15: WEB corpus: co-occurrence of different scientific and common variants for Sander lucioperca

Co-occurrence	**No of docs.**	**% Co-occurrence**
Sander lucioperca and Stizostedion lucioperca	10	34%
Sander lucioperca and pikeperch	22	55%
Sander lucioperca and pike-perch	15	68%
Sander lucioperca and zander	12	44%
Stizostedion lucioperca and pikeperch	18	62%
Stizostedion lucioperca and pike-perch	10	45%
Stizostedion lucioperca and zander	11	41%

The numbers of co-occurrence in Tables 6.14 and 6.15, show us that, as suggested by the graphs, in effect while vernacular variants have higher percentages of co-occurrence with one of the scientific variants, there does not seem to be a particular link between any scientific variant and a vernacular variant. The dispersions do show that at least in these two corpora that there is a tendency towards pikeperch in scientific writing and that the contexts in which Sander lucioperca is used would indicate that it is the preferred term. The limited number of hits for the variants in this entry mean that to further explore this issue, new corpora with documents that focus on this fish family would enable us to better to investigate the way the terms - larger frequencies would provide the basis for more reliable results.

Relation network graphs

The relation network graphs were much clearer in this nomenclature profile, given the more limited number of elements (nodes) and frequency of relations. This therefore meant that it was unnecessary to perform so many layers of filtering. The results here can serve as a profile for these corpora as regards the nomenclature terms used to describe this species. Figure 9.80 refers to the JEFF corpus and 9.83 to the WEB corpus.

Consistency/agreement The graphs for each corpus show Percidae as a hub (closeness centrality of 0.5-1, neighbourhood connectivity in the bottom third of the range and an edge count of over three). In both graphs the hierarchy of Percidae agrees with that of the knowledge representation resources.

Pikeperch is also identified as a hub in both corpora, in which Sander lucioperca and Stizostedion lucioperca are identified as children of this term. In the WEB corpus, Sander volgensis is also included, which will be talked about in the gaps section. In the WEB corpus zander as a vernacular variant is also linked to both Stizostedion lucioperca and Sander lucioperca, but with stronger links to the latter, the accepted name.

Gaps A possible gap was identified in the WEB corpus, as regards the Sander volgensis node (see Figure 9.83), which from the graph appeared to also be linked to the term pikeperch. This constitutes an example of overlap in common variant usage across different species within the same family, and indicates broader meaning within common variants. It also indicates where ambiguities may lie. The concordance shows that there are further qualifying terms to separate the two pikeperch. Further work could go into investigating overlaps in terminological meaning particularly regarding common variants (See Figure 9.84).

Ambiguities and disagreements In the JEFF corpus, as seen in relation graph, pikeperch was also linked to the Gymnocephalus cernuus and gudgeon. However, when investigating the concordances, this was an error of the method (see Figure 9.81 and 9.82). This demonstrates the issue when working with very low frequency terms, as each of these only had one hit.

6.3.3 Summary of findings

Nomenclature profiling study 2 focused on an infrequent taxonomic entity with signs of synonymy usage in the corpus. The infrequency of hits in the study make it difficult to draw definite conclusions, but a study like this could be used as an indication of potential focuses for investigation in future work with corpora compiled to focus on the nomenclature term in question. The infrequent number of hits make it more difficult to use the graph visualisations and filter techniques to remove spurious hits. However, concordances are a useful tool here because the low numbers of hits makes it easy to check each individual instance manually. Only some very general observations can be made in this case, but they may be worthy of further exploration with a focused corpus.

- It is possible that there is some disagreement as to the accepted name in the case of Sander lucioperca
- The JEFF corpus was more consistent in vernacular variant usage whereas the WEB corpus showed more variability
- In general the resources showed fewer variant options for this species
- The Sander volgensis case in the WEB corpus highlights the possibility of overlap between vernacular variant usage and multiple species

6.4 Nomenclature profiling study 3: Salmo trutta

Salmo trutta was chosen for the third nomenclature profiling study because of the frequency with which it appeared in the corpus. Unlike the other two profile studies, this term was not identified as having a linked scientific name synonym, despite the taxonomic resources recording very hight numbers of scientific variants. While it was expected that the high frequency of the term would yield results, it was thought that the relative absence of scientific synonym usage may result in a less interesting profile. This third study is an example of an exploratory profiling the nomenclature terms for a species without any indication of specific usage, in the case of a variant which occurs frequently.

6.4.1 Salmo trutta in the taxonomic resources

Comparison of the accepted names

The accepted names in the respective resources have the same qualities as in the previous studies: both ITIS and CoL agree and present exactly the same accepted name, Salmo trutta (Linnaeus, 1758) and the VTO once again presents the accepted name without authorship, Salmo trutta. In this case the name with authorship does not appear as a synonym in the VTO.

In the CoL there were also three examples of Salmo trutta (non-Linneaus 1758), in which it was classified as a misapplied name. These were:

- Salmo macrostigma (Duméril, 1858)
- Salmo pallaryi Pellegrin, 1924
- Salmo tigridis Turan, Kottelat & Bektaş, 2011

In each case, it would appear that taxa, originally all thought to be within the taxon concept of Salmo trutta, have later been identified under these new names. However, this distinction was not highlighted in the other resources, nor were there examples of these nomenclature in the corpora.

Comparison of the synonyms

Despite the apparent absence of hits for scientific variants in the corpus data, the resources demonstrated a wide range of variants, plus a large amount of variation between the resources as regards the scientific variants, similar to that in the Oncorhynchus mykiss study. There are 143 variants in total, with 116 scientific nomenclature variants and 27 common names variants. There are three exact matches (1 scientific (the accepted name for the species), 2 common variants) and 58 no matches (33 scientific, 25 common). There are then 81 partial matches as regards scientific nomenclature, which follow the same basis as in the previous study, in which the CoL presented a name with the authorship, whereas the VTO includes the entry without this information. The common variants which occur across more than one resource are: brown trout (CoL and ITIS) and sea trout (CoL and VTO). A full list is in Appendix E.

6.4.2 Salmo trutta in the JEFF and WEB corpora

Frequency and incidence with ITIS, VTO and CoL

Again similar terms occupied the top ranking in both corpora, although Salmo trutta in the JEFF corpus ranks 3$^{\text{th}}$, whereas in the WEB corpus it ranks 4$^{\text{th}}$.

As with the other studies, the WEB corpus has a wider coverage. The JEFF corpus includes 8% of the variants (12 of 142) and the WEB corpus includes 15% (22 of 142). As regards scientific nomenclature, the JEFF corpus only includes examples of Salmo trutta, the accepted name, whereas the WEB corpus has examples of 3 names other than the accepted name. These are very infrequent. The WEB corpus also exhibits more variety in the usage of common variants, with 18 different variants being identified in comparison with 11 in the JEFF corpus.

Table 6.16: Table looking at Salmo trutta variant name mentions across JEFF and WEB corpora

Resource	Name variant	Frequency in JEFF corpus	Ranking JEFF corpus	Frequency in WEB corpus	Ranking WEB corpus
CoL	trout	5801	1	7843	1
CoL/ITIS	brown trout	3274	2	3159	2
CoL	lake trout	527	5	1412	3
VTO	Salmo trutta	1212	3	1240	4
CoL	brook trout	1181	4	832	5
CoL/VTO	sea trout	317	6	758	6
CoL	whiting	22	9	49	7
CoL	sea-trout	28	8	24	8
CoL	salmon trout	7	10	21	9
CoL	finnock	0	13	18	10
CoL	sewin	0	13	17	11
CoL	brownie	0	13	11	12
CoL	whitling	0	13	11	12
CoL	loch leven trout	0	13	8	14
CoL	river trout	4	11	7	15
VTO	Salmo fario	0	13	7	15
CoL	peal	1	12	6	17
VTO	Salmo lacustris	0	13	2	18
CoL	blacktail	33	7	1	19
CoL	herling	0	13	1	19
CoL	orange fin	0	13	1	19
VTO	Salmo levenensis	0	13	1	19

Table 6.17: JEFF and WEB corpus frequency comparison (per million words)

Resource	Name variant	JEFF corpus (freq per mill)	Ranking JEFF corpus	WEB corpus (freq per mill)	Ranking WEB corpus
CoL	trout	1136.48	1	1278.19	1
CoL/ITIS	brown trout	641.41	2	514.83	2
CoL	lake trout	103.25	5	230.12	3
VTO	Salmo trutta	237.44	3	202.09	4
CoL	brook trout	231.37	4	135.59	5
CoL/VTO	sea trout	62.1	6	123.53	6
CoL	whiting	4.31	9	7.99	7
CoL	sea-trout	5.49	8	3.91	8
CoL	salmon trout	1.37	10	3.42	9
CoL	finnock	0	13	2.93	10
CoL	sewin	0	13	2.77	11
CoL	brownie	0	13	1.79	12
CoL	whitling	0	13	1.79	12
CoL	Loch leven trout	0	13	1.3	14
CoL	river trout	0.78	11	1.14	15
VTO	Salmo fario	0	13	1.14	15
CoL	peal	0.2	12	0.98	17
VTO	Salmo lacustris	0	13	0.33	18
CoL	blacktail	6.47	7	0.16	19
CoL	herling	0	13	0.16	19
CoL	orange fin	0	13	0.16	19
VTO	Salmo levenensis	0	13	0.16	19

Greater ambiguity in the WEB corpus could be indicated by this wider variety of variant usage. In addition, the JEFF corpus seems to be more consistent in its usage of brown trout and Salmo trutta. The dispersion data and relation network graphs later in the section provide further analysis in the relation between the two terms.

Table 6.18: Ratio of scientific nomenclature use to common variant use in the Salmo trutta profile

	SCI	**COM**	**% SCI**	**% COM**
JEFF	1212	5394	18%	82%
WEB	1250	6336	16%	84%

The ratio of scientific nomenclature to common variants is even more marked in this profile, as Table 6.18 demonstrates.

Table 6.19: JEFF and WEB corpora: variant coverage across all knowledge resources of Salmo trutta

	Total	**% coverage**
No of variants: VTO, CoL, ITIS	143	
Number of variants (SCI)	116	
Number of variants (COM)	27	
JEFF total	12	8%
WEB total	22	15%
JEFF number SCI	1	1%
WEB number SCI	4	3%
JEFF number COM	11	41%
WEB number COM	18	67%

Both corpora demonstrate a much greater coverage for common name variants, but while in the case of Oncorhynchus mykiss there was quite high coverage of scientific variants, here is it very low. The lack of authorship means that there are no matches in either the CoL or the ITIS, which elicits the same response as for the other case studies. What is perhaps the most surprising again is how widely used various common names are across both corpora, with the limited use of different scientific nomenclature variants, even accounting for the authorship. This provides good support for consistent nomenclature usage in these cases, but highlights potential ambiguities in vernacular variant usage.

Table 6.20: JEFF and WEB corpora: VTO coverage of Salmo trutta

	Total	**% coverage**
No of variants: VTO	51	
Number of variants (SCI)	50	
Number of variants (COM)	1	
JEFF SCI match with VTO	1	2%
JEFF COM match with VTO	1	100%
WEB SCI match with VTO	4	8%
WEB COM match with VTO	1	100%

Table 6.21: JEFF and WEB corpora: CoL coverage of Salmo trutta

	Total	**% coverage**
No of variants: CoL	92	
Number of variants (SCI)	65	
Number of variants (COM)	27	
JEFF SCI match with CoL	0	0
JEFF COM match with CoL	11	41%
WEB SCI match with CoL	0	0%
WEB COM match with CoL	18	67%

Table 6.22: JEFF and WEB corpora: ITIS coverage of Salmo trutta

	Total	**% coverage**
No of variants: ITIS	2	
Number of variants (SCI)	1	
Number of variants (COM)	1	
JEFF SCI match with ITIS	0	0
JEFF COM match with ITIS	1	100%
WEB SCI match with ITIS	0	0%
WEB COM match with ITIS	1	100%

Scientific nomenclature variant usage

As Salmo trutta was the only scientific nomenclature variant to be used in the JEFF corpus, dispersion can only confirm whether the term is well-dispersed through the corpus, which it appears to be. Salmo trutta is also very strongly preferred in the WEB corpus, so dispersion is not particularly useful here either. The dispersion of the scientific variants used in each corpus are shown in the dispersion graphs in Figures 9.85 and 9.86 for the JEFF and WEB corpora, respectively. Salmo trutta seems to be the most consistently used scientific term of the ones chosen in the profiling studies, although the conversation with P4 in the evaluation also indicates that the usage of this scientific term may be more ambiguous than would appear at first glance (see Chapter 7).

Common variant usage

The frequency dispersion profiling started to show some more interesting results in the comparison of common variant usage. In both the JEFF and the WEB corpus, particularly in the cases of the most frequent common names, there was a lot of co-occurrence between one or other variant, particularly brown trout with other variants. Tables 6.23 and 6.24 show the relative dispersion and co-occurrence figures for the JEFF corpus, whereas Tables 6.25 and 6.26 show the same figures for the WEB corpus. Figures 9.87 and 9.88 show this information in a graphical way for the JEFF and WEB corpora, respectively. This is at document level, so it may not mean a direct relation between the two terms, but it indicates a possible area for investigation. The terms that mainly coincided with brown trout were sea/brook/lake trout, which suggested that the variations could be reflective of their habitat.

Table 6.23: JEFF corpus: dispersion of single common name variants. No. of docs represents the documents in which the variant appears, and % of docs the percentage this represents of the total number of documents in the corpus.

Single variant dispersion	**No of docs**	**% of docs**
Brown trout	270	45%
Lake trout	49	8%
Brook trout	121	20%
Sea trout	50	8%
Sea-trout	13	2%

Dispersion data was used to identify co-occurrence of these common names with Salmo trutta, to check whether it was likely that all these common names were used as synonyms with Salmo trutta. Here a common pattern in the co-occurrence rates is observed. Both the JEFF corpus (see Table 6.27) and WEB corpus (see Table 6.28) show similar patterns of terminology

Table 6.24: JEFF corpus: co-occurrence of common variants between brown trout and other variants. No of docs represents the number of documents in which the variants to co-occur, then % of docs refers to the percentage of the total possible documents in which they could co-occur, referring back to the previous table.

Comparative dispersion	**No of docs**	**% of docs**
Co-occurrence of all five variants	0	
Co-occurrence of brown trout and lake trout	24	49%
Co-occurrence of brown trout and brook trout	99	82%
Co-occurrence of brown trout and sea_trout	45	90%
Co-occurrence of brown trout and sea-trout	13	100%

Table 6.25: WEB corpus: dispersion of single common name variants

Single variant dispersion	**No of docs**	**% of docs**
Brown trout	355	33%
Lake trout	166	15%
Brook trout	179	17%
Sea trout	99	9%
Sea-trout	11	1%

Table 6.26: WEB corpus: co-occurrence of common variants between brown trout and other variants

Comparative dispersion	**No of docs**	**% of docs**
Co-occurrence of all five variants	0	
Co-occurrence of brown trout and lake trout	96	58%
Co-occurrence of brown trout and brook trout	133	74%
Co-occurrence of brown trout and sea_trout	81	82%
Co-occurrence of brown trout and sea-trout	10	91%

usage. Brown trout, sea trout and sea-trout co-occur the most, although in the JEFF corpus this is most pronounced with the sea trout variants, both of which nearly exclusively occur in documents in which Salmo trutta also appears. The lower percentage of co-occurrence of lake trout and brook trout indicate that these terms could be used in other contexts to not describe Salmo trutta. This could suggest that sea-trout and sea trout are more closely collocated with Salmo trutta and brown trout, and that the other variants are more ambiguous in their usage, which was supported by the later conversation with P4 (see Chapter 7). The lower percentages all round in the WEB corpus could be the result of an increased emphasis of non-academic texts which mean that common names are used more frequently instead of their scientific variants, or that the common names are used more in the context of other scientific names. This phenomena was further explored through the relation network graphs.

Table 6.27: JEFF corpus: co-occurrence of various common name variants with Salmo trutta

Comparative dispersion	**No of docs**	**% of docs**
Co-occurrence of brown trout with Salmo trutta	233	86%
Co-occurrence of lake trout with Salmo trutta	22	45%
Co-occurrence of brook trout and Salmo trutta	90	74%
Co-occurrence of sea trout and Salmo trutta	48	96%
Co-occurrence of sea-trout and Salmo trutta	13	100%

Table 6.28: WEB corpus: co-occurrence frequent common variants with Salmo trutta

Comparative dispersion	**No of docs**	**% of docs**
Co-occurrence of brown trout with Salmo trutta	270	89%
Co-occurrence of lake trout with Salmo trutta	71	43%
Co-occurrence of brook trout and Salmo trutta	104	58%
Co-occurrence of sea trout and Salmo trutta	82	83%
Co-occurrence of sea-trout and Salmo trutta	9	82%

More infrequent vernacular variants also did not seem to exclusively appear alongside Salmo trutta in the WEB corpus. Table 6.29 shows the number of documents more infrequent vernacular variants appeared throughout the WEB corpus, whereas Table 6.30 shows the percentage of times these terms co-occurred with Salmo trutta in the same document. This adds to the argument of possible ambiguities in the usage and classification of vernacular variants, which is explored further in the next section.

Table 6.29: WEB corpus: document distribution of infrequent common name variants against Salmo trutta

Name variant	**No of docs**
Salmo trutta	303
Loch leven trout	4
Peal	5
Herling	1
Finnock	9
Brownies	10
Orange fin	1
Sewin	8
River trout	5
Salmon trout	13
Whiting	29
Whitling	4
Blacktails	1

Table 6.30: WEB corpus: co-occurrence dispersion of infrequent common variants with Salmo trutta

Co-occurrence of name with Salmo trutta	**No of docs**	**%**
Loch leven trout	1	25%
Peal	2	40%
Herling	1	100%
Finnock	5	56%
Brownies	1	10%
Orange fin	0	0%
Sewin	1	13%
River trout	1	20%
Salmon trout	2	15%
Whiting	5	17%
Whitling	3	75%
Blacktails	0	0%

Relation network graphs

The relation network graphs for Salmo trutta were used in this case particularly to identify whether the vernacular variants identified as linked with Salmo trutta were also used to refer to other species within the Salmonidae family.

As with the other profiling studies, the size of the arrows relates to the strength of the relation (number of hits, or if mentioned otherwise, the salience of the relation). Where the graphs are filtered it is specified. The size of the nodes refer to its closeness centrality. Consistency, gaps and disagreements relate to the validation of the VTO ontology aspect of the profiling.

Consistency/agreement Figure 9.89 is filtered for salience 9.5 and provides a clear picture of the relations identified in the JEFF corpus. This picture clears further when filtered for only 2 or more hits (see Figure 9.90), although some of the interesting links are removed in this interaction. Brown trout is by far the most strongly related term (by frequency) to Salmo trutta, but that also sea trout is related as a broader term (parent). Interestingly, brook trout is linked as a more specific term to both Salmo trutta and Salvelinus namaycush in the graph. It should be noted that each of these relations occur only once in the corpus. When the concordances were accessed brook trout was logically linked to Salvelinus fontinalis. In contrast, lake trout was linked to Salvelinus namaycush. In the graph lake trout is not linked to Salmo trutta at all, and appears linked to Salvelinus namaycush. These results plus a more general commentary regarding vernacular name use were explored with experts in the expert evaluation (see Chapter 7).

The WEB corpus presents a very similar picture to that seen in the JEFF corpus (see Figure 9.91). In this case the WEB corpus was filtered for 2 or more hits and salience over 11, because this gave a clear picture that maintained most of the relations of interest. Trout mentions were removed as in the Oncorhynchus mykiss profiling study.

In the WEB corpus, as in the JEFF corpus, brown trout, sea trout and Salmo trutta are all linked. It is also interesting to note how the other common variants appear: lake trout and brook trout will be discussed further in a later section, but brownie, loch leven and sewin all appear linked to brown trout. As explored in the Oncoryhnchus mykiss profiling study, the hierarchy through taxonomic ranking is clearer in the WEB corpus: Euteoleostei and Salmonidae is identified through Salmo salar.

Gaps An interesting piece of information that appeared in the WEB corpus is the link between Salmo trutta and sea trout, with Salmo trutta as the parent of the relation (indicating that Salmo trutta is a more general term). The concordance shows that sea trout is a term used specifically for the anadromous form of the species (terminology-wise).

This was particularly interesting because in the JEFF corpus the sea trout came out as the parent of Salmo trutta, which is how vernacular variants tended to identify in the graphs. In the JEFF corpus the concordances identified it being used as the vernacular specifically of Salmo trutta, without mentioning the anadromous form. This is important when thinking about the relativity of meaning and how the framing of a word can change where in a hierarchy, or knowledge representation model, an object may sit. The meaning of sea trout and possible interpretations was discussed with P4 in Chapter 7, in which P4 agreed that it could be considered a child of Salmo trutta because it is only a part of the Salmo trutta population.

In the WEB corpus the term sewin appears to have the same meaning as sea trout, if we look at the concordance line for the hits (see Figure 9.92). This is a contextual thing. This is interesting to comment on, because it means that the terms themselves, although presented as synonyms to a scientific taxon label, may only be used in specific contexts such as this one. This was not identified in the JEFF corpus.

Disagreements and ambiguities In both the JEFF and WEB corpora, as observed in the first nomenclature study profile, trout is a more general term than specifically Salmo trutta. There is no need to repeat the information here as it is explained in the previous study.

Looking back at both Figures 9.90 and 9.91 we can see that lake trout and brook trout, in contrast with the information given in the CoL, appear to be linked to Salvelinus namaycush and Salvelinus fontinalis, respectively, but not Salmo trutta. In the JEFF corpus the link between brook trout and Salvelinus fontinalis is only seen in the corpus that has not been lower-cased (see Figure 9.93). The concordances confirmed that these were legitimate links, as did a cursory Google search, although please see Chapter 7 for a full discussion on the subject with experts.

In the WEB corpus, Salvelinus alpinus appears linked to both brown trout and Salmo trutta, but looking at the concordances, these links were misidentified from tables that have been converted in the text files, and so have not been properly processed by the Word Sketches. The lack of grammatical context provided in single words in tables undermines the Word Sketches here.

6.4.3 Summary of findings

Nomenclature profiling study 3 focused on an example of scientific nomenclature that was frequent in the test corpora, but which seemed to demonstrate more singular scientific nomenclature usage. Through the use of frequency and dispersion analyses, an initial profile was drawn up to identify possible areas of interest and to gain an overall view of the behaviour of the different scientific and vernacular variants throughout the corpus. These findings were then further explored through the relation network graphs to see what relations between terms were identified and confirmed or disproved. Some findings from this analyses were:

- Salmo trutta is the current accepted name
- There seem to be some semantic differences between the different common variants that do indeed appear to be linked to Salmo trutta as synonym
- Some of the common variants listed in the CoL seem to be linked to scientific nomenclature/taxonomic entities rather than Salmo trutta
- Some common variants seem to have, on the one hand, a more specific meaning, but their distribution in areas in which Salmo trutta may not appear could mean that, on the other hand, they are also more general terms (at once only referring to anadromous forms, but also being used to refer to multiple different taxa, for example)
- The discussion with P4 revealed that these species are not truly anadromous because they have no reason for going so sea such as migration, but that the common names reflect the way they physically change to go to sea.

6.5 Guidelines

As an outcome of these nomenclature profiling studies and the work preceding the application throughout this thesis, I developed some guidelines for anyone who may want to apply the methods described herein. They are split up into three sections: the first contains general method guidelines. This section can be split into corpus creation and processing, resource data identification and preparation and the analysis techniques employed in the profiling analysis. The second looks specifically at data representation (filter and framing) and how this and the different analysis techniques are used draw different information from the data. The third looks more specifically at specific profiles and what sort of patterns to expect in the case of studying a taxonomic entity of certain characteristics.

1: Method guidelines

Corpus creation and selection

- Choose a subject matter/species/family on which to focus your corpora
- Identify suitable data and collate following copyright regulations

- Process suitably according to the data representation considerations below to be able to analyse to required granularity (i.e. with metadata for publication date, author, publication type, etc.)
- For more detailed dispersion graphs ensure that document length is available for each document in the corpus

Resource identification and application

- Choose a/multiple suitable resources according to analysis requirements
- Pull a total list of names to tag the test corpora
- Perform an analysis of the different resources to provide a comparison if using multiple resources

Data preparation

- Build tagged corpora using adapted Sketch Grammar
- Consider lower-casing and other pre-processing steps with the following considerations:
    1. Lower-casing will improve the number of hits (as well as other pre-processing steps such as deleting extra spaces, removing line breaks, etc.)
    2. Lower-casing does impact on Word Sketches because of altered tagging but does not seem to have had a great impact in this case (further work)
- Use script to pull Word Sketches for all the names in the list from the chosen resources
- Transform Word Sketches into edge lists for manipulation

Dispersion and frequency analysis

- Perform dispersion and frequency corpus analysis techniques on the data
- Use raw and normalised frequency and ranking to make intra- and inter-corporal comparisons of name usage
- Intra-corporal ranking will show the preferred terms within a corpus and inter-corporal ranking will demonstrate if these preferences are stable or not across the different test corpora

- Use raw and normalised frequency to evaluate the weight of focus of these particular species or name variants in the respective corpora
- Use dispersion analysis to evaluate intra-corporal synonymity of different name variants by the levels of co-occurrence at the necessary levels of granularity (scientific variant co-occurring with single or multiple common names, multiple common names co-occurring or appearing alone in different contexts)
- Compare dispersion analysis results from each corpus to evaluate stability or lack thereof across the corpora

Network graph manipulation and analysis techniques The relation network graph analysis was a key part of the analysis and included the identification of characteristics in the graph that could be used for profiling and meaning disambiguation. During the research I also identified that in certain cases the classification of specific nodes (entities) or edges between nodes (relations) could be ascertained through the shape of part of the graph itself. These formations have been described as “hubs”. The criteria for identifying as a hub is set out in Table 8.1.

Table 6.31: Identification of node characteristics through hubs and graph positioning

Node identification	**Meaning**	**Node description**	**NC**	**CC**	**Edge count**
species (except parr)	Classification (as species); disambiguation (many species level names also genus level in other context)	Hub - outer node	High (top fifth or top half of range)	Low (0 - 0.04)	Under 3
genus; common; general	Identification of central nodes of hubs - classification level.	Hub - central node	Low (bottom third of range)	High (0.5-1)	3 or over
common; general (collective and life-stage); family; order	Linking node between different parts of the graph (classification)	Link node (not very well defined)	High (top fifth)	High (0.5-1)	3 or over

- Filter for frequency to eliminate potentially spurious results
- Filter for salience to then fine-tune some results
- Use selective filtering upon identification of a spurious result or an overly general result that obscures other results
- Use hubs to identify variants which serve to group other variants under their meaning
- Use corpus analysis and concordance to check validity of any infrequent hits and know whether to discard or not
- Focus on infrequent results by removing the frequent ones and then use salience to look at the strongest links
- Highly linked nodes indicate, in graphs which have reasonable frequencies, the accepted and most commonly used forms of the term
- Nodes which are separated from the rest of the graph may be specifically collocated with the other node they are linked with, may have a different meaning to the highly linked nodes, or may not be in common usage in the corpus in question but could be elsewhere. Worth further investigation.

2: Data representation and analysis techniques

An important aspect of the thesis has been investigating the choices relating to data representation. This has taken two different focuses: data framing which has focused on the multiple nature of scientific nomenclature and vernacular terms, and how these can be framed as single or multiple units.

Corpus pre-processing: data representation considerations

- Original no processing
    1. Emphasis of each step of the taxonomic hierarchy (species, genus, family, etc.)
    2. Links in the graph through shared species level or below terms (such as Linnaeus). Not present in the unified corpus.
    3. Accentuation of role of genus in grouping
    4. Comparatively more relations identified overall
- Unified term as one

1. Emphasis of taxonomic entity (species, taxon concept) as a unit
2. Accentuation of role of vernacular variant in grouping
3. Comparatively fewer relations identified overall

Data representation: frequency and salience filtering The filtering methods used in the analysis were frequency and salience. The profiling studies revealed the following about the different relations highlighted by each filter and how they could be used in combination to tailor the results.

Frequency

- Frequent, highly connected terms are highlighted
- These will usually equate to accepted names, or vernacular variants
- Initially very high numbers of relations identified, which quickly slope off making it an easier measure for large amounts of data
- Can use in combination with salience to reduce total number of relations but focus more specifically on some other, less frequent, occurrences in the text

Salience

- Infrequent, less connected terms are highlighted
- These will usually equate to non-official variants, species not a main focus of the test corpora, scientific variants

3: Profiling meaning from the analyses

Comparative dispersion patterns (dispersion and frequency analysis)

1. Frequent and broadly distributed
    - These are the names which are in most common use (so one would expect these to be the accepted names, and most usual/recognised common names for these species
    - If there is variation from that then it indicates either discrepancies of opinion as regards the naming, or changes or specific domains in which different terms might be used (this could be investigated with corpora which control for the variable required)
2. Frequent and specific

- Indication of author-, time-, domain-specific usage of a term, depending on where the term is concentrated

3. Infrequent and specific

    - Less used or not accepted names (this may be in the context in which the corpus is focusing)
    - As regards common names, either ones used in specific areas or for specific purposes
    - These can be time-, author-, domain-specific so this should be borne in mind
    - Indicate former names, which are now outdated (unless looking at historical corpora)

4. Scientific names

    - Frequent, but common names still more frequent (indicate a trend to mention the scientific name but then talk general using the common name)

5. Accepted scientific name

    - Most common usage of scientific names
    - Well-distributed across corpora of different types
    - Highly connected (node hub, or at least with many incoming/outgoing edges) in graph representations

6. Common name variants

    - Expect to be widely used from what has been seen in the test corpora
    - From the test corpora it would seem that overall ratio about 80:20 to common names in comparison with scientific nomenclature
    - Most used common name to be most highly distributed
    - General use common name highly connected (node hub, or at least with many incoming/outgoing edges)
    - Where connected to various scientific nomenclature being a hub can be understood as more general than the species level names used to represent
    - Common names used in specific settings to co-occur with a specific scientific variant or in a specifically in a certain type of document
    - Where used in specific settings in the network relation graphs will only have links to these variants

7. Specific terms – domain, time, language specific (for example)

    - Where name changes for a species have taken place, this is often indicated in corpora by multiple variants occurring at the same time (in the same documents)
    - Previous accepted names most likely to be found in references section of academic corpora
    - In non-academic corpora more variety of both scientific and common names expected (less consistency)

Relation network graphs: node and edge identification

1. Scientific name variants

    - Accepted name variants, which constitute a focus of the corpus, are more connected
    - Any scientific variant not a focus of the corpus will not be well connected
    - Infrequent variants may only be linked to specific terms
    - Species-level names may be surrounding hub nodes of higher hierarchy (common names or other)
    - High frequency scientific names (ones in common use in the corpora) should be identified using frequency
    - Rarer, less commonly used variants can be found using salience

2. Common name variants

    - Frequently used common name variants likely to be hubs
    - Often link to multiple scientific names
    - Better to look for through frequency

3. More general terms

    - Usually source node, not target node
    - Tend to be at the top of the hierarchy, having exclusively outgoing edges
    - Hub nodes

4. More specific terms

    - Usually target node, not source node
    - Tend to surround hub nodes
    - Tend to have incoming, not outgoing edges

6.6 Chapter Discussion

These nomenclature profiling studies have served to provide an application of techniques developed throughout the research project, and traditional existing corpus analysis techniques, to feed into a critique of existing ontological and other knowledge representation resources as well as perform a first look at terminological and conceptual stability of the terms within these resources.

The analysis of three different existing resources revealed the vast variability in nomenclature variants for some taxa, and also questioned the the homogeneity of meaning among them. The nomenclature profiling studies have applied corpus analysis techniques and the relation network graphs to look at the use of these terms in context and empirically to shed light on nuances of meaning and usage, and identify any significant differences between the two test corpora.

There was a large amount of concordance and stability across the two corpora on all three accounts. While the WEB corpus revealed greater coverage overall as regards the focuses of the profiles, it is also larger. The increased variety could also in part be related to its not exclusively academic focus. To investigate this quality in future it would be necessary to develop corpora that are more specifically and carefully created to represent different domains in this way to make a better comparison.

Frequency and dispersion analysis was used to identify possible areas of interest for further study. Anomalies were also identified through this technique, and checked for veracity through the concordance lines and relation network graphs. The co-occurrence aspect of this part of the analysis also indicated whether terms were likely to be mutually exclusive, near synonyms or if they were used to describe nuances of meaning. As regards scientific nomenclature variants, the co-occurrence or not could indicate whether both are still in current use or if one or other is used only in specific circumstances or the past. To better study both these features in the future corpora with documents marked-up for the different sections of each document would be helpful.

The relation network graphs allowed for a deeper analysis of the semantic relations between the terms themselves, to draw some conclusions or infer some probable outcomes as regards the specificity of a term, its real link or not to the taxonomic entity focus of the study, and also find out whether there are other hidden meanings within the term itself.

Finally, after having performed the nomenclature profiling studies, guidelines were developed that collated learning from throughout the course of the thesis, providing a guide by which anyone who wishes to replicate this sort of study with the means to do so. It gathered together the aspects of traditional corpus analysis used in the thesis with the methods developed here in relation to semantic profiling using relation network graphs and frequency and dispersion analysis.

Chapter 7

Phase 4: Expert evaluation and outreach

The final phase of my research, Phase 4, relates to the relevance cycle of the design science process. Chapter 5 dealt with the technical evaluation of the method developed in this thesis, whereas this evaluation was used to used to gain further insight into specific issues, such as ones relating to the data itself and its subsequent analysis, to obtain feedback as to validity and applicability of the method developed to the area of biodiversity, as well as to identify possible avenues for future applications. For this, a focus group design was chosen to discuss nomenclature-related queries with experts in biodiversity-related fields, as well as present them with the results and conclusions of my research, to evaluate the usefulness of the data extracted and also provide an evaluation of the conclusions drawn. The main evaluation and outreach consisted of a small focus group comprised of professionals who specialise in a variety of biodiversity roles. Three professionals took part in the focus group, two of whom are researchers in ecology at the University of Brighton (P1, P2) and the other is a tech lead in the informatics team at the Natural History Museum (P3). A fourth person, another researcher at the University of Brighton who specialises in fish ecology, was supposed to participate in the focus group but was unable to due to issues arising from the Covid-19 pandemic and resulting closure of the university (P4). A more informal conversation took place with P4 at a later date. This conversation was used to explore some of the more specific questions as regards the thesis data, as P4 was the only specialist with specific expertise in fish species. The Covid-19 pandemic meant that both the focus group and chat had to take place remotely. All the information relating to the focus group can be found in Appendix F. This includes pre-focus group questionnaires, the focus group and informal chat transcripts, which were formulated as described in the corresponding section in Chapter 3.

The chapter is set out as follows:

- Participant professional roles
- Resources
- Identification of species in data
- Scientific nomenclature versus vernacular variants
  - Scientific nomenclature
  - Vernacular variants
- Usefulness and applicability of method developed
- Conclusion

Each of these sections include a discussion of the opinions and thoughts expressed by the various participants as regards each of the above themes, including problems, rules and usage and best practice. There is then a subsection which considers specific examples from the data extracted in Chapter 6as well as some more general examples relating to the method developed in Chapter 4. The final sections consider the opinions about usefulness and applicability of the method developed and then an overall conclusion to the chapter.

7.1 Participant professional roles

All four participants responded to a pre-focus group questionnaire, which was used to guide questions in the focus group itself (see Appendix F for full responses). This questionnaire included information about their professional roles as it was deemed important to be able to see if there were differences between people's perspectives and experiences. The roles as identified by the participants were: three with researching roles (P1, P2, P4), two of whom also identified as having teaching roles (P1, P4) and one a scientific software engineer (P3). Three identified as working in ecology (P1, P2, P4), of which two also selected the category of biology and biodiversity (P1, P2) and one added a conservation focus (P1). The last identified as working in the area of informatics (P3). The other data gathered from the pre-focus group questionnaires has been included in the main body of the chapter because each relates to specific themes identified within the focus group and other discussion.

7.2 Taxonomic resources

This section considers participants' usage of various taxonomic resources, and any comments they had regarding the accuracy of resources or problems they may encounter. The section proceeds to analyse participants' comments relating to findings in this thesis about variations in resource representations and examples where the data in this thesis contradicted the taxonomic resource entry.

7.2.1 Usage

Half of the participants used the Catalogue of Life and also half used the Encyclopedia of Life in their work. Other resources used included: FishBase [67], AmphibiaWeb [33], Wikidata [219] and GBIF [77], IOC World Bird list [75], Handbook of Birds of the World (HBW) and Bird Life Taxonomic Checklist [20], Clements Checklist of Birds of the World [40] and the IUCN Red List [1]. The researchers tended towards using resources which have a specific focus on their species of interest and turn to more general resources such as the Catalogue of Life when having to go beyond their organism groups of primary interest (P1). P3 tended towards more general overarching frameworks such as GBIF because of the capacity to map data which had been based on multiple sources or taxonomies to the same place.

Resources were used by all participants to check name variant status. Three participants also used them to check the taxon classification of a name variant, and two used them to check the taxon classification of name variants across data sources. Only P3 used resources to map data to, whereas P4 responded that he also used resources to annotate data. When this answer was explored further in the guided conversation, it would appear that P4 had understood the question differently to that intended. P4 had not understood the question to be referring to the automatic annotation of data. The discussion both in the focus group and the conversation with P4 revealed that P1, P2 and P4 use taxonomic resources to manually check data, and in fact that their work is predominantly manual in its approach.

7.2.2 Quality

In general the participants held the resources they use in high regard, although two participants noted that the information in resources could be ambiguous (P1, P3), and P1 said that conflicting information between resources could sometimes cause problems.

When this question was explored in more detail in the discussion, signs of quality were identified as being endorsed by reputable agencies in their area of specialisation, such as Bird Life International (P1), as well as in the case of GBIF, linking various taxonomies or linking back to specimen data (P3). There was also support for the quality of the Catalogue of Life taxonomy specifically through its application at the Natural History Museum in their ScratchPads [215] platform, which “allows scientists to create their own taxonomies and describe their species and specimens” (P3). Catalogue of Life is available through ScratchPads and P3 stated that it is one of the most popular imports and “we have done some analysis on how much they have modified it after the import and it’s not hugely modified afterwards. So, it seems the scientists using the platform seem quite happy with it as well”. The general feeling was that reputable resources were well-curated and comprehensive (all), and that the important thing was to use a resource that was suitable for one’s needs. One of the researchers (P1), for example, commented “A lot of my research links in with extinction risk, so I use the IUCN Red List data quite a bit.[...] But that is what I use on a kind of global scale. But I also use a British taxonomy as well if I am focused on more national based data. For me it depends on the regional scale, actually. So whether I am doing sort of more global research or looking at more national checklists.” This statement gives an indication as to the multiplicity of the resources, and the fact that different resources serve different purposes, which will be considered later.

However, there were points made about issues with resources, such as a published resource, called the Wilson Reeder Taxonomy [221], had misspelled some of the nomenclature (now corrected) (P3), in the edition published “11 years ago” (P3). The edition cited here is the most recent as the dates did not match exactly. There is another edition due to be published soon. Issues related to erroneous resources included the propagation of these misspellings in the literature and problems this caused for mapping purposes, which highlighted the need for well-spelled taxonomies. P3 also highlighted that many resources are out of date, “I know lots of this sort of name resolution services. I mean, some of them were created about 10 years ago, they’re still live, but the names in them are now out of date a lot of the resources won’t actually resolve the currently accepted name which propagates the inaccuracies”. P1 also mentioned that there is a lag between updates in one resource to updating a secondary resource which uses the first as a baseline (see pre-focus group questionnaire responses, Appendix F). Overall the discussion identified problems of multiplicity, time and outdated resources, which all contribute to the further propagation of misspellings or outdated information. All three focus group participants agreed that overcoming these obstacles is not difficult for a seasoned scientist because they have the knowledge and the tools to effectively search and identify the correct usage. In contrast, they agreed that for scientists starting out, or people working in different fields this is not necessarily the case (see Appendix F focus group transcript, p.6). Scientific nomenclature is used by everyone from scientists to the man on the street, which would indicate that these variations in usage will always exist - as they will continue to exist in the legacy literature.

It was interesting to follow this focus group with the chat with P4, who did highlight continued problems using resources because of their multiplicity. In his case there seemed to be a level of distrust in the reliability of resources: "I think we need a more reliable resource to go to. I mean, I say the one I tend to refer to is FishBase because I'm familiar with it... How accurate it is I have no idea and actually if you look at each fish entry it can actually be quite confusing and if you go country by country sometimes you see, you're like wow even some of the Latin names are different". He also mentioned a specific case in which he had a journal article returned for corrections because of nomenclature usage, and that he "realised just how much is out there that's incorrect, because you know I'd checked some of those species, you know I normally use FishBase and they're still wrong in there some of them". This highlights the issues of keeping resources up to date given the changes to taxonomic decisions. P4, together with the other participants in the evaluation, highlighted the increased difficulties that people new to the area (students, inexperienced researchers) had navigating taxonomic resources and identifying the right name. In contrast with the other three participants, P4 did not feel that the choices were clear for experienced researchers. This response may be linked to the area in which this researcher works. No conclusions can be drawn but the commentary both in the focus group about the domain of fish, because of the variety of non-academic domains it covers (P2: "I would say maybe fish in particular where you have all kinds of strange papers and stuff on specific areas. And how whether they aren't necessarily using the same names as other people. You do see that in some literature where some of the more applied stuff perhaps written by types of academic, or not academics even, you might see more name variation") and then P4's comments would suggest this possibility, and suggests that this could be a useful application of the method developed in this thesis, to evaluate name variation across different domains.

7.2.3 Resource issues identified in this thesis

My research had uncovered three instances in which the Catalogue of Life had seemingly misclassified vernacular variants to a scientific taxon: the inclusion of brown trout under the Oncorhynchus mykiss entry and the inclusion of both brook and lake trout as vernacular variants for Salmo trutta. There was no fish specialist in the focus group, however as regards Salmo trutta and brown trout, P1 said, "I'm hoping that people wouldn't use the term brown trout to describe a rainbow trout. It might happen in terms of, you know, case [...] of misidentification". When I explained again to confirm that it was the Catalogue of Life that had classified it in that way, and my data showed the link only between the Salmo trutta and brown trout, not with Oncorhynchus mykiss, P1 responded, "I don't know why that would be the case. I don't know.".

When catching up with P4 at a later date, he in fact was not at all surprised by any of the seemingly misclassified terms. The two Salmonid species that were studied in this research are

Salmo trutta and Oncorhynchus mykiss, both commercially important fish. P4 described the frequent miscommunication relating to these and other species due to scientific nomenclature that differs from country to country, misidentification (as highlighted by P1) arising from farmed fish escaping into the wild, and the use of vernacular names in some areas which adds to ambiguity.

Lake and brook trout are not Salmo trutta but according to P4 are involved in further confusion in usage in the literature. He recounted that, "there is a very famous case in Yellowstone lake, where the lake trout, I think they introduced lake trout and then they pretty much wiped out the brook trout [...] And when I was reading it to give the lecture, the two names were used interspersedly". This provided support for that found in the test data, and also highlighted further issues on the identification level, which then impacts on the literature and nomenclature usage and understanding.

The information both from the focus group (P1-3) and conversation with P4 provides evidence that the method developed in this thesis could be used to check for coherence of usage between taxonomic resources and data. It could be used to identify where there are disagreements between resources and the data, and researchers could use the lexicographic method to identify where the issue lies: the resource or the usage, as well as identify possible patterns in usage. P4 gave direct support to this argument. When he described these difficulties, he stressed that anything that can help to pick apart the inconsistency between resources and usage could be of help (see Appendix F.7, p.14).

7.3 Identification of species in data

Participants in the pre-focus group questionnaire responded to say they used resources to check name variant status, to check the taxon classification of a variant in one resource and across resources. This is all linked to the identification of species in data, and refers back to the fluidity of the biological taxonomy.

7.3.1 Difficulties

The literature review highlighted the difficulties in identifying species in data. This issue was discussed at length with the focus group participants. Many of the issues raised were linked to the fluidity of taxonomies, although automatic processing issues were also mentioned. The latter did not constitute a primary focus of the discussion because only one participant (P3) had experience in automatic processing, but it would definitely be a focus for further exploration in the future.

Taxonomy fluidity

Taxonomy fluidity refers to the changes that taxonomies undergo in the hypothetical process of biological taxonomy, both on a terminological and organisational level. The issues surrounding taxonomic fluidity that emerged throughout the focus group were separated into the following themes: splitting and lumping, taxonomy alignment and transparency or tracking issues. Tracking can be understood to mean the ability to follow the course or series of changes that a specific species undergoes as regards nomenclature. The term was used because of the terminology used by experts in the focus group and chat. Transparency, for the purposes of this analysis, is a closely related term also used frequently throughout the focus group, which is used to refer to the explicit identification of taxonomic resources to be transparent in the definition of a species used, or lack thereof when this sort of information is lacking. The splitting and lumping phenomenon in biological taxonomy means that one species can pass to be classified as two or more (splitting), or vice versa (lumping). Finally, taxonomy alignment refers to the capacity to align different taxonomies which have different organisational structures because of this fluidity.

The conversation with P4 will be dealt with separately as there seemed to be more underlying distrust in accurate identification leading to problems in the nomenclature usage in the literature that did not emerge in the focus group discussion.

Tracking or transparency The ability to track, often inextricably entwined with transparency in one's work, seemed to be the most fundamental issue faced by all three focus group participants in this area. This was because it was identified as the key to determining whether one is able to identify the species within the specific taxonomy as intended by an author or not. On the subject of tracking, P3 said, "Trying to track changes in redescriptions over the years, particularly older ones that aren't in GBIF. But on the whole, it's quite easy to track down, once you... [PhD researcher: understand?]. Yeah, yeah, exactly. And I find Wikipedia and Wikidata are useful as well. Because there's often references to things." This reinforced the earlier message that the problems arise mainly through a lack of knowledge or experience because errors are likely to occur when you are unaware of the pitfalls (P2, p.6). All participants agreed that familiarity with the process and specific knowledge about the species concerned was central to the issue of correct identification.

Transparency in taxonomy usage was closely linked to tracking. P1 explained that "when [...] reading scientific articles they don't always clearly highlight where the data, or what taxonomy they are using. But a number of journal articles do. So it can be quite mixed when looking at scientific literature". If the taxonomies are not provided in name of transparency this leads to a lot of detective work and sometimes assumptions being made (P1). All participants agreed that this was important. Without transparency as to the taxonomy used it may be impossible to know how to handle the data. P1 explained that "if you are given [...] a trait database, you are given some trait data, what species and what taxonomy is being assigned to it. Because you may need to pool that data if you are using a different taxonomy or might have to try and split the data somehow otherwise. It can get quite difficult if you are interested in trait and characteristics data but you don't have that transparency about what is being used in taxonomy as well." P3 highlighted the issue in relation to climate modelling saying, "if we can't actually tie down exactly what species it is and identify the traits to go along with that species, the models break down and the data too. Yeah, so it is vital" (see Appendix F, pp.8-9). This all provides evidence to the link between nomenclature and taxonomies, and how the ambiguities in language used impact the realities of research.

Splitting and lumping Splitting and lumping was mentioned because it is related to the fluidity of the taxonomy and importance of transparency. P1 and P2 emphasised the impact of splitting and lumping in conservation research, because of the clear impact on the way species are classified (threatened, of concern, etc.) because the way splitting and lumping affects their population size and range, which in turn has far-reaching implications on how species are treated (see Appendix F, pp.8-9). P1 stressed that "it can get quite difficult if you are interested in trait and characteristics data but there isn't that transparency about what is being used in taxonomy as well", because of the difficulties this causes in making proper judgements as to how to handle data, as described in the previous section.

Taxonomy alignment Taxonomy alignment was a concern for P3. In relation to differences in taxonomies, he mentioned a specific project with "the Wilson Reeder mammal taxonomy. It was published 11 years ago. And so trying to join up the data that was published under that taxonomy, with what is now considered the standard, which is the ASM, mammal diversity database. And the number of redescriptions and synonyms... So trying to match up those two different taxonomies, even for a mammal taxonomy, yeah is incredibly complicated." This highlighted the real issues in relation to automatic processing and informatics in general, because mammal taxonomies are not assumed to be as unstable as some other taxonomies, as is implied in P3's comment. The researchers did not highlight this as an issue for them. They tended to work with one taxonomy or another depending on their focus within a particular study. This contrast serves to highlight an important point about how different specialists use resources in different ways. While P1 and P2 tended to use taxonomies manually and used them to refer to one or another depending on the task they are undertaking, P3 was interested in amalgamating and inserting data into larger, integrated platforms for later usage, using automated means. This is important because it highlights the different processes and where the difficulties lie in each case.

Automatic processing

Identification of species was also mentioned in relation to automatic processing. The issue identified here was linking the usage of broader terms in parts of descriptions to specific species' mentions earlier in the text, both in relation to using vernacular terms and also familial terms in the main body of a text. P3 mentioned that "with trait mining in publications, because they kind of set the scene before describing the species sometimes. It's actually with older publications. Then use the vernacular to describe the landscape with some useful traits thrown in. Other than that we don't actually use vernacular names very much." The project that P3 was talking about referred to historical literature, but these characteristics were also identified in my research. This should form another focus for future exploration in relation to my work. The method could potentially help tackle these issues by linking the names to more general or vernacular variants as shown in Chapter 6. The use of the vernacular in this way in some domains was supported by the other two participants.

Accurate identification of specimens

The veracity of species' identification was an issue that emerged as a great source of ambiguity in the conversation with P4 that did not emerge in the focus group. The principle source of this ambiguity was identified as the difficulty in the accurate identification of taxa in the first place. This was identified coupled with the compounding issue of multiple names for the same taxon and as a consequence the impact this has on research usability. This was a subject that we returned to repeatedly during the conversation.

The conversation with P4 indicated that the issue of accurate identification of specimens in the domain of fish is very relevant. This was linked to the unreliability of using physical appearance to identify taxa, coupled with the fact that physical identification is often the only resource available where genetic testing is not feasible. P4 gave various examples of the unreliability of physical features for identification. These included the variability of humans, and how appearance can be drastically changed by habitat, either in the example given relating to black crayfish, "you've got crayfish that in normal riverine systems they are quite clear and light coloured and then you get these random black ones that live in silt and they just grow bigger because they have more detritus and it's kind of black and then people think they are different species". He also described how different trout species "silver-up" when they go out to sea. P4 described how this can result in the misidentification of species, "I think it was something like 32 species of catfish they found, but then when it came down to it they were only about 9 different species. I mean they look different, because some of them were darker, bigger, blacker, and then when they actually did the molecular testing on them, they whittled it down to about 9 species, they reckon", which can result in an excess of different scientific names and creates confusing environment for communication. It seemed from the conversation that the changes species undergo in different habitats also contribute to a multiplicity of vernacular names to describe these different expressions of the same species, once again obscuring the panorama of what is what. "Everything is so different, you can't compare. People are using different methods, different even the same chemicals, they're using different tissue, different temperatures, different concentrations, different timings, you can't then compare that against, you know. And it's the same with this. You can't, it's actually very difficult to actually marry two papers up when they are using different terms."

However, the unreliability does not end there. As a result of the difficulty of identifying species through physical attributes, and a lack of transparency in the process, P4 seemed to have a distrust for the names used in the literature. P4 described how he has "actually excluded literature from [a paper], because I'm not sure, because I've thought, actually I've got no evidence, I've got no proof, there's no images to be able to say actually that is Salmo trutta or that's Salmo salar or whatever. And actually what people are calling these things is just random." Among the warning signals for doubting the veracity of information in a paper, P4 signalled: taxa apparently in areas that their range do not cover or in environmental conditions under which they would not survive. He went on to describe real problems that arise from this uncertainty in the correct identification of organisms in practice, in which he said, "I think misinformation out there on what those species are doing. You know, what they are. And certainly because I've worked in aquaculture, you know there are a lot of discrepancies in aquaculture, as well in what species they are farming, unless they've been translocated out the country, um put somewhere else, there are very different techniques for different species. And that has caused some issues as where people have tried to pick up techniques for a species that they relate as the same species that have been farmed, and then gone back and it hasn't worked". Statements like these highlight the problems that exist at every level of the chain: people do not know what species they are in charge of, this leads to inaccurate reporting and then techniques or assumptions are made on the basis of this reporting but in fact the application is not suitable for the species. P2 did mention aquaculture and anglers in the focus group as a domain in which there may be less than consistent use of the nomenclature, and P4's remarks would support that. As P2 also pointed out, people reporting who do not specialise in taxonomy may have significant impact on usage (see Appendix F, focus group transcript p.12). Further investigation into patterns of nomenclature usage in more academic, or more taxonomic-focused domains in contrast with commercial domains may be an avenue worthy of exploration.

7.3.2 Good practice

Good practice in avoiding these issues was explored. An important aspect of good practice to permit the identification of species, the specific variant and exactly what species an author is referring to, is the paper trail: allowing readers to identify this through a number of mechanisms that were identified by the participants.

One of these is the use of authorship: this refers back to the description of the species as per the author of that taxonomic description. This seems to be applied variably, but the importance of this varies, as will be explored in the scientific nomenclature section of this chapter.

Another good practice is that of referring to the taxonomy being used when writing a paper or creating a database. All participants agreed and P1 said, "I am reviewing a paper and they don't tell me what taxonomy they are following I will put a comment and ask them what taxonomy they were following – for that transparency like you said." Even so, this practice does not seem to be applied throughout and sometimes the reader is left to try to investigate which taxonomy might have been used. As an addition to this, P1 said, "I use a taxonomy that is endorsed and used by BirdLife International – which is the custodian of the bird section of the IUCN Red List. So I tend to always go to their most up-to-date checklist, based on the taxonomies they use", highlighting the importance of up-to-date and reliable resources. This was supported by P3 in his comments about the corrections of previous versions of taxonomies and also that many name resolution services are built, but subsequently become outdated and therefore propagate errors. P2 mentioned "with the papers I have been working on recently they tend to all defer to GenBank basically so [...] is tied specifically to the genome, so it will be an identifier 16rsDNA,. . . RNA or something. And that will be encoded in GenBank specifically with a name and now will link back to that. And I mean that certainly works when you know what you have molecularly. But at least with lots of amphibian and reptile papers they are supposed to be able to say what it is specifically." Genetic testing would also be relevant to mitigating the problems highlighted by P4, although this would only work on current or future work, and would not be applicable to legacy literature and may not always be possible. This is explored further in the next section.

7.4 Scientific nomenclature versus vernacular variants

The scientific nomenclature discussion, as well as the vernacular variant discussion, were split into the broad themes of rules versus usage. Figure 9.94 provides an overview of the split (% per section), according to the amount of the focus group discussion dedicated to each part. The conversation with P4 was excluded because it did not follow the same outline as the focus group discussion.

Figures, 9.95 and 9.96 show the different weight of the discussion given to rules and usage respectively for scientific nomenclature and vernacular variants. The scientific nomenclature required an in-depth discussion of both rules and usage, whereas for vernacular variants the discussion focused nearly entirely on usage.

7.5 Scientific nomenclature: rules versus usage

There are many rules related to scientific nomenclature, which are discussed here as regards the comments made during the focus group. In this section rules and usage were amalgamated because of the need to contrast rules with usage. This format allowed for a more cohesive argument.

7.5.1 Authorship

Authorship is defined as the description of the author plus the year as part of the name. P2 stated that the Latin name with the authorship was actually the official name, not solely the Latin, although the nomenclature code states, "the name of the author does not form part of the name of a taxon and its citation is optional, although customary and often advisable" (Article 51) [102]. Extra grammatical features, such as brackets around the author and date or not also provide information as to whether the species has been previously described or not (Article 51.3 of Zoological Nomenclature Code [102]). The focus group participants questioned to what extent these rules were known across the whole scientific community, and the resulting impact on their usefulness if not widely known, and therefore inconsistently applied (P2).

In practice it seems that authorship is often dropped. I identified in my research that this was the case in the test corpora, and when questioning participants, there was discussion and general consensus that the usage of authorship was varied. All participants agreed that the impact of authorship inclusion or exclusion on ambiguity depended very much on context. The general feeling was that taxonomic descriptions must and would have authorship associated with it, but that other areas in their experience sometimes did not. The areas mentioned were entomology (P3) and conservation (P1). Whether the usage or not of authorship caused ambiguity was also dependent on contextual factors: participants agreed that it was necessary to avoid ambiguity if dealing with a name or field which was constantly changing and P3 mentioned that it was specifically useful in data mining where particular discrepancies in a name trying to be tracked down, but that otherwise it is not a problem. On a more ethical note, P1 mentioned the transparency and also acknowledgement issues, as by omitting the authorship fails to acknowledge the describer ("don't know, just trying to think. I guess the potentially the lack of transparency and lack of acknowledgement. But a lot of journals I publish in don't explicitly ask that information to be provided").

7.5.2 Genetics

Information from the focus group participants highlighted the increasing tendency in some domains to require genetic codes to describe new species, as well as provide this code in any paper referring to a species with this information. P2 mentioned this referring to good practice. P3 added to this with his experience working on a paper in which a new species had been discovered and that they "were struggling to get names published because we don't have the genetic breakdown. So they are now refusing to describe new species without the DNA."

While this only works for current and future work which has access to said information, it does constitute a disambiguating factor in identifying species in the literature and shows how current practice and rules are evolving with the methods available for identification.

7.5.3 Nomenclature code

Compliance and difficulties relating to the nomenclature in some way were discussed at length during the focus group,. The conversation touched on many of the arguments identified in the literature review. Although the nomenclature code is very clear in the accepted name for a particular taxon, not all scientists respect these decisions for various reasons. Non-compliance with the code is not always wilful however, with difficulties in getting access to consistent up-to-date information being a prime culprit. Other issues identified were complications of spelling and the question of terminological consistency versus taxonomic accuracy. These are all discussed in the following sections.

Taxonomic resource issues

This issue was touched upon in the resources section earlier in the chapter. P3 highlighted that many name resolution services were outdated because they had been created years ago and no longer resolved to the most up-to-date accepted name. The Zoological Nomenclature Code is retrospective and retroactive [102], however many taxonomic resources are not always kept up to date because of lack of resources or finished projects that get forgotten. This means that people may unwittingly use outdated names. P4 also mentioned the issues he had with outdated resources, leading to the return of papers for correction, "I realised just how much is out there that's incorrect, because you know I'd checked some of those species, you know I normally use FishBase and they're still wrong in there some of them".

P4 complained of problems finding specific variants in the resource he uses. He described problems when doing research because this means he is unable to link a paper which uses a previous name to the current accepted name "I had this paper from like the 70s and things have changed quite a lot from there it was one of the first feeding, one of the first proper feeding studies that had been carried out on this group of sharks and it's actually really difficult to find the name of the shark in any literature because it had changed back in the 70s, you know? But that was quite a challenge [...] because then you've got a Latin name, that's linked with a vague common name, in Australia. And that means that it's actually very difficult to marry those two species together".

Spellings and names

Spellings were highlighted in the literature review, and also in the both evaluation session. While there are many rules, such as "whether you have parentheses around the name or not whether something has synonyms, where there is a further grammatical layer to that as well. In Linneaus, if it has has brackets around it for example Perca Flavescens. If it doesn't have brackets it means that the species has not been relisted or reclassified. And if it does have brackets, it means that it was was originally described as something else" (P2), and there are "very specific rules for how you say take certain things like say like a species named after a toponym, a place name would be neuter normally and so that ending that you take with end in a certain syntax or suffix" (P2). However, the feedback from participants in general was that these rules are not well-known and understood, with P2 adding that "a lot of people yeah, a lot of people just like it is about adding an I or S on the end of this but it is actually much more complicated than that". P2 specifically talked about "genders" and the Latin naming as "another language", stressing that in naming "you get genders in Latin and things you get like neuter, masculine feminine that can slip through and it does in a lot of papers where you get stuff grammatically misnamed but then that gets stuck because that's what it's called". He went on to talk about how "someone may try to fix that because they've got a knowledge of Latin and yeah... that's my opinion of that. There are very specific rules for how you say take certain things", which linked to the quote used previously relating to toponyms and the impact of gender on the syntax required. These complications seem to allude to the sorts of issues seen in the first nomenclature profiling study (see Chapter 6) as regards "Salmo gairdnerii" and "Salmo gairdneri". P4 could not add anything specifically to the discussion as to why the incorrect spelling of a synonym would be more common than the correct spelling, but P2's assertion that "a lot of people just like it is about adding a double I or IS on the end of this but it is actually much more complicated than that" indicates that this may be a case of what is happening there, and demonstrates that the method developed in this thesis could be used to look for similar patterns in the literature in other contexts.

The spelling issue is linked to the principle of priority because if someone names something incorrectly then that is still the name, even if it makes no sense, or as mentioned here, is grammatically incorrect. So the issue of misspelling is not clear from the offset because it depends on what perspective you view the misspelling (nomenclatural or grammatical). This will be described further below.

P1 supported P2 in his observations and said, "On the same thing I guess a lot of a lot of scientists and I'm including myself because I didn't you know, study Latin at school or anything I started using scientific names at university and but I was never really trained up in you know how scientific names are properly constructed and decided upon". Names are just a tool for most scientists to describe what they are talking about, but they do not have an in-depth understanding of how the name came about, what the bits of it mean. Therefore they rely on resources to ensure they are using the correct spelling, but as the P3 highlighted, misspellings get copied from specimen labels, redescriptions and so on. He also mentioned "in the new version they are actually commenting on the fact that the Wilson and Reader taxonomy misspelled some of the species' names", so authoritative versions result in the propagation of misspellings, as with papers that may have misspelled species' names because of the reasons highlighted above, or simply because of "typos" (P2). P1 described the situation as a case of "Chinese whispers".

Somewhat unsurprisingly as a result, misspellings seem to be a wide-spread problem, although all participants said that misspellings are uncommon and never occur more frequently than the accepted name variant for a taxon (see Appendices, pre-focus group questionnaire results). Three of the four participants said that misspellings could cause ambiguity in their work. As the above discussion highlighted, what a misspelling actually entails may be interpreted differently in different situations. When asked to define what they understood by synonyms and misspellings in the focus group, none of the participants answered directly, but P1 highlighted "With birds, one of the quite frustrating things, I don't know if it is the same with other taxonomic groups, but with birds quite often with synonyms there are very subtle differences. It could just be the last couple of letters in the specific name. I know if I happen to search and happen to have used one of the spellings over another I can, you can easily miss a species. So sometimes there can sometimes be very subtle differences, that are synonyms, recognised synonyms I am not really answering your question but it's something that I have noticed is particularly prevalent. If something ends in an A or a US. It can be quite subtle but it can made a big difference when you search for species, unless you use a wild-card search function which is what I tend to do nowadays. But yeah they can be quite subtle differences." This goes a way to highlighting the difficulty in agreeing on a definition. I identified that some resources would classify what P1 refers to as synonyms, as misspellings. This demonstrates the lack of consensus in relation to misspellings and synonyms, as well as multiple grammar issues that arise from the fact that the naming of species and use of the nomenclature is handled by so many people without a knowledge of Latin. These issues surrounding grammar and Latin were particularly highlighted by P2, who also mentioned the Principle of Priority, described below.

Principle of Priority The principle of priority was also brought up in the discussion, which is where the first accepted name of a species goes. This is the "oldest available name that applies to it" (Article 23.1) [102]. P2 highlighted that, "if the original classifier 200 years ago originally called it something that didn't make any sense, in terms of etymology, the description of what the word means. You still get stuck with that, that's the official name for it. It's the oldest that takes it. So you might have something that is incorrect – I can't think of any examples but there are some good ones – where something is actually called something that it isn't, in the Latin." This, he described can cause a number of problems: one it can propagate misspelling or misuse, it can mean that "if you read the name and you understood some of it, you would misconclude without further checking" (P2) and finally "someone may try to fix that because they've got a knowledge of Latin", but because of the Principle of Priority the first name stands. All of this seems to contribute to a plethora of spellings and variants, which seem to be here to stay.

Consistency of usage

As has been highlighted throughout the whole thesis, the taxonomy of species is being constantly added to, revised, reordered. There are many different perspectives of these taxonomies and the discussion of the consistency of usage seems to be central to some of the inconsistency in usage across different scientific realms and different authors.

P2 gave the example of the Acacia family, in which half of perhaps the Acacia tortilis species have been split and reclassified as a different genus (apparently Vachellia). P2 stressed that "sometimes when the species name changes people don't use the new names because of local usage and like you said recognition [...]. Half of acacias are no longer acacias. They've got split and half renamed to another genus [...] but all of the people in the field or the field guys and the people who know what an Acacia is, and we will know what an Acacia is probably. So they stick with the old usage just for ease for the, again I get annoyed at papers, even some tree books I've seen for example, have said although the new classification that there should be this we are going to stick with the historic names because everyone knows what it is." So there seems to be an opinion in some areas that maintaining consistency is more important than what the rules say. This was something that resonated with the other participants who agreed that in some cases, no matter how clearly wrong, people may have their "favourite name" and continue to use them (P3). These comments give weight to the idea of inconsistent usage in the literature and the need to be able to link species to other mentions within a dataset to be sure what they are talking about.

While P2 expressed his exasperation for those who maintain the status quo despite new scientific evidence, P4 stressed the difficulties he experienced because of changing nomenclature usage. P4 complained of problems finding evidence to link previous and current names to each other across and within papers, which he described as "a problem, I guess, in that when you are citing literature, if you go, if you go with the most up to date name in the actual piece that you're working on, but then you reference back, they are two different species names. So the title of the article, this is the problem I had, I went back to the editor, to ask them what to do. Because you're telling me that that species name is now wrong, but the citation, or the reference is now telling me that that was that species. How to reference that back, I've actually gone back to authors previously, when I have seen this, where people have, and I've thought, oh, that's funny, that paper isn't about that species, and you get back to them, and they say, oh no, I did check it but the nomenclature has changed and basically that's no longer valid. So that's really bloody confusing now because you know, I'm working off an older name or not marrying the two up and I've found that quite a challenge when I'm writing." P4 also talked about the issue of correctly identifying changed nomenclature usage as referring to the same species and how it can result in incorrectly identifying species as two, when they are actually referring to the same thing "Salmo gairdneri, that's the one that caught me out when doing my first trout paper. It caught me out. That's the one that I thought was a different species and I thought it was a different species, because it's not a Salmo and the editor came back to me and said, this is the same species. So I felt very stupid. But I'd given it as a separate case of another species. So that one always jumps out to me when I see that". The issue of being able to prove two papers are linked without sufficient information about the taxonomic history of a species seems to be real, and indicates a lack of transparency as was highlighted in the focus group. The two different opinions from participants also show why and how different decisions may be taken and how this feeds into inconsistency in the usage of nomenclature in the end.

Another aspect of nomenclature inconsistency related to geographical idiosyncrasies in usage, despite the global reach of the nomenclature code. P4 also mentioned a geographic inconsistency of usage (linked to inconsistent identification, further undermining his trust in the literature), such as Salmo trutta being "pretty much the accepted brand [UK] but you go across the Atlantic and so many people are using it interspersely with other species that are related to America". This comment also suggests that many fish that may not really be Salmo trutta are being attributed to this species incorrectly. Problems in geographical specificity of scientific nomenclature usage indicates a problem of consistency that causes ambiguities. As regards the results in nomenclature profile study one, it indicates that "Parasalmo mykiss" may be a term used in Russia. Again, it would seem that the domain of fish, and particularly the species that were the focus of my research, may be a particular focus of inconsistency because of the commercial element to the work. The participants in the focus group suggested that it was "anglers, fisherman [...] doing strange things" (P2) or "author specific" (P2).

To summarise the scientific nomenclature section, the focus group and other discussion served to highlight the issues identified in the literature view and give a more personal view to why there are so many different applications of such a defined naming code. There are always multiple perspectives to take in the decision-making process and not all people involved in using nomenclature have the same levels of knowledge or come from the same perspective. The discussion also served to provide further evidence to suggestion explanations for some of the results out of nomenclature profiling studies in the previous chapter.

7.6 Vernacular variants

The final discussion topic centred around vernacular variants. Only P1 specified that she came across vernacular variants very frequently in her work, with the others responded to this between sometimes to very infrequently. However, all had agreed that vernacular variants can cause ambiguity in their work and there was a very fruitful discussion about vernacular variants, their usage and characteristics that could help to identify these issues better.

7.6.1 Rules

Participants agreed that, in contrast with scientific nomenclature, they were unaware of any rules governing the usage of common names. P2 said, "what are the authorities for it. Because in papers when you describe a species, you don't even need to give a common name. You can suggest one but people don't have to use it. Yeah. I'd say, for all species descriptions I've been involved in. You explain why you pick the Latin name. And I think only in a minority of cases to present a common name. People might use one using the Latin name. But often they don't, I suppose. I'm not sure really how it works to be honest." In support of this P1 highlighted that, "I don't know how, for example, the IUCN Red List or Birdlife International, how they decided on the common name that they use. For birds. But I don't know what the... because, for some of them I question because I would use a slightly different common name for some of the birds". This is interesting as while usage was recognised in the scientific nomenclature to sometimes stray from the accepted form, despite very clear rules to the contrary, here no one even knows how or why names are chosen. This adds to the confusion, and as P2 highlighted, they are regional and may have multiple terms in different languages or not exist in others. The other participants agreed with this comment.

7.6.2 Usage and importance

This indicated that vernacular variants pose a different problem. Vernacular names vary in importance depending on domain. P1 highlighted that "a lot of species, especially those that have been newly described, invertebrate species, or plant species as well, but just they just have a scientific name, they might obviously, have a very niche vernacular name that local communities use but not a globally used common name", and was supported by P2. This argument was taken further in the fact that the regionality of common names make it hard to communicate on a global level about specific species, as the terms differ and sometimes don't exist in other areas. P3 mentioned that he only rarely works vernacular names.

However, as was indicated by the introduction, there are areas in which the use of vernacular variants can be important, such as citizen science. Despite the problems, participants seemed to agree that using common names may encourage "more occurrence records to be deposited" (P3), and P1 added that "I encourage people to use iNaturalist and things like that. Of course there is no expectation that citizen scientists are trained taxonomists, so yeah, in those situations use of the common name is a given. But with iNaturalist you can upload an observation, have the common name but then you'd have the online iNaturalist community helping to get it down to a proper scientific observation" (P1).

In commercially important domains, such as fish and aquaculture, both the test corpora for this thesis and the discussion with P4 would suggest the prevalence of usage of common names, as well as pitfalls to this use. This was supported by the discussion in the focus group, in which P1 described how she used them a lot in teaching, in combination with scientific names, highlighting the "history" of vernacular names in avian taxonomy but also stressed that they are very multiple and "are very regional, very localised names as well" (P1).

Their usage seems to vary, but all participants seemed to agree that scientific nomenclature and vernacular variants go hand in hand. P1 said, "I would never for example, I wouldn't be allowed to anyway, open and publish a paper where I just use common name, I would always have the scientific name used alongside", and reiterated that the two names go "side by side" (P1) in a later part of the discussion. P2 agreed, stating that in papers, "I use the scientific names but then I will lapse into vernacular or common names for readability. But I will define them before that. If you have a particularly horrible Latin name you try to use it less. If you are writing about E-coli, in a different field, it's easy to talk about E-coli. But yeah, only for readability. But it's always, I just stick to scientific names" (P2). Finally, P3 said that generally he did not use vernacular names much in his work, but that he was currently working on a project in which the vernacular name was important for descriptions, which follows on from what the other two participants said. In relation to this research, perhaps the most important finding is the recognition by the participants that scientific nomenclature and vernacular variants work as a team. P3 even has trouble working with the parsing of data because of the usage of vernacular variants in texts, in which “we’re trying to parse a description often their species description is grouped at familial level. And so you need to have a semantic understanding of the entire narrative. Because they’ll have the specifics of a species then a broader, high level description away. But that is some publications. Makes it harder to extract the data. But for a reader I guess it makes a lot of sense” (P3) This highlights one of the essential issues with natural language and computer parsing, but which as regards names and usage my research could help to address.

This teamwork between vernacular variants and their scientific counterparts in the disambiguation of usage: whether highlighting multiple vernacular names linked to a single term or many scientific terms linked to the vernacular. The next sections describe issues identified in the focus group and how the method developed in this thesis may be able to respond to these issues.

7.6.3 Ambiguity

The focus group discussion identified that ambiguity in relation to vernacular names was context dependent. When used in conjunction with a scientific name, the ambiguity is removed or mitigated because the narrative makes it clear which species is being referred to, as P1’s statement above confirms. Participants also agreed that vernacular names can, on face value, communicate more information than their scientific counterparts, as P1 said, “in terms of a species identification, in terms of what it what it looks like its appearance or its locality”. This, depending on context, can mean that vernacular names actually provide more information than scientific variants.

However vernacular names are still ambiguous scientifically speaking: there are no rules governing the usage or definition of common names, so P2 warned that “you have to be sceptical of them” (P2). P1 supported this by saying, “some people use a given common name to represent different species” (P1).

So usage is not regulated by any body and therefore is even more multiple, fluid and therefore there is a greater margin for error. It is also less scientifically specific, with many terms that are used in relation to multiple species, as P1 explained in relation to citizen science: “when people say I have seen a gull, or a [...] black bird. That is very hard. When you have multiple gull species, and which species you actually saw.[...] So yeah, particularly for citizen science it can be very hard to know if they are using the correct name, and if they use a generic name what can you do with that realistically” (P1). This statement highlights not only ambiguity inherent in the usage of nondescript names, but also the resulting utility and also questions as to the veracity of the information.

The conversation with P4 also revealed how prevalent the usage of vernacular names is in some domains, and the problems that this can cause. An example given was the tendency to use vernacular names in the purchasing of species for laboratories "We still don't know whether we've got the shrimp that we wanted. They are called cherry shrimps, but the variation we have got across the shrimp that we got, we're not convinced that they are all cherry shrimps, that they are all the same species of shrimp, because they are so different [...] my colleague and I just said, you know, they don't even look the same, you know, but these are classed as cherry shrimp, and I can't remember the name of them, we've only just got them in. And we were like, are these even the same species? I mean they don't look like it but they have been classed as that, because, you know, they're red.". This was an issue that occurred despite using certified suppliers. The lack of genetic analysis, combined with the great variability of physical specimens sowed seeds of doubt. On the basis of the conversation with P4, common name usage seems to be prevalent in the areas of aquaculture, more than the other participants who work with other species such as bird species, amphibians and insects. Fish have another trait which stumped the participants in the focus group. In the data I had identified multiple vernacular variants for Oncorhynchus mykiss (rainbow and steelhead trout) (see Section 6.2) and also brown trout versus sea trout in relation to Salmo trutta. Through my own investigation I had identified that they seemed to be used for the freshwater and seawater versions of the fish. When this topic was floated in the focus group, the participants were stumped, they did not think it was possible to have two names used interchangeably for the same fish that are the same species in the same paper. However, P4 clarified that sea trout is Salmo trutta (brown trout), when it goes to sea. Steelhead trout is the seawater version of rainbow trout, also. While vernacular names are often used to describe physical appearance, and in the focus group discussion highlighted some of the uses of this, this can also cause ambiguity. In relation to both sea trout and steelhead, it is because of the way the fish "silver up" to go to sea, although they are the "same species genetically", but this could lead to differing identification. P4 noted that with pikeperch, this leads people to think "[Sander lucioperca] are hybrids of those two species". Finally of interest relating to this multiplicity is related to recognising these multiple terms as the same species. This, again, goes back to underlying knowledge and awareness protecting from misunderstanding. P4 said that "to start with I thought [steelhead] were a different species. Not that I needed to worry because I don't really work with rainbows, but it was only when someone said they're just the equivalent of sea trout [to brown trout]".

Much of the interesting findings in the nomenclature profile studies were related to identifying how vernacular names mapped onto specific scientific nomenclature names, and the participants agreed that without the context the names were too ambiguous. All the examples here serve to show possible applications of the relation network graphs to mitigate this ambiguity, because they can be used to show the specificity or lack therefore of a specific common name in context (such as that identified the broader definition of trout, the dual interpretation of sea trout, or which shows steelhead and rainbow trout as common variants of Oncorhynchus mykiss). This could be particularly useful for those new to the domain, or non-experts.

7.7 Usefulness and applicability of method developed

My research was well received overall by the focus group participants. They thought that the visualisations were good in general and also as a way of analysing "complex data" (P1), with P3 adding "anything you can do to visualise that's good." P1 added that she thought "it's certainly got some utility to it to be able to explore and to break down and zoom in and zoom out on the different the different levels, so to speak". Both P1 and P2 highlighted the importance of adding extra levels for context, giving the examples of time plus other sorts of thematic analysis, which was already planned for further work. P4 said that it would be very useful to be able to see where "if you see how they derived those names that would be really useful, because that would reduce the work we would have to do working that down", alluding to the lack of transparency in papers and if my method could help to uncover some of that logic it would help. P4 specifically mentioned that he does not "care what common name they use now, all I want to know is the Latin name, and actually the Latin name isn't sound. And that's the underpinning, and I'll tell my students all the time. It doesn't matter, why do we have taxonomy, why do we have taxonomic names, this nomenclature that basically tells us, because there's such discrepancy in the common names." Considering the use of common names in our discussion, and the apparent use of common names in some of the domains he works in, as he said, something which could help to map out the common names being used in conjunction with which scientific names would help to at least unpick the inconsistencies in usage of which he complains.

Only two people responded to the evaluation post-focus group questionnaire (see Appendix F for a link to the full results). To highlight the most interesting responses. Both respondents (P1, and P2) said that they thought the profiles developed provided a practical approach to dealing with the ambiguities identified, and Figure 9.97 shows which ones they respectively considered them to be useful to respond to.

In general they both thought the method developed in this thesis could be useful for them in their work, although neither gave any suggestions as to how. They also overall evaluated the session positively.

7.8 Conclusion

The focus group was unable to take place in ideal circumstances as a result of the Covid-19 pandemic. As a direct result of the pandemic it was necessary to conduct the focus group via a remote channel and one of my participants was unable to participate. Another missed 20 minutes of the group because he had to speak to NHS 111 in the middle of the evaluation.

Despite this the focus group served to gain valuable insight into the realities for this small, diverse yet cohesive group of people. It served to provide support for a number of the assertions made in Chapter 6 as regards usage, provide further information about the problems or ambiguities supposed by certain trends and also dispel some ideas about how important these are or not, depending on context. It has served to develop ideas as to future applications for the method and the most interesting/relevant paths to follow for future study, which will be explored fully in discussion chapter (Chapter 8).

In evaluation of the focus group itself, it was the first time I held a focus group and I failed to follow up sufficiently on some of the questions which would have provided further support for my line of investigation, particularly as regards taxonomy alignment. However, participants seemed to be quite positive as to the usefulness of my work and definitely liked the visualisation used.

Chapter 8

Discussion and conclusions

This chapter brings together the major findings of the research set out in this thesis, which aimed to "employ computational lexicography and natural language processing techniques to identify, extract and group nomenclature according to its usage in the biodiversity literature and use contrasting corpora and existing knowledge representation structures to perform a systematic empirical analysis of these conceptualisations". This aim was created to respond to the lack of any research into the actual usage of nomenclature in the biodiversity literature, despite the multiple issues identified in the literature review relating to nomenclature usage, which arise from the lack of a gold standard taxonomy and recognition that there is no one agreed stance as to the biological taxonomy. A lexicographic approach was taken because of the empirical nature of this approach, and the way that it aims specifically to look at word use in context, to emphasise usage above tradition or expectation.

At the beginning of the PhD, as mentioned in the introduction, a second aim which consisted of exploring the characterisation of trophic interactions in the biodiversity literature was presented, through the application of these same techniques. Preliminary research into this second aim was explored in the pilot phase of the thesis, but was not pursued in the latter stages because of complexities identified in achieving the aim as regards profiling nomenclature. Further work into interactions will be considered in the future.

The research questions related to the final aim of the research were therefore:

1. How does empirical corpus-based analysis use the linguistic evidence in the biodiversity literature to model the hierarchical relationship between species?

2. How does the knowledge representation model extracted in research question one compare with other knowledge representation approaches currently being employed?

3. How do conceptualisations between different corpora vary quantitatively (number or trends of mentions) and qualitatively (contextually or links between different mentions)?

The objectives of the research to respond to these questions were:

- model the hierarchy of relations between nomenclature references/units of nomenclature that is identified within a specific corpus (by extracting the relevant information) (RQ1)
- create a graph/tree hierarchy image of this model to compare to the ontological structure for validation and evaluation purposes (RQ2)
- produce an evaluation method to compare the relations identified for precision, recall (quantitative measures) and differences (quantitative and qualitative measures) between the different expressions of knowledge (RQ2)
- perform comparisons between the hierarchies extracted between different corpora and ontologies of choice to evaluate the conceptual stability of nomenclature references (RQ3)

The work in this thesis has culminated in a method by which to empirically extract ontological structures from text data for validation (which can be applied to cross-corporal comparisons, or ontology-corpus comparisons as required). The main contributions of the research can be grouped into three broad themes: data representation, knowledge integration, and lexicography and terminology. The following sections first outline the major findings of the research, and then look at the contributions that these findings make to each of the aforementioned areas. Afterwards there is a section that explores the strengths and limitations of the research, followed by one dedicated to future work.

8.1 Major findings

The research followed a design science model in which different phases were used to explore different aspects of the problem and iteratively feed back in to develop the method described above. Phases 0 and 1 focused on understanding the data, the approach and possibilities of the Word Sketches and different ways of presenting, or framing, the data in order to respond to Objectives 1 and 2 of the research. The validation and evaluation of this method responded to Objective 3. The evaluation was also used to demonstrate some of the differences between the frequency and salience parameters in the types of profiles each of these filters highlighted. The application stage applied these methods to formulate guidelines by which to interpret the patterns identified in the data both using traditional corpus analysis techniques and the relation network graphs produced from the Word Sketch information. The comparison between taxonomic resources demonstrated how the same data can be presented in different ways.

The following is a breakdown of the research findings:

- Method development

  1. Data manipulation: filter parameters
  2. Data manipulation: data framing
  3. Graph representation: relation network graphs

- Method formalisation and application

  1. Nomenclature profiling studies
  2. Guidelines

The guidelines, repeated at the end of this section but originally described in Chapter 6, encapsulate different stages within the method developed throughout this thesis and incorporate both existing corpus analysis and lexicographic techniques with new, novel approaches to the profiling of meaning, in this case specifically in the context of the use of scientific nomenclature and related terminology. They bring together various aspects of the findings as described below to present a deliverable from the thesis research, and as such are referred to as necessary in the overall discussion.

8.1.1 Data manipulation: filter parameters

In creating the profiles, it was necessary to explore different possibilities for representing the data. To do this, the extracted information was manipulated and analysed in different ways, to draw conclusions as to how each parameter affected said representation. Frequency and salience were the filtering parameters. These parameters were found to be effective used either in isolation or together to change the focus of the profile extracted.

The exploration in Chapter 4 and evaluation in Chapter 5 of these parameters demonstrate that, at least in the test corpora, each type of filter highlights nodes and relations of different types, as demonstrated by the divergence of the results within increasing salience. Frequency was found to highlight more frequent, highly connected terms: terms that were related to families of species which were a principal focus in the corpus, whether these variants were scientific, misspellings or vernacular. Frequency in the test corpora (JEFF and WEB) highlighted nomenclature references related to the Salmonidae family. Salience, on the other hand, highlighted less well-connected mentions and terms that were less frequently found in the corpus, in this case fish species that were less of a focus of the corpus but also many invertebrate species. This indicates that salience serves as an interesting filter measure specifically because of its capacity to highlight more infrequent, but strongly related terms. Both salience and frequency encountered high numbers of synonyms and out-of-scope entries, but in the results with the salience filter, these were particularly prominent. This could be related to the fact that salience could be seen to highlight the least connected terms (which can equate to terms that are either not the focus of a corpus, or that are rarer, often being irregular terms, or unaccepted terms). These results suggest that salience should be used to extract information about the variety of infrequent mentions within a corpus and the relations between them, whereas frequency should be employed in cases where a profile of the more iconic families or species within a corpus is required.

These fed into the guidelines under the heading of "Data representation: frequency and salience filtering", were developed on the basis of these analyses, to guide anyone wishing to apply these techniques in the future.

These guidelines can be applied through the method developed in this thesis for profiling nomenclature usage, and in guiding the output in line with the purpose of the research. For future, further work into these parameters, the following should be taken into consideration. The equation used to calculate salience, the collocation association measure applied in Word Sketches, highlights the strength of the relation according to the number of times the words appear together, in contrast with the number of times the words appear in other contexts. In the case of working exclusively with fairly technical terms such as the scientific nomenclature, the salience scores for all relations were fairly high and often actually were describing words that only appeared in the context in which they had been identified. The strong tendency towards very infrequent, less connected relations may be related to the characteristics of scientific nomenclature as described above. If working with word pairs that had a more varied range of salience scores (for example if the method was extended to look at general descriptive characteristics of the species) then the characteristics of the filter parameter may shift.

8.1.2 Data manipulation: data framing

The other parameter as regards data representation was the idea of data framing. This choice had to be made in the pre-processing stage. Scientific nomenclature, is, by nature, multi-part. The scientific nomenclature is an "artificial language for scientific taxonomy" [87], which is used to describe the supposed hierarchy of species through their families and orders etc. Binomial nomenclature represents the species' name, in that it states the genus (first word) and the species (second word). Then terms relating to further up the ranking of taxonomic hierarchy are represented by single word terms. This structure of naming for species meant that the model was a suitable focus for this type of exploratory investigation. In the preliminary exploration it was useful to leave each word as a term in its own right, to highlight each step in the hierarchy (from species level all the way up). However, unifying the terms allowed the information to be analysed from a different perspective, in which a species was considered an entity in its own right, rather than a sum of its parts.

These findings, first described in Chapter 4 of the thesis, went into the development of the guidelines in the section entitled, “Corpus pre-processing: data representation considerations”, and highlight how the same information can be presented in different ways, and manipulated depending on the needs of the researcher or purposes of the research. They also provide further evidence that supports the argument in the introduction and literature review as regards the relativity of meaning. The use of the parameters set out here would serve anyone wanting to apply this method in the future. These findings would help them to decide how to process their data and later filter it to extract the results in accordance with the research focus.

8.1.3 Graph representation: relation network graphs

The literature review talked about the importance of turning data into knowledge and the complexities in this task. An integral part of the method developed in this thesis was based in converting the information held within Word Sketches into graphs that could be used to do that, which I have called relation network graphs. The previous sections have explored some of the ways that these graphs can be manipulated to highlight different features of the data extracted.

Specifically relating to profiling and meaning disambiguation, during the research I also identified that in certain cases the classification of specific nodes (entities) or edges between nodes (relations) could be ascertained through the shape of part of the graph itself. These formations have been described as “hubs”. The criteria for identifying as a hub is set out in Table 8.1, presented in the guidelines in the “Network graph manipulation and analysis techniques” section, and was developed in Chapter 4 of the thesis.

In the original corpus, when this set of criteria was applied to the data with low filtering thresholds, the majority of the hubs identified were genus level, and surrounding nodes were species level. When the threshold for frequency was increased, this resulted in hubs that were increasingly more focused on common and general terms (vernacular variant and general terms such as species).

In contrast the unified corpus always identified hub nodes as general or vernacular variants as the main, surrounded by the genus_species binomial. The results were consistently accurate in that respect.

These findings were used as part of the application stage in testing the findings of the earlier stages of the research, and can be applied in future research to help to guide profiling analyses as done here. Future research in this area would look to expand on the findings to create a more robust, automatic system of disambiguation. The next section explores the application phase of the thesis, which was used to demonstrate the functionality of the method and take steps to formalise it for application beyond this thesis.

8.1.4 Nomenclature profiling studies

The culmination of the previous steps of analysis was to apply the method developed to data to profile nomenclature use on real bodies of texts. It also resulted in the development of the following guidelines to facilitate future application of the method developed throughout this thesis. These nomenclature profiling studies were used to compare, evaluate and validate existing knowledge representation resources and also provide a method by which the stability or consistency of nomenclature usage across corpora could be compared, to demonstrate differences and similarities in data representation across different resources and data.

Each of the nomenclature profiling studies, described in Chapter 6, started with an analysis of the profile for a specific nomenclature entry in three different knowledge representation resources. This analysis, while manual, served to identify the commonalities and differences between three reputable resources. This analysis served to demonstrate the argument in the literature review about the multiplicity of representations as regards the scientific nomenclature [70, 192], by presenting and analysing actual examples. It also served the basis of the profiles with which to compare the profiles extracted from the test corpora data.

The nomenclature profiling studies focused firstly in creating a profile of each nomenclature entry in the test corpora using the methods described above and then used these profiles to compare the representations within the data with the profiles of the same entries from authoritative resources chosen from the literature. The criteria for comparison were based on these four categories:

- Commonalities
- Broader/narrower meaning
- Gaps
- Contradictions

These categories were chosen to be able to perform a two-way evaluation of both the corpora in question and the chosen resources, in response to various issues identified in the literature: the lack of empirical evaluation of the resources, the incompleteness or one-sidedness of resources [66, 175], the claim that common/vernacular variants are only relevant in citizen science or non-academic literature [191].

In the area of data representation, these profiles served to make a comparison between the corpus data and the taxonomic resources, highlighting points of commonality, disagreement, gaps and ideas for future investigation. Specific findings called into question some of the assertions made in the literature review, such as the relevance of vernacular variants in the scientific literature. The test corpora contained high rates of vernacular usage. This was confirmed in the expert evaluation (see Chapter 7) and seems to relate to the domain-specificity of usage patterns in data. On the basis of this process, guidelines were created to formalise the method and allow for replicability of the method in the future.

This aspect of the analysis is reflected in the third and final section of the resulting guidelines, reproduced from Chapter 6, below. The third and final section, entitled "Profiling meaning from the analyses", focuses on the patterns that would be expected in the data for different sorts of names. The nomenclature profiling studies focused on three different nomenclature entries, which displayed different behaviour as regards their nomenclatural and other variant usage. The subjects chosen for each study were based on the original findings about their usage patterns: high frequency and variability of usage (NPS1); low frequency and some variability of usage (NPS2); high frequency and low variability of usage (NPS3). The nomenclature was analysed in its own right, then the findings used to develop preliminary guidelines as to patterns to expect in different entities with different sorts of behaviour for anyone wishing to apply my method to further work. The studies combined traditional corpus linguistics analysis (frequency and dispersion analysis) with the network relation graphs described in the previous section of this chapter and elsewhere in the thesis.

The comparative dispersion patterns can be used to identify whether terms are frequent and homogeneously used throughout a corpus, or if the frequency is specific to one part of the corpus. This is to check for names which may appear to demonstrate accepted name qualities but that are actually heavily weighted to one or few sources, for example.

The guidelines also set out specific patterns to identify in the relation network graphs. These graphs can be used to see the relations identified between different terms or variant mentions. The guidelines here can be coupled with the use of the hub disambiguation criteria to discern meaning, taking into account whether the corpus being studied is looking at binomial or more terms with each part in isolation (original) or as a single, unified unit (unified). These guidelines represent the formalisation of a method by which people can now start to use this approach to identify characteristics and profile nomenclature usage across a corpus. The next section looks at applications within the area of data representation for these methods were applied.

8.1.5 Guidelines

The guidelines first produced in Chapter 6 are reproduced here in full.

1: Method guidelines

Corpus creation and selection

- Choose a subject matter/species/family on which to focus your corpora
- Identify suitable data and collate following copyright regulations
- Process suitably according to the data representation considerations below to be able to analyse to required granularity (i.e. with metadata for publication date, author, publication type, etc.)
- For more detailed dispersion graphs ensure that document length is available for each document in the corpus

Resource identification and application

- Choose a/multiple suitable resources according to analysis requirements
- Pull a total list of names to tag the test corpora
- Perform an analysis of the different resources to provide a comparison if using multiple resources

Data preparation

- Build tagged corpora using adapted Sketch Grammar
- Consider lower-casing and other pre-processing steps with the following considerations:
  1. Lower-casing will improve the number of hits (as well as other pre-processing steps such as deleting extra spaces, removing line breaks, etc.)
  2. Lower-casing does impact on Word Sketches because of altered tagging but does not seem to have had a great impact in this case (further work)
- Use script to pull Word Sketches for all the names in the list from the chosen resources
- Transform Word Sketches into edge lists for manipulation

Dispersion and frequency analysis

- Perform dispersion and frequency corpus analysis techniques on the data
- Use raw and normalised frequency and ranking to make intra- and inter-corporal comparisons of name usage
- Intra-corporal ranking will show the preferred terms within a corpus and inter-corporal ranking will demonstrate if these preferences are stable or not across the different test corpora
- Use raw and normalised frequency to evaluate the weight of focus of these particular species or name variants in the respective corpora
- Use dispersion analysis to evaluate intra-corporal synonymity of different name variants by the levels of co-occurrence at the necessary levels of granularity (scientific variant co-occurring with single or multiple common names, multiple common names co-occurring or appearing alone in different contexts)
- Compare dispersion analysis results from each corpus to evaluate stability or lack thereof across the corpora

Network graph manipulation and analysis techniques The relation network graph analysis was a key part of the analysis and included the identification of characteristics in the graph that could be used for profiling and meaning disambiguation. During the research I also identified that in certain cases the classification of specific nodes (entities) or edges between nodes (relations) could be ascertained through the shape of part of the graph itself. These formations have been described as "hubs". The criteria for identifying as a hub is set out in Table 8.1.

Table 8.1: Identification of node characteristics through hubs and graph positioning

Node identification	**Meaning**	**Node description**	**NC**	**CC**	**Edge count**
species (except parr)	Classification (as species); disambiguation (many species level names also genus level in other context)	Hub - outer node	High (top fifth or top half of range)	Low (0 - 0.04)	Under 3
genus; common; general	Identification of central nodes of hubs - classification level.	Hub - central node	Low (bottom third of range)	High (0.5-1)	3 or over
common; general (collective and life-stage); family; order	Linking node between different parts of the graph (classification)	Link node (not very well defined)	High (top fifth)	High (0.5-1)	3 or over

- Filter for frequency to eliminate potentially spurious results
- Filter for salience to then fine-tune some results
- Use selective filtering upon identification of a spurious result or an overly general result that obscures other results
- Use hubs to identify variants which serve to group other variants under their meaning
- Use corpus analysis and concordance to check validity of any infrequent hits and know whether to discard or not
- Focus on infrequent results by removing the frequent ones and then use salience to look at the strongest links
- Highly linked nodes indicate, in graphs which have reasonable frequencies, the accepted and most commonly used forms of the term
- Nodes which are separated from the rest of the graph may be specifically collocated with the other node they are linked with, may have a different meaning to the highly linked nodes, or may not be in common usage in the corpus in question but could be elsewhere. Worth further investigation.

2: Data representation and analysis techniques

An important aspect of the thesis has been investigating the choices relating to data representation. This has taken two different focuses: data framing which has focused on the multiple nature of scientific nomenclature and vernacular terms, and how these can be framed as single or multiple units.

Corpus pre-processing: data representation considerations

- Original no processing
    1. Emphasis of each step of the taxonomic hierarchy (species, genus, family, etc.)
    2. Links in the graph through shared species level or below terms (such as Linnaeus). Not present in the unified corpus.
    3. Accentuation of role of genus in grouping
    4. Comparatively more relations identified overall
- Unified term as one

1. Emphasis of taxonomic entity (species, taxon concept) as a unit
2. Accentuation of role of vernacular variant in grouping
3. Comparatively fewer relations identified overall

Data representation: frequency and salience filtering The filtering methods used in the analysis were frequency and salience. The profiling studies revealed the following about the different relations highlighted by each filter and how they could be used in combination to tailor the results.

Frequency

- Frequent, highly connected terms are highlighted
- These will usually equate to accepted names, or vernacular variants
- Initially very high numbers of relations identified, which quickly slope off making it an easier measure for large amounts of data
- Can use in combination with salience to reduce total number of relations but focus more specifically on some other, less frequent, occurrences in the text

Salience

- Infrequent, less connected terms are highlighted
- These will usually equate to non-official variants, species not a main focus of the test corpora, scientific variants

3: Profiling meaning from the analyses

Comparative dispersion patterns (dispersion and frequency analysis)

1. Frequent and broadly distributed
    - These are the names which are in most common use (so one would expect these to be the accepted names, and most usual/recognised common names for these species
    - If there is variation from that then it indicates either discrepancies of opinion as regards the naming, or changes or specific domains in which different terms might be used (this could be investigated with corpora which control for the variable required)
2. Frequent and specific

- Indication of author-, time-, domain-specific usage of a term, depending on where the term is concentrated

3. Infrequent and specific
    - Less used or not accepted names (this may be in the context in which the corpus is focusing)
    - As regards common names, either ones used in specific areas or for specific purposes
    - These can be time-, author-, domain-specific so this should be borne in mind
    - Indicate former names, which are now outdated (unless looking at historical corpora)
4. Scientific names
    - Frequent, but common names still more frequent (indicate a trend to mention the scientific name but then talk general using the common name)
5. Accepted scientific name
    - Most common usage of scientific names
    - Well-distributed across corpora of different types
    - Highly connected (node hub, or at least with many incoming/outgoing edges) in graph representations
6. Common name variants
    - Expect to be widely used from what has been seen in the test corpora
    - From the test corpora it would seem that overall ratio about 80:20 to common names in comparison with scientific nomenclature
    - Most used common name to be most highly distributed
    - General use common name highly connected (node hub, or at least with many incoming/outgoing edges)
    - Where connected to various scientific nomenclature being a hub can be understood as more general than the species level names used to represent
    - Common names used in specific settings to co-occur with a specific scientific variant or in a specifically in a certain type of document
    - Where used in specific settings in the network relation graphs will only have links to these variants

7. Specific terms – domain, time, language specific (for example)
    - Where name changes for a species have taken place, this is often indicated in corpora by multiple variants occurring at the same time (in the same documents)
    - Previous accepted names most likely to be found in references section of academic corpora
    - In non-academic corpora more variety of both scientific and common names expected (less consistency)

Relation network graphs: node and edge identification

1. Scientific name variants
    - Accepted name variants, which constitute a focus of the corpus, are more connected
    - Any scientific variant not a focus of the corpus will not be well connected
    - Infrequent variants may only be linked to specific terms
    - Species-level names may be surrounding hub nodes of higher hierarchy (common names or other)
    - High frequency scientific names (ones in common use in the corpora) should be identified using frequency
    - Rarer, less commonly used variants can be found using salience
2. Common name variants
    - Frequently used common name variants likely to be hubs
    - Often link to multiple scientific names
    - Better to look for through frequency
3. More general terms
    - Usually source node, not target node
    - Tend to be at the top of the hierarchy, having exclusively outgoing edges
    - Hub nodes
4. More specific terms
    - Usually target node, not source node
    - Tend to surround hub nodes

- Tend to have incoming, not outgoing edges

The following is a breakdown of the research findings:

- Method development
    1. Data manipulation: filter parameters
    2. Data manipulation: data framing
    3. Graph representation: relation network graphs
- Method formalisation and application
    1. Nomenclature profiling studies and guidelines

8.2 Data representation: domain contribution

The previous sections have set out how different parameters and the techniques developed in the thesis can be used to represent data extracted from test corpora in relation to nomenclature mentions to create a profile for a specific nomenclature term, which can be manipulated in different ways to emphasise specific features of this term. Many of these findings are directly concerned with data representation: how to present data, how different filters or processes can change the formulation or presentation of the data, and how network graphs can be used to produce visualisations of this data. This can be applied directly to the profiling of nomenclature terms in different corpora as is, through semi-automatic means. Future analysis, using the existing methods, could be used to look at patterns in nomenclature usage to compare and contrast different domains, authors, geographies and times. These are all areas that were highlighted in the literature review and reinforced in the focus group (see Chapter 7 as areas of interest for the area of biodiversity. This represents a practical application in learning more about any patterns of usage within this domain, or the many fields that make up this domain to analyse differences and similarities in usage, where there may be particular contradictions. Specifically this has the application of being able to compare ontological structures extracted from real-life text data against existing ontological resources or other text data, to compare for consistency or divergence in meaning as a validation tool, which will be explored further in the knowledge integration section.

The focus group discussion and the conversation with P4 also indicated that my research could be particularly applicable in that relating to the visualisation of complex data. While participants felt that experienced experts do not have problems with the nomenclature in their area, this sort of visualisation could be useful for students or experts from other domains to make them more aware of the different variants in usage, their patterns of usage and any peculiarities in meaning. This could be particularly relevant to avoid the issues of thinking that name variants are different species, for example. The same could be said about vernacular variants and providing clarity as to the meaning and usage of specific variants, again for people new to the domain or non-experts.

The findings and in relation to knowledge integration will be discussed in the next section, as well as possible applications.

8.3 Data and knowledge integration: domain contribution

The previous sections looked at the methods developed, the first steps to formalising these methods and how these methods can be used to profile nomenclature references within corpora to identify patterns of usage and how this contributes to the field of data representation. This section focuses on the application of these methods and their contribution to the field of data and knowledge integration.

The importance of being able to evaluate ontologies against corpora against the aforementioned criteria is two-fold. In the absence of a gold standard ontology or other knowledge representation resource [66], there are always queries as to the completeness and accuracy of any existing resource. This is further emphasised by the ambiguity inherent in the biological taxonomy and scientific nomenclature as described at length in the literature review [192]. To add to this argument is the need to make a choice in the representation or classification of information for most single hierarchy resources [175]. Therefore the method developed was applied to evaluate the knowledge representation resources chosen in response to the gaps identified. Ontologies are used for integration and search purposes, so this method represents the first steps on the path towards the development of an integrated method which can be used to decide whether a corpus or corpora are suitable for integration, whether a specific ontology could be used to map said corpus, as well as highlight any logical issues or gaps in chosen taxonomic resources.

The method developed can be used to identify a suitable or best fit match of a corpus against the chosen taxonomic resources through calculating the % match and coverage of corpus data. This could then be used to choose a suitable resource, according to the purpose of the mapping.

While the identification of similarities, gaps and contradictions serve in their own right to profile a specific nomenclature term in a corpus, they can also be used to identify potential gaps or nuances of meaning and highlight potential issues in the integration of corpora using specific resources. In the field of knowledge integration it could be used to apply to identify gaps in taxonomic resources, areas of contention for integration or areas where further work needs to be done.

The method developed in this research is relevant to data integration and knowledge representation because of the way it can highlight nuances of meaning, and identify where there are contradictions between the source data and either an ontology or other corpora, in the way it could be used to prevent the inappropriate imposition of ontologies onto data. This method can be used as is to provide information about the percentage match between a certain ontology and corpus, or multiple corpora, and produce an evaluation such as the one produced in this thesis in Chapter 6. The method here could also be used as a basis for the development of an automatic process to identify suitable terminologies or ontologies for a specific data set, or to evaluate the suitability of a particular ontology to map certain data. It could also be used to assess the cross-compatibility of multiple datasets for integration or if another approach should be taken.

These methods could also be applied to different areas in which there are multiple ontologies, to check for suitable resources for the mapping or integration of data onto them and, as with the data representation, to check for gaps or contradictions between the data and the ontological resources.

A first application could be to build further corpora in the area of freshwater fish, or another species, and repeat the nomenclature profiling studies to test on a different or the same ontology. The corpora could be built in a way to allow for automatic processing of domain, geography or publication date to provide further nuance to the analysis.

8.4 Lexicography and terminology: domain contribution

The research described in this thesis constitutes a contribution to the domain of lexicography as regards our understanding of Word Sketches and how they can be applied. The development of this method firstly demonstrated that it was possible to adapt Word Sketches, the lexicographic dictionary entry summary feature of Sketch Engine, to produce such a graphic representation of links between nomenclature references. As described in the literature review, taking a linguistic approach to this sort of task is far from new. Hearst patterns [90] have been used frequently in tasks such as ontology creation/automation [12], as well as other linguistic patterns such as those seen in the Ecolexicon project [128]. The representation of collocations in the form of graphs was first proposed by Phillips in 1985, but they were manual and it is only recently that they have gained more importance using computer manipulations [29]. Superficially these are similar to word maps used for a thesaurus, however collocation graphs and networks work on the basis of associations between words in discourse. My research could be considered to straddle these two areas because it selects only specific entities for analysis which gives it a more thesaurus-like quality, particularly given the subject matter. However, the graphs/networks are extracted on the basis of the behaviour of these words in discourse.

This is the first time, to my knowledge, that Word Sketches have been adapted in this way. It is also the first time, to my knowledge, that these techniques have been applied to produce comparison characterisations of scientific nomenclature usage between existing knowledge representation resources or cross-corpora. The significance of the way Word Sketches have been adapted in this instance can be considered from three perspectives: the possibility of expanding the application of Word Sketches to tasks beyond lexicography, the possibility of using the methods developed here to investigate terminology stability or change in specialist domains from a lexicographic perspective and also how the results from the research set out in this thesis could be used to inform work in the area of word embeddings in the future. These will all be explored further in the future work section.

8.5 Strengths and limitations of the research

The research possesses a number of strengths and weaknesses, which are explored here. The method evaluation demonstrates the reliability of the method across two corpora, as well as having shown how different parameters can be used to alter the perspective of the same data. This provided support for the argument made in the literature review and throughout the thesis as to the importance of framing information. The manipulation of parameters can be used in future work to highlight specific features of interest or look at specific datasets from a specific perspective. It has also clearly highlighted the importance of common name variants in extending research beyond names to interactions, which could be of vital importance.

The evaluation consisted of both a technical and an expert aspect, which served to provide validity not only from a technical perspective but also gave weight to the argument as to the usefulness of the method developed, as well as clarifying the conclusions drawn in the cases of ambiguity in the results. However, the research is only the beginning. Only two corpora were analysed, they are from similar domains (although one with a purely academic basis), and contained very little metadata. In future research, to be able to generalise better, it would be necessary to test on more disparate corpora. The addition of more metadata, such as publication date, location, document length, author, language and annotate the different sections of each document would result in a more fine-grained and more useful analysis, as was commented in the focus group.

Continuing on this vein, as this research aimed at developing the method in a first instance, the relations used were simplified to only parent-child relations. The granularity of the relations identified also require more work to try to identify sibling-sibling relations and avoid so much grouping of like items. This could be a focus of future work.

8.6 Inconclusive or surprising results

The data was limited and so no conclusions could be drawn on some occasions. Observations of what was found could be made and suggestions as to the possible meaning. Conversations with experts provided support for my assertions but, as is the case with corpus analysis, only what is there can be analysed, what is not cannot. P4, the fish expert, was not surprised by any of the results.

8.7 Further research

There are a number of possible avenues for further research arising from this thesis. Firstly, there are various avenues that could be explored in developing the method. Various applications were mentioned in the above sections which could be performed using the method as developed in this thesis, but to improve the output and make the method more usable for a wider population a number of steps could be taken. Further research to increase the granularity of the relations identified would be one route to take, as would taking steps to further automate the process to make it accessible to a wider audience. Using corpora that have been built with temporal and thematic metadata would be an essential part of any future work, which was specifically mentioned in the focus group as an important step to making the method applicable to the scientific community. Expanding the research into other domains within biodiversity or beyond would be another important next step to test cross-applicability of the method.

Besides this, research could now move forward to explore the extraction of trophic interactions within a dataset in more detail, taking up the aim that was dropped in the pilot stage of the research. This could be used to further explore the links between vernacular variants and their scientific counterparts and how this information can be leveraged to facilitate processing of descriptions involving interactions in natural text, as mentioned by P3 and the difficulties there.

From the perspective of how the usage of Word Sketches could expand in the future, the research presents an opportunity for adapting the features of Sketch Engine to be used in specific domains outside that of lexicography and linguistics, particularly in the areas of data representation and integration, as highlighted above.

The research in this thesis also represents a contribution to lexicography in itself as regards possible applications of the method to terminology mapping for lexicographic purposes. This would require further work to explore the applicability of the method, but differences between word meaning in corpora over time or in the use of terminology from different domains could aid in the creation of lexicons for specific domains or mapping change in this way. While collocation graphs are already used in terminology and lexicography, this method provides a formalised process to follow in relation to the comparison of similarities and differences that does not as yet exist in these areas.

The possibility of applying this method to inter-lingual terminology usage, both in the area of scientific nomenclature (in which it could be used to compare scientific and vernacular variant synonymy across geographic and linguistic borders) and also beyond this in terminology on a broader scale. In the area of biodiversity, the geographical element to varied usage came up repeatedly and was reconfirmed by all the different participants at one time or another, so this would seems to be an important avenue to follow. As regards terminology in general, while there are collocation graphs [29] already being employed in linguistic analysis, to the researcher's knowledge there has been no research into the comparison of these across languages, despite this being a big question in translation studies, the consideration of true inter-lingual synonymy.

Finally, throughout the course of the thesis, the similarity between Word Sketches and word embeddings has become apparent to me. Both are based on specifically defined features of word use in the contexts in which they appear. While the features in word embeddings are defined in their thousands by computers, in the case of Word Sketches they focus on collocational and grammatical behaviour to define said position. Word Sketches obviously do not present a word as a single vector within a space, however, the way that the collocations are separated into grammatical relations helps to provide a more nuanced understanding of the different parts of what makes up this characterisation to some extent, and as we have seen, can be used to define where it fits in a graph. The adaptation of Word Sketches in this research has demonstrated how they can be adapted to take account of different semantic groups to focus efforts on a specific area to be modelled. The graphic visualisation helps by setting out what the relations between each term or unit are in a visual, and therefore more integrated manner. There may be possibilities in the future to consider the work set out in this thesis in taking a different perspective towards the creation or use of word embeddings that may help in potentiating their use in more specific fields, which requires smaller amounts of data, or in fields where there may be inherently ambiguous usage of terms.

8.8 Conclusions

This research aimed to aimed to “employ computational lexicography and natural language processing techniques to identify, extract and group nomenclature according to its usage in the biodiversity literature and use contrasting corpora and existing knowledge representation structures to perform a systematic empirical analysis of these conceptualisations". The aim was created to respond to the lack of any research into the actual usage of nomenclature in the biodiversity literature, despite the multiple issues identified in the literature review relating to nomenclature usage, which arise from the lack of a gold standard taxonomy and recognition that there is no one agreed stance as to the biological taxonomy. A lexicographic approach was taken because of the empirical nature of this approach, and the way that it aims specifically to look at word use in context, to emphasise usage above tradition or expectation. As a result, the work in this thesis culminated in a method by which to empirically extract ontological structures from text data for validation (which can be applied to cross-corporal comparisons, or ontology-corpus comparisons as required).

The research followed a design science model in which different phases were used to explore different aspects of the problem and iteratively feed back in to develop the method described above. Phases 0 and 1 focused on understanding the data, the approach and possibilities of the Word Sketches and different ways of presenting, or framing, the data in order to respond to Objectives 1 and 2 of the research. The validation and evaluation of this method responded to Objective 3. The evaluation was also used to demonstrate some of the differences between the frequency and salience parameters in the types of profiles each of these filters highlighted. The application stage applied these methods to formulate guidelines by which to interpret the patterns identified in the data both using traditional corpus analysis techniques and the relation network graphs produced from the Word Sketch information. The comparison between taxonomic resources demonstrated how the same data can be presented in different ways.

Through the meeting of the research objectives, the research resulted in a method that permits the extraction of an ontological structure from text data, using empirical evidence. This method enables comparison of the structure extracted from the text data against existing resources to validate such resources and also compare structures extracted from different text data corpora to look for cross-corporal stability or lack thereof. It was particularly useful in identifying variations in vernacular variant usage, both for disambiguation purposes and in contrast with some categorisations of the respected ontological resources chosen as a focus of the research.This method can now be applied to further corpora within the areas of biodiversity to explore the area further, particularly where there are questions or complications as to the usage of vernacular name variants. This should be done with a mind on various controlling factors such as domain, geography, author and time.

This research demonstrates a new application of Word Sketches beyond lexicography, which opens up a number of different possibilities for future work. Both the main body of the research and the initial exploration into trophic interaction demonstrated the high prevalence of usage of vernacular name variants in the domain of freshwater fish and the relevance of the links between these and scientific names, which should be explored in future work to possibly make moves forward in natural language processing of these subjects. The research also demonstrates how the use of graph visualisations can be used to disambiguate data (not new, but just in this specific context), and the usefulness of this within the area of biodiversity.

The method developed in this thesis can be applied to a number of areas, such as mapping of corpora (text data) to specific resources, the evaluation of suitability of one resource or other for a specific corpus, or to measure the level agreement in term usage (which is an indication of agreement in conceptual meaning) across corpora or between an ontology and text data. Future application should if possible rely on corpora delimited for time, author, geography, with metadata relating to specific sections of the document to allow for a more granular analysis of differences.

Possibilities for future work are multiple. The main focus of the research was the development of the method, although preliminary application was used to demonstrate the functionality and evaluate the appropriateness of the method. Therefore, the continued development of the method would be an ideal first step. This would mean that various features of the method could be improved, such as relation granularity, further automation of the method to make it more practical, as well as adapt the method for broader applications. Further work could be performed to explore the differences between salience and frequency on different data sets to confirm the results in this thesis and also see if there are inter-domain or inter-lingual differences that may affect the outcomes of these filters.

The method could also be adapted to other domains entirely, with specific terminological issues, or for use as a terminological tool within lexicography. These adaptations could focus on monolingual applications or there could be further research into inter-lingual comparisons. This would be a particularly interested avenue to explore as regards vernacular variants in the scientific literature, going on the comments from the focus group and conversation as regards geographical-specificity even when talking within the realm of English.

Chapter 9

Graph booklet

The following sections include all the graphs and network diagrams from the three results chapter (Chapters 4-6) and also from Chapter 7 for easy reference to the main body of the text while consulting. Graphs are in colour throughout the booklet, so if viewing in black and white this will affect your ability to read the graphs effectively. As regards the Cytoscape relation network graphs, all the Cytoscape files are available in the Appendices folders referenced to the specific phase of the research. This will allow anyone interested to look at the data themselves and see how the relation network graph images were extracted.

9.1 Chapter 4

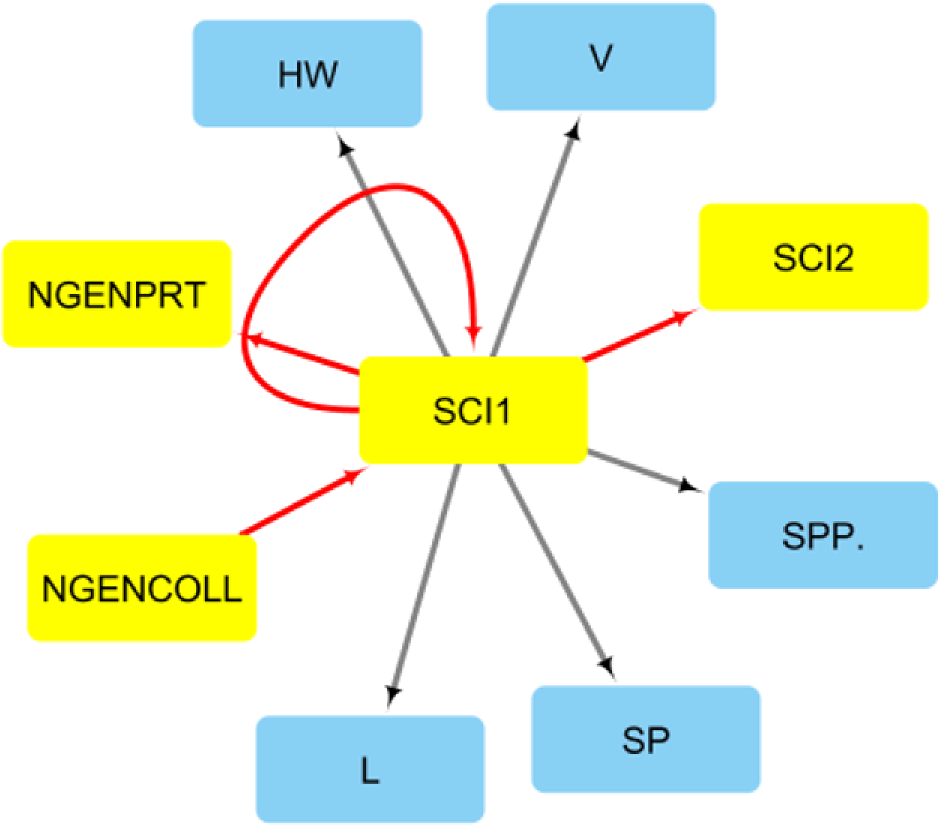


Figure 9.1: Simple graph showing hierarchical relations identified by class [in colour]

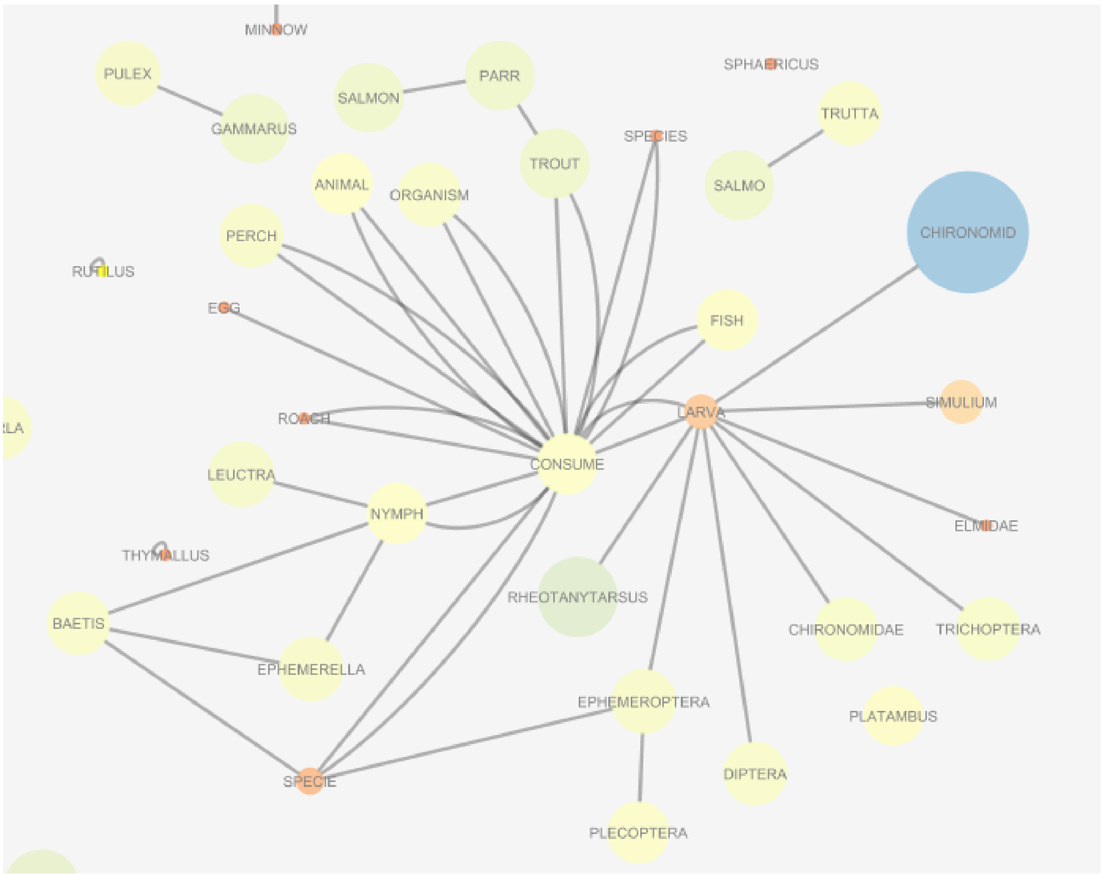


Figure 9.2: Zenodo corpus: demonstration of the importance of common and general-type words for trophic interaction extraction. The term consume, which represents the trophic interaction words ini the graph is linked to terms such as trout, nymph, larva, which are in turn linked to scientific nomenclature such as Chironomidae

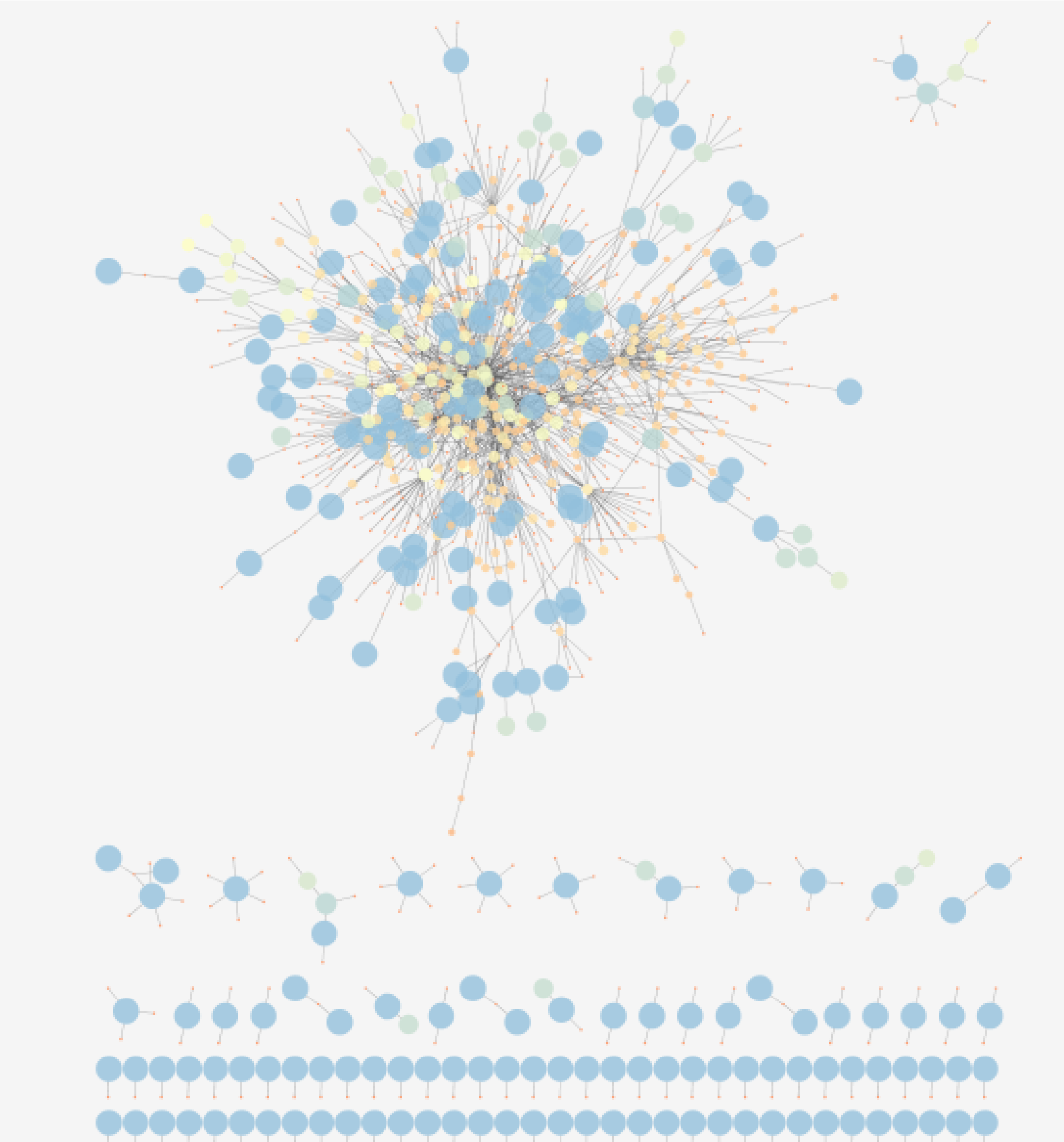


Figure 9.3: Graph visualisation of Word Sketch relations between nomenclature terms with a frequency of hits over 5. Here the number of nodes and relations makes it impossible to read much from the graph.

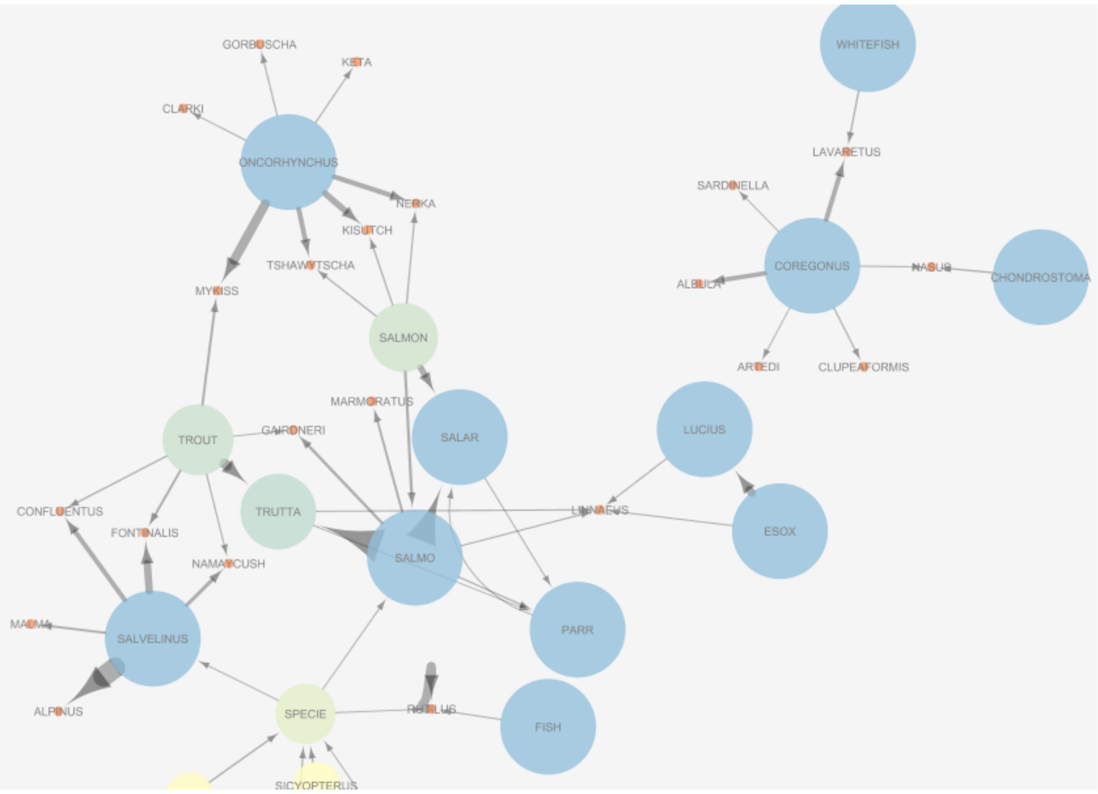


Figure 9.4: Graph visualisation of Word Sketch relations between species mentions with a frequency of hits over 10. Here the filter has reduced the number of nodes and relations so links between the different nodes can be seen, such as the links between Salmo as a parent and trutta and salar as children, with salmon also being the parent of salar and trout being the parent of trutta.

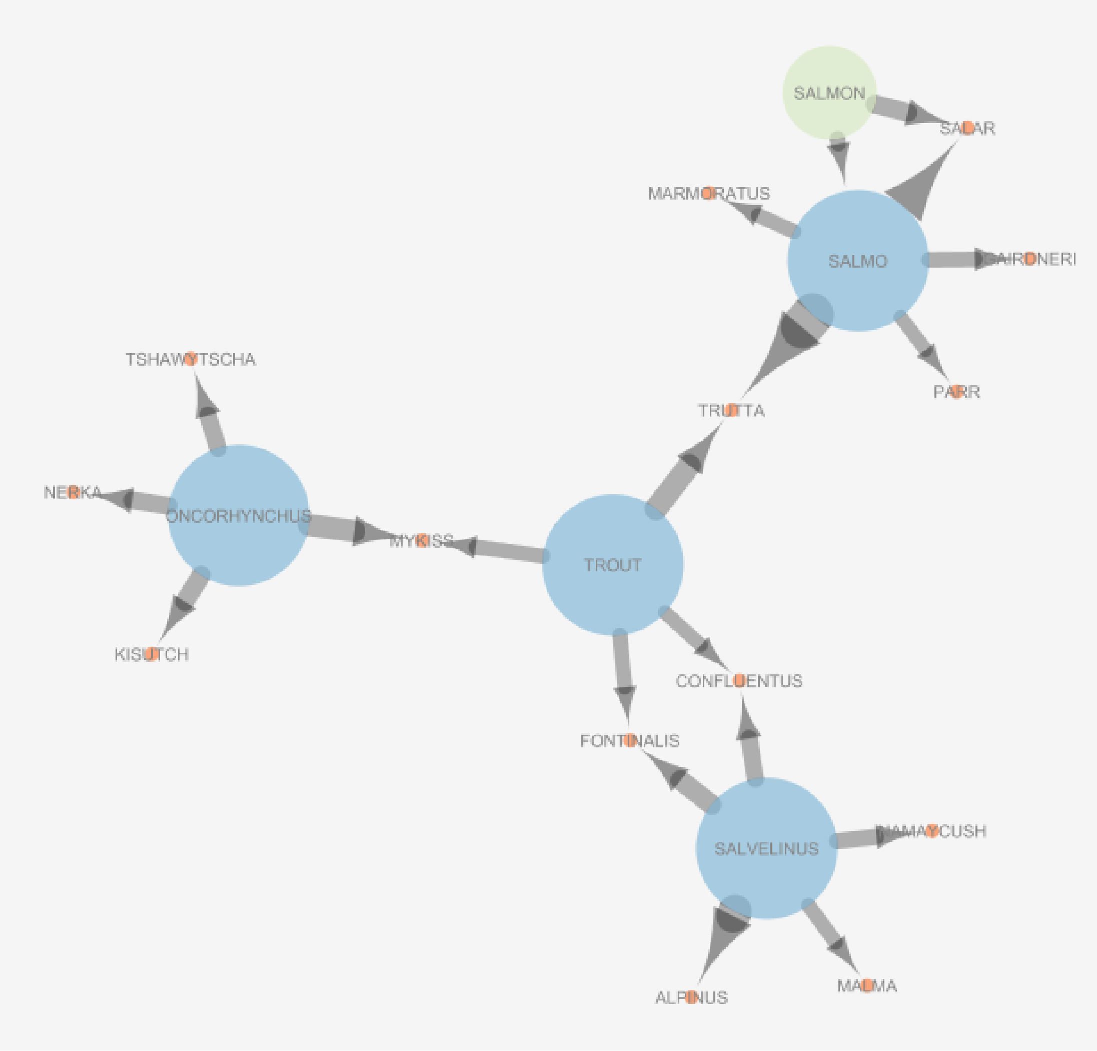


Figure 9.5: Graph visualisation of Word Sketch relations between nomenclature terms with a frequency of hits over 20. The over-20 filter leaves a much clearer picture still, with clear hubs of genera (Salmo, Oncorhynchus and Salvelinus), surrounded by species-level names (such as fontinalis, salar, trutta, kisutch). The arrows show the parent-child relation between genera and species-level terms. Trout is seen as a linking parent term over various species. In this instance the link goes through the species-level term to then link out to the genus.

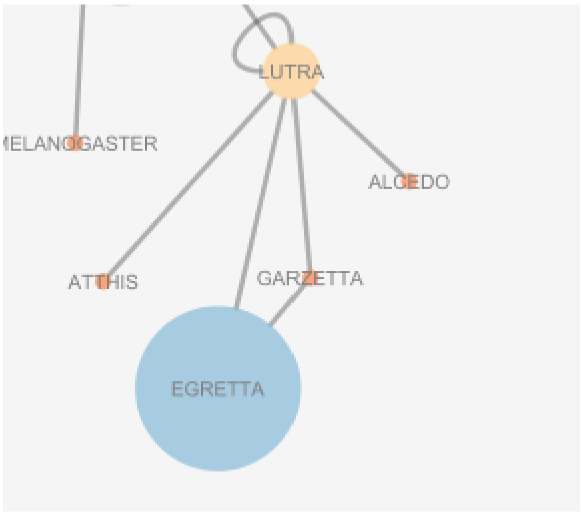


Figure 9.6: JEFF corpus: lutra-egretta relation. Lutra in this instance refers to otters, whereas egretta a species of water bird. This shows how relations between words can be identified because different species share similar habitats and so may be mentioned in the same contexts.

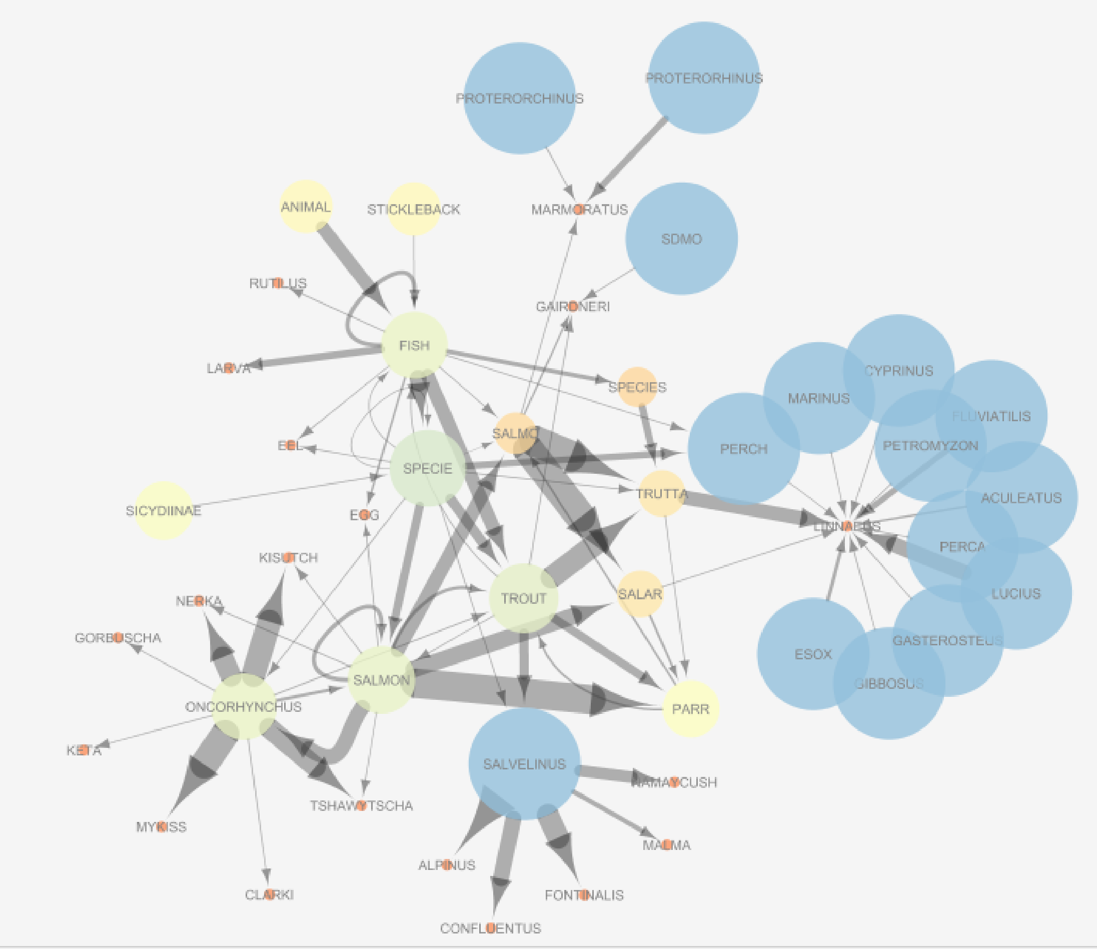


Figure 9.7: Salmo part of the JEFF corpus, salience 9. Trout, salmon and respective genera (Salmo, Oncorhynchus, Salvelinus) nodes link as parents down to species level nodes as children. The child nodes often surround the genus nodes in circles.

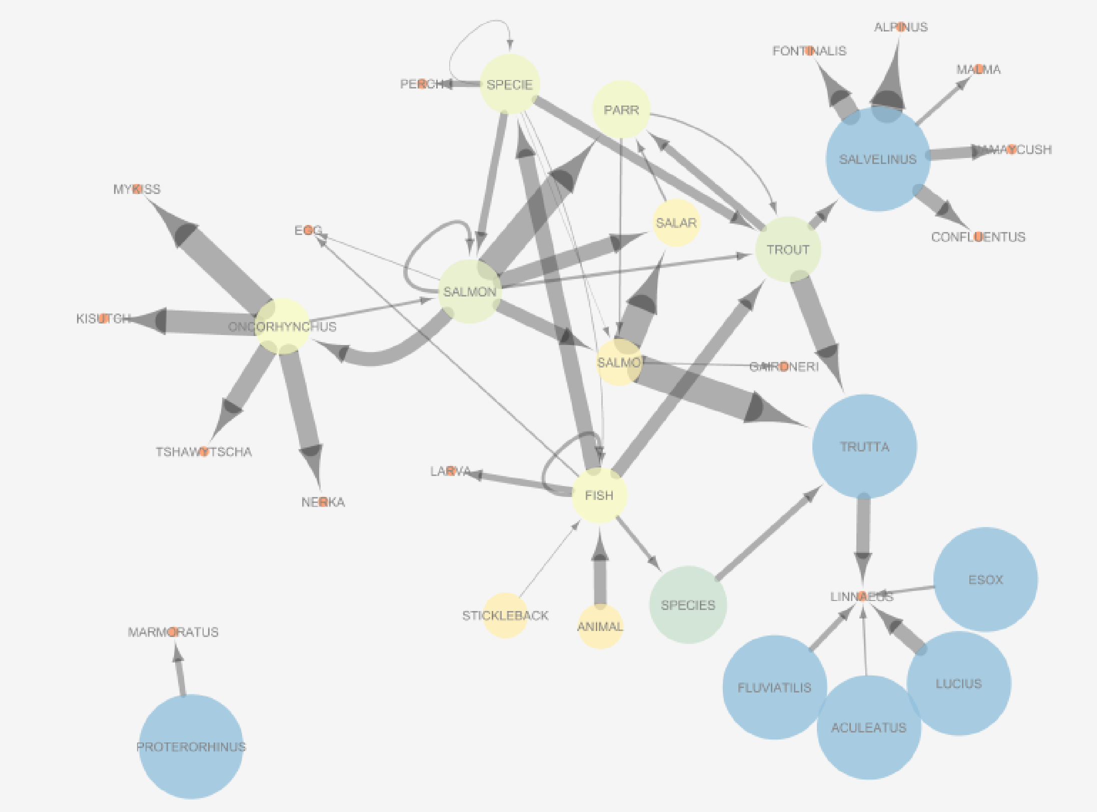


Figure 9.8: Salmo part of the JEFF corpus, salience 10. Links between different nodes are clearer than in the salience 9 filter network graph. Here general terms such as “species” and common names such as “trout” serve as linking nodes that link multiple parts of the graph. Genus nodes such as “Oncorhynchus” and “Salmo” are seen to be the parents of species-level nodes such as “trutta”.

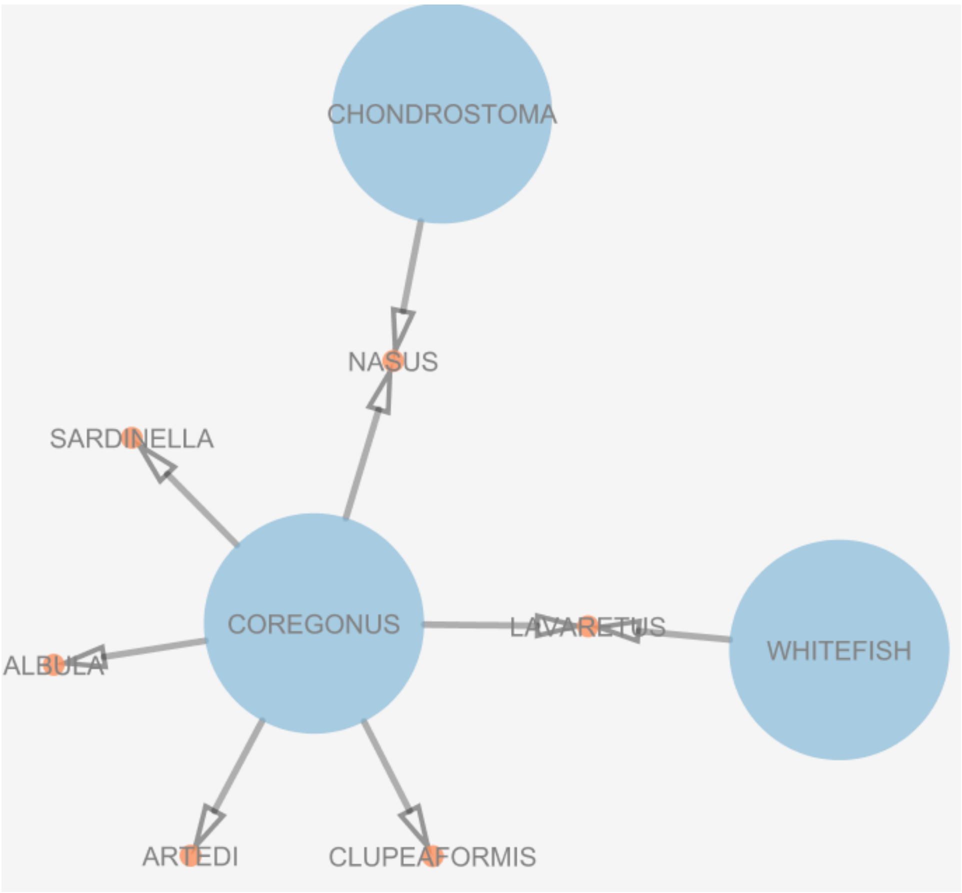


Figure 9.9: Original JEFF corpus: graph visualisation of Coregonus (genus) hub with surrounding species level nodes (hub plus surrounding node equals binomial nomenclature item), filter frequency 10 (higher CC, larger dots)

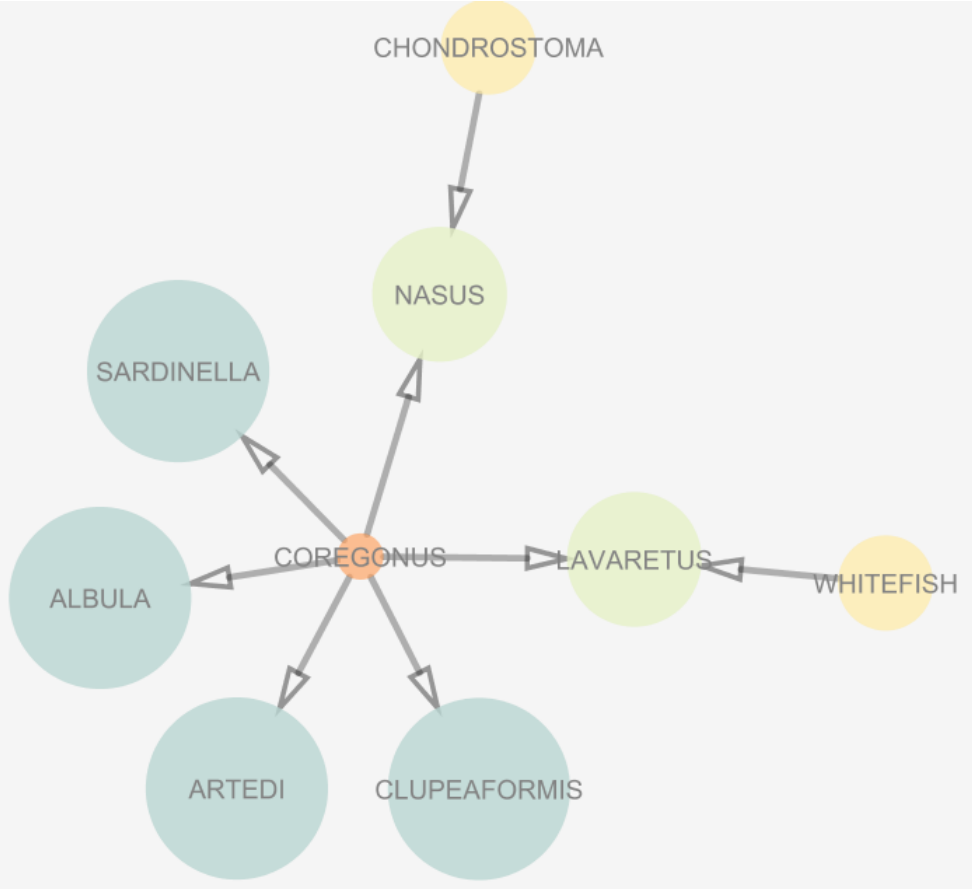


Figure 9.10: Original JEFF corpus: graph visualisation of Coregonus (genus) hub with surrounding species level nodes (hub plus surrounding node equals binomial nomenclature item), filter frequency 10 (higher NC, larger dots)

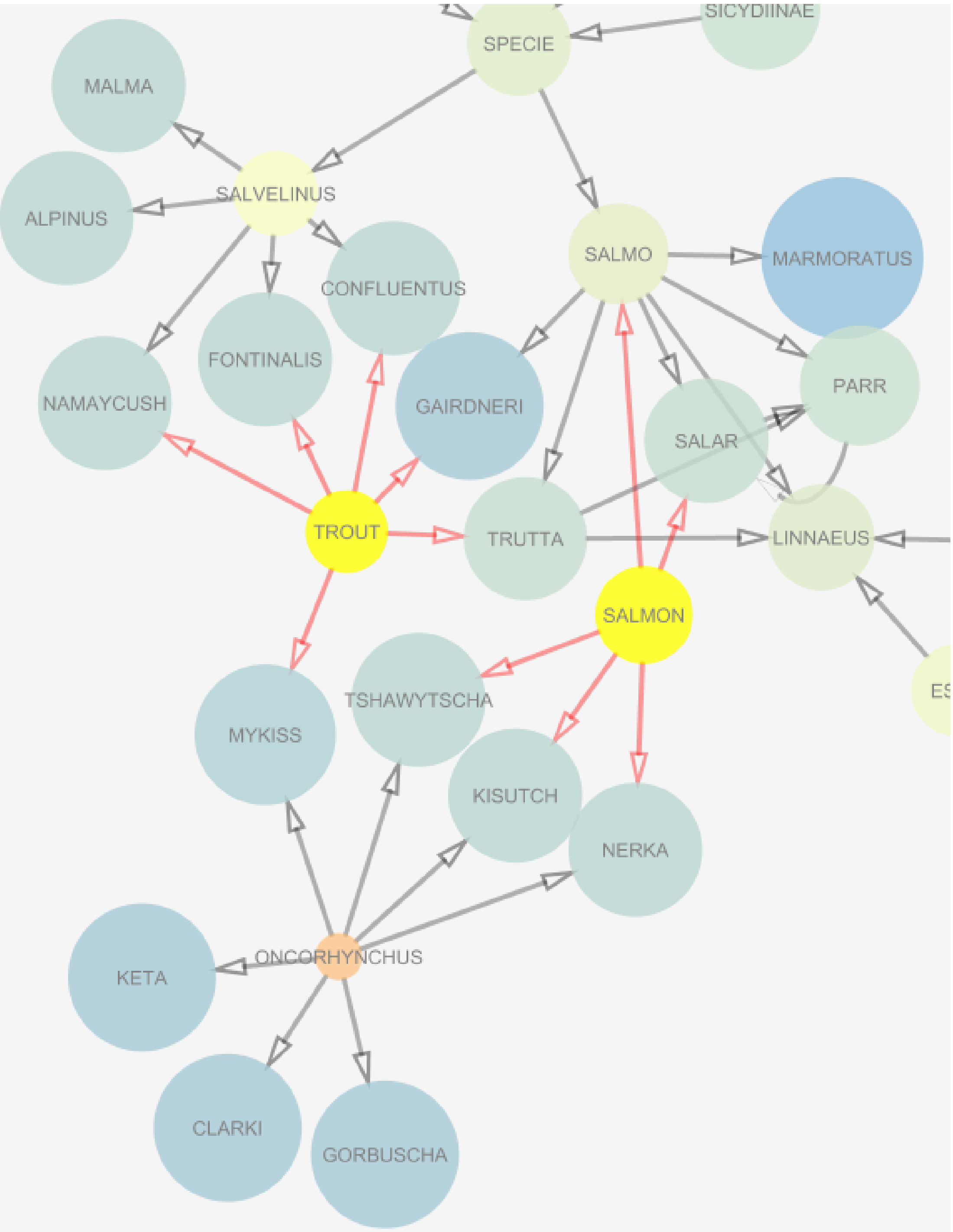


Figure 9.11: Original JEFF corpus: graph visualisation of linking nodes (nodes highlighted in yellow, relations highlighted in red), filter frequency 10 (higher NC, larger dots)

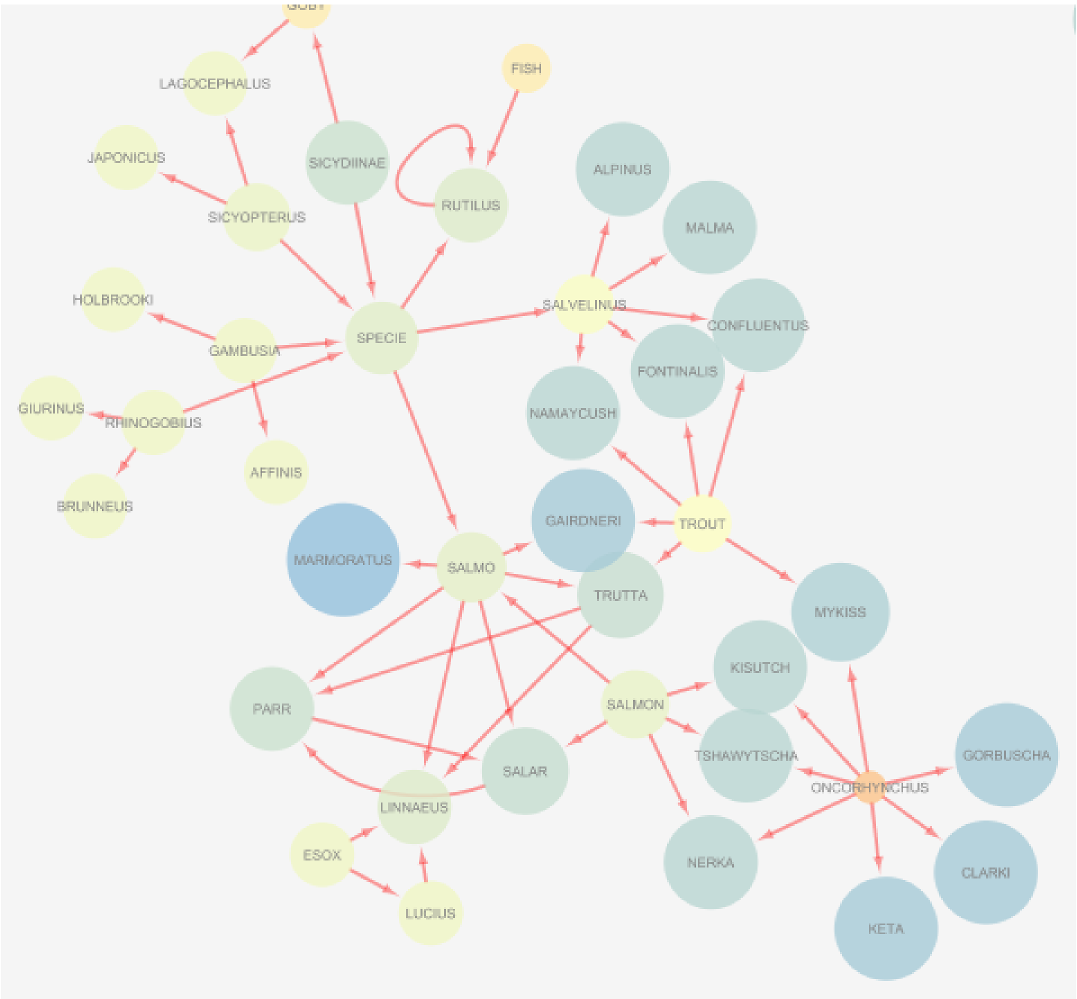


Figure 9.12: Original JEFF corpus: filtered for frequency 10. Characterised by species-level term nodes surrounding genus-level hub nodes. Common terms tend to link different species that share a common name through the lower species-level term nodes (see salmon, trout, Salvelinus, fontinalis, Salmo, trutta, Oncorhynchus, mykiss, nerka)

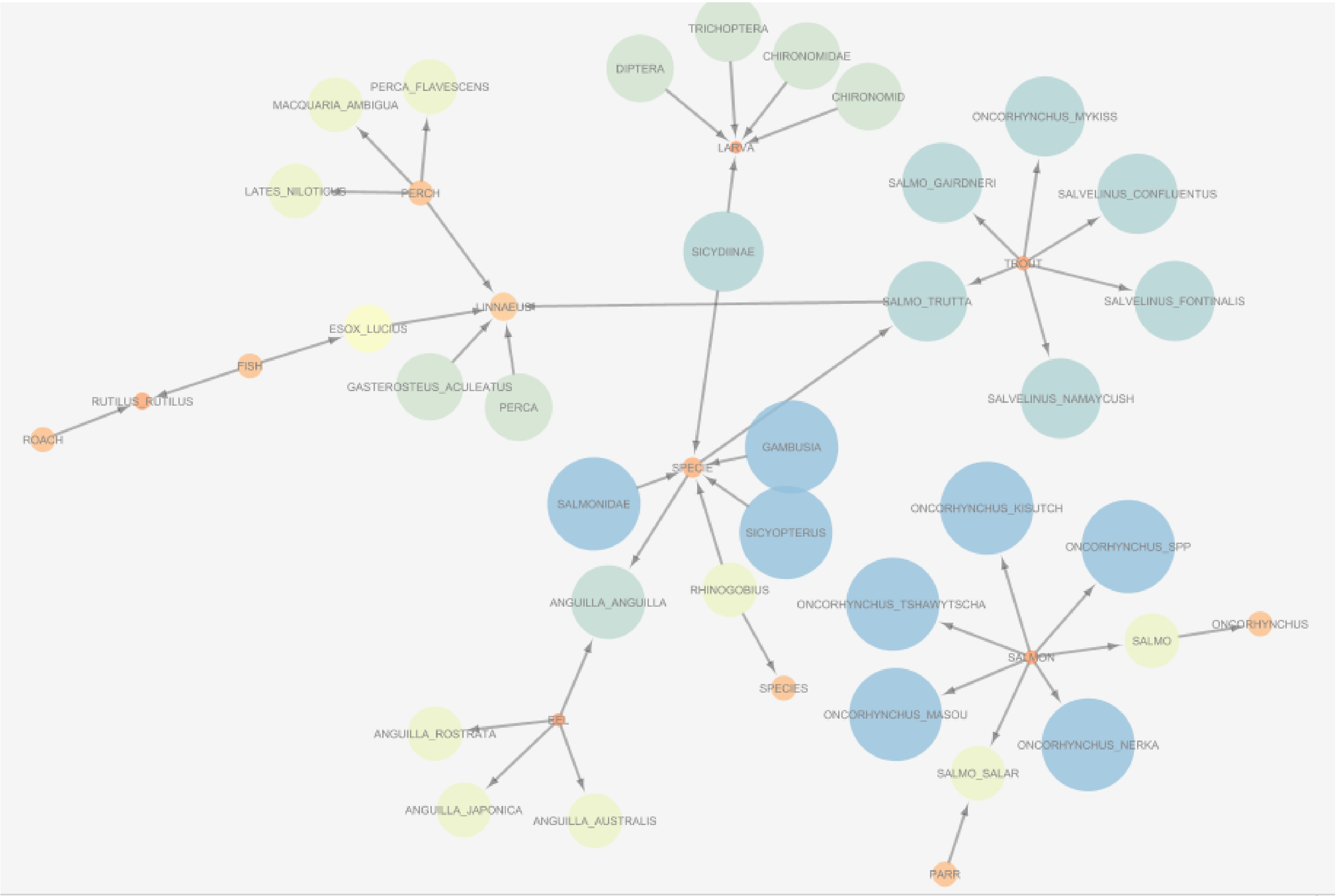


Figure 9.13: JEFF unified corpus: filtered for frequency 5. Characterised by hubs of species (binomial nomenclature) around common variants and general terms - see eel, perch, salmon, trout

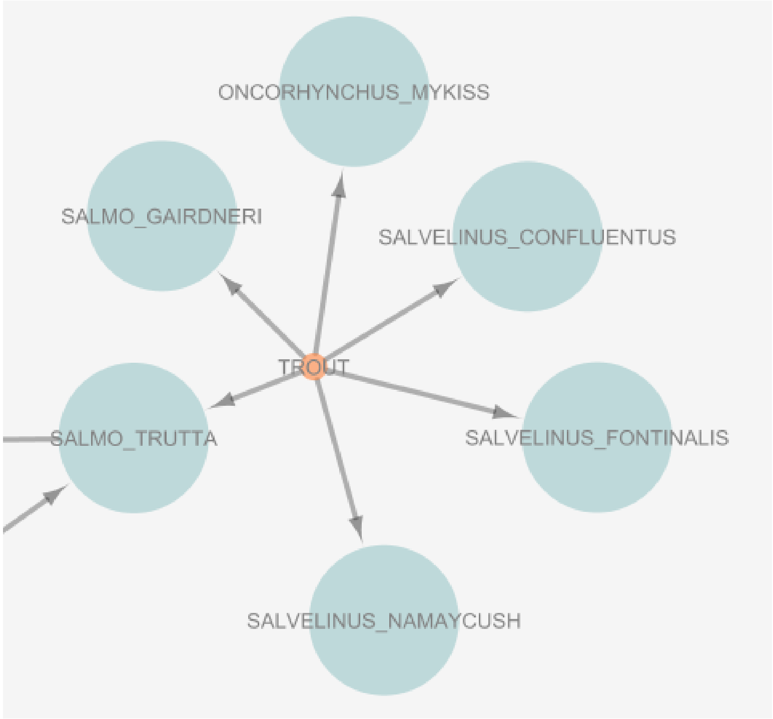


Figure 9.14: Graph visualisation of trout hub in species as unified corpus, filter frequency 5 (higher NC, larger dots)

9.2 Chapter 5

The following graphs relate to that described in the methodology validation and evaluation in Chapter 5.

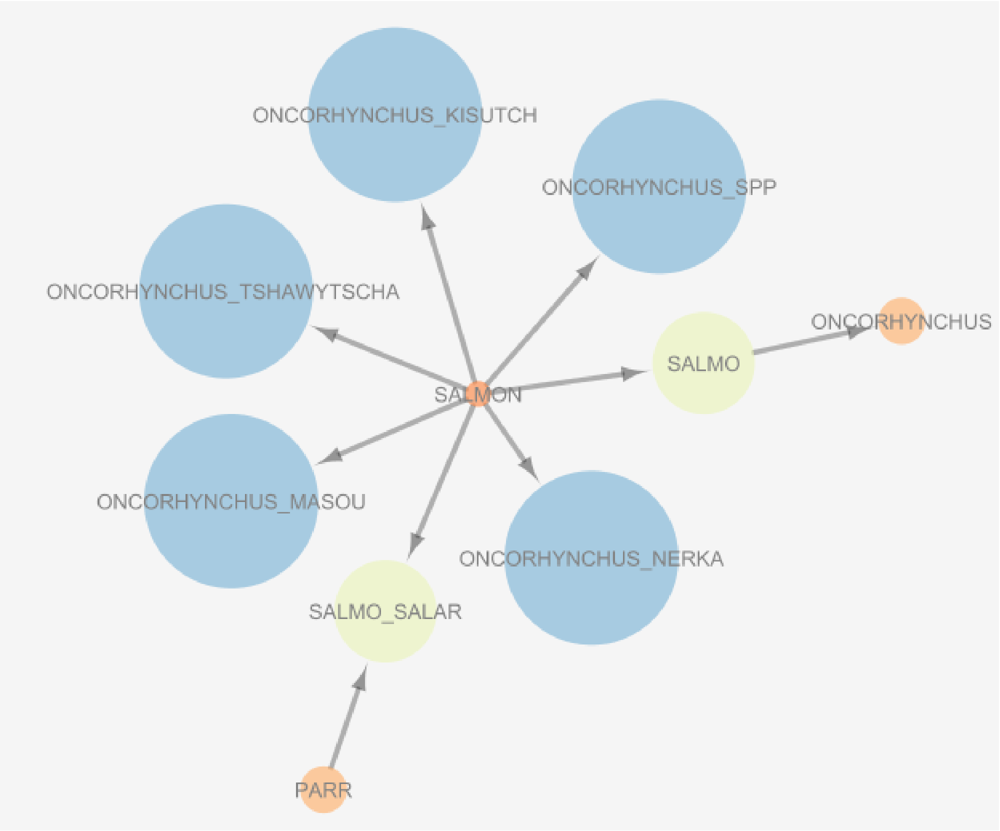


Figure 9.15: Graph visualisation of salmon hub in species as unified corpus, filter frequency 5 (higher NC, larger dots)

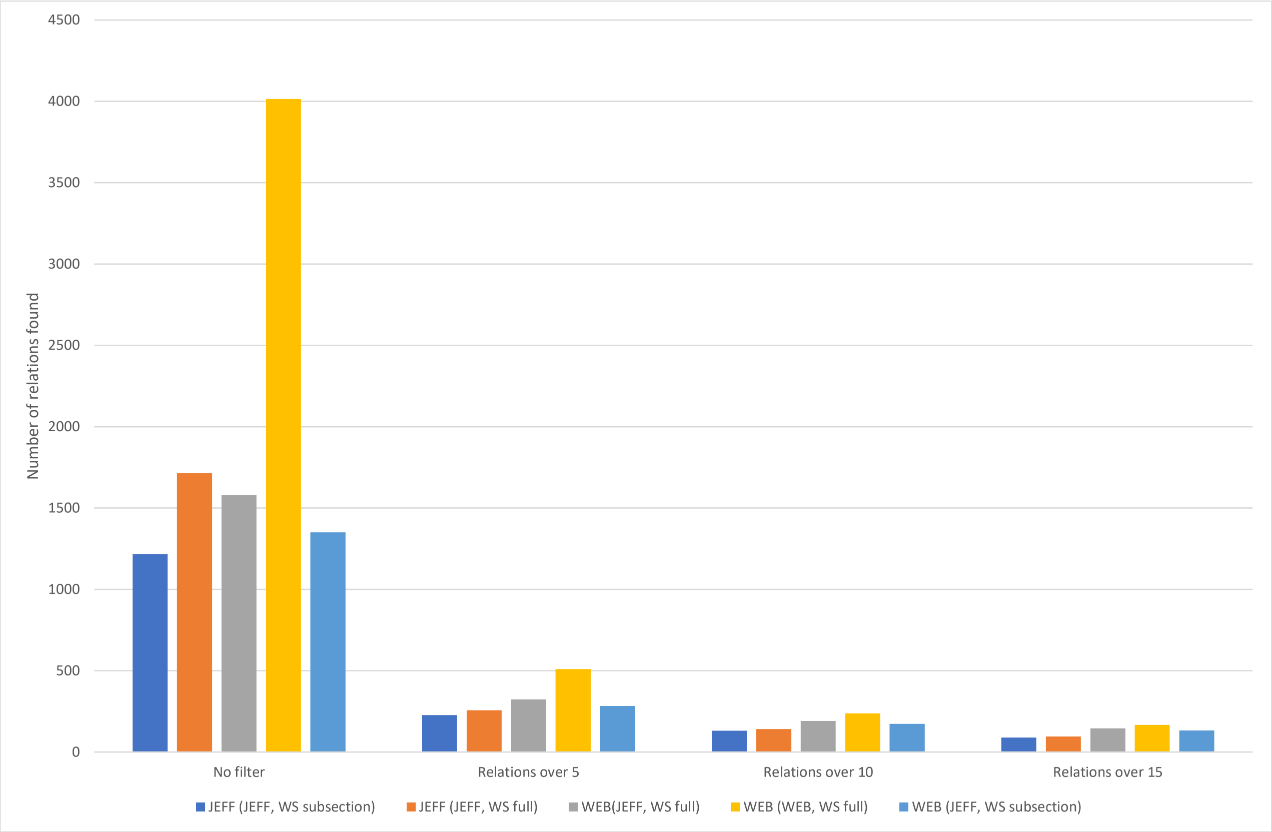


Figure 9.16: Number of nomenclature pairs found in each corpus per annotation/Word Sketch list [in colour]

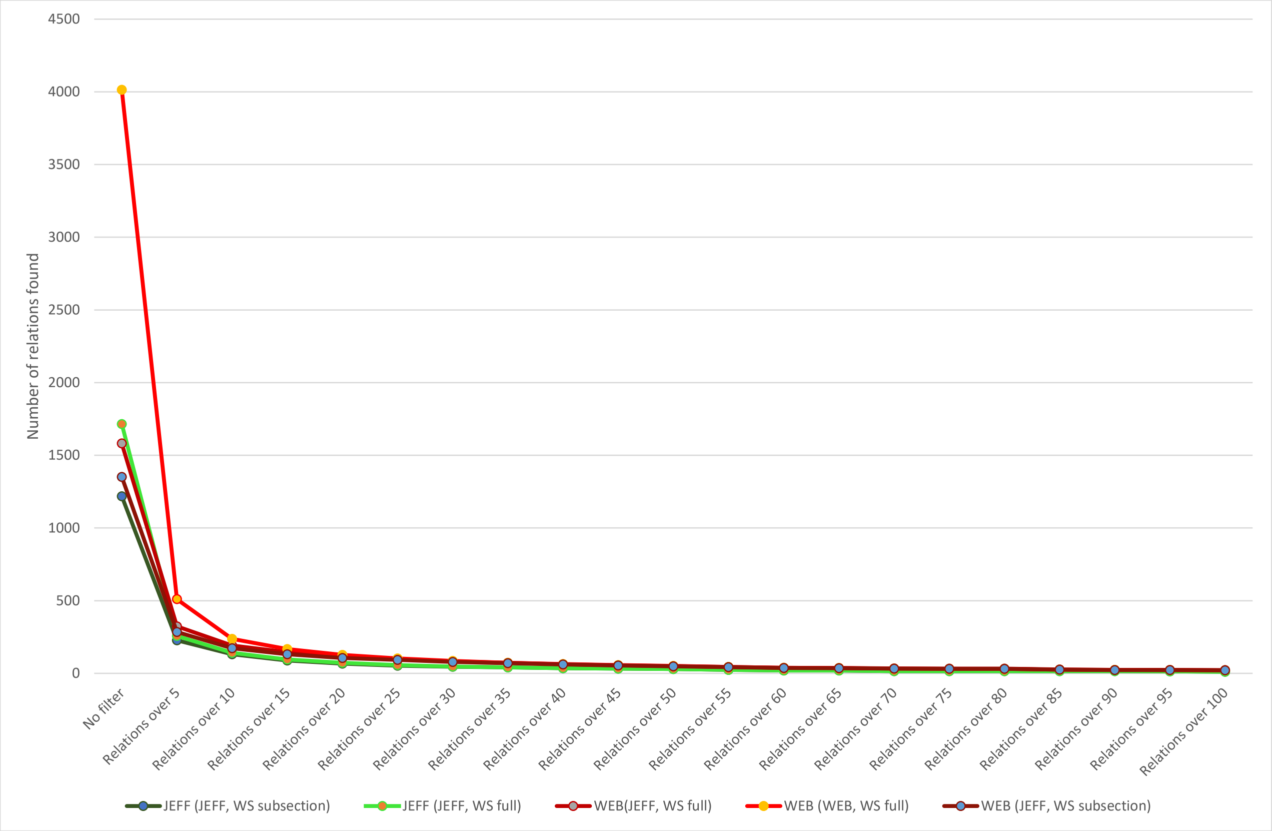


Figure 9.17: Trajectory of number of nomenclature pairs found in each corpus per annotation/Word Sketch list, with increasing frequency threshold [in colour]

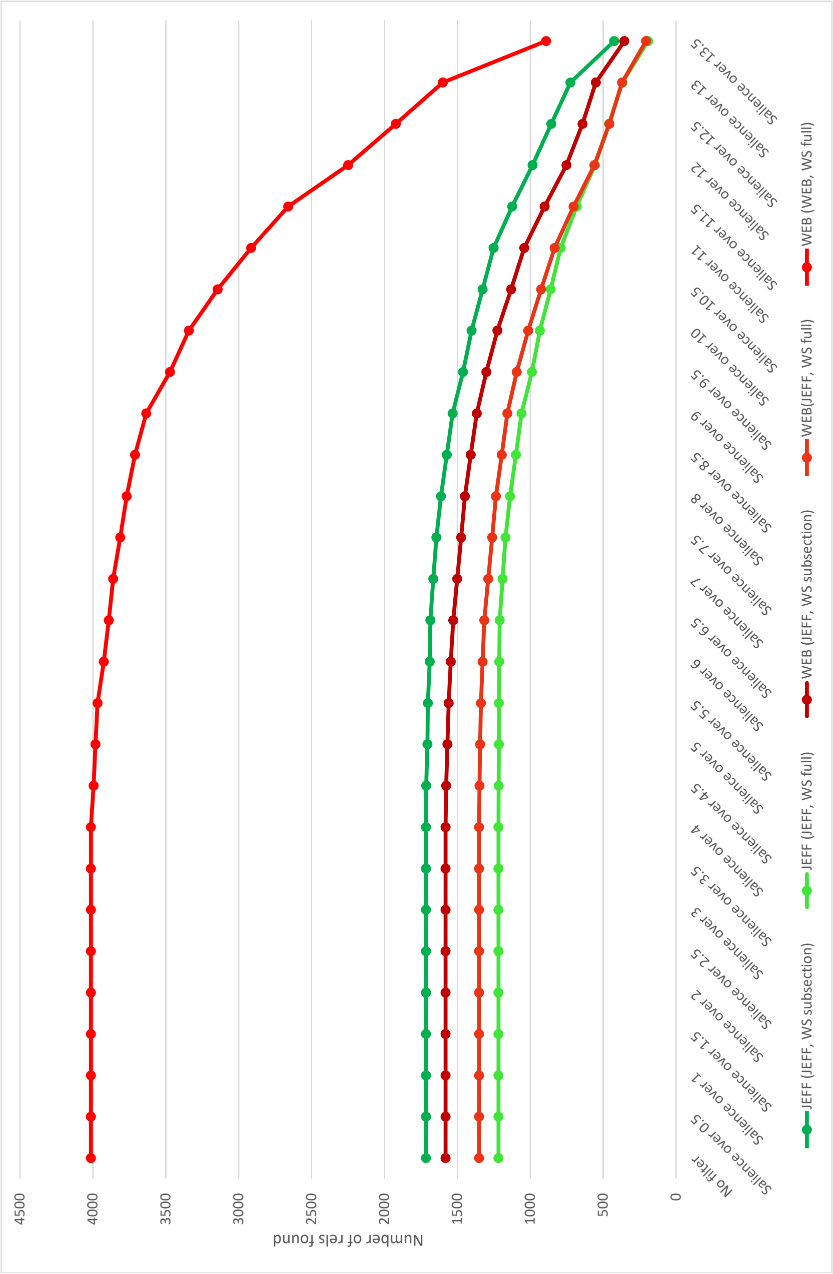


Figure 9.18: Trajectory of number of nomenclature pairs found in each corpus per annotation/Word Sketch list, with increasing salience threshold [in colour]

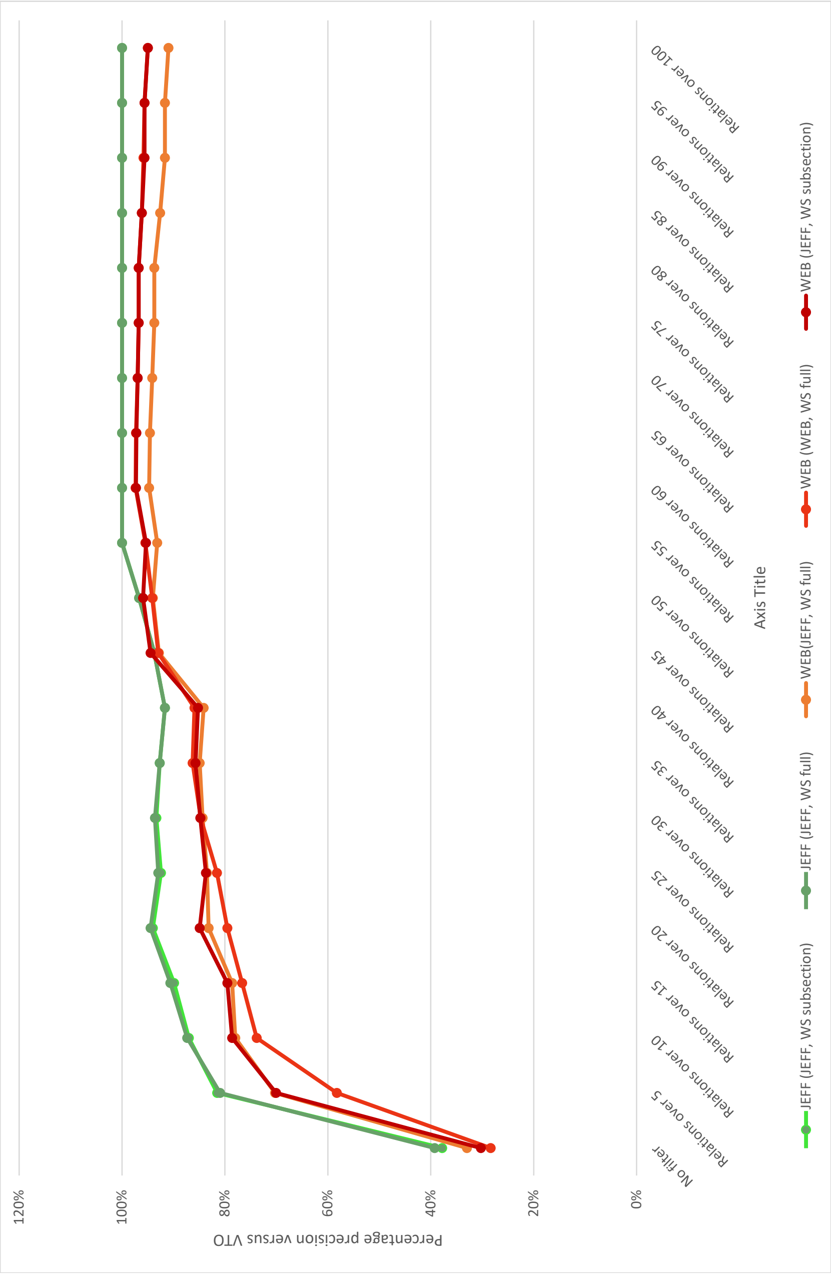


Figure 9.19: JEFF and WEB: precision vs VTO (frequency filter) [in colour]

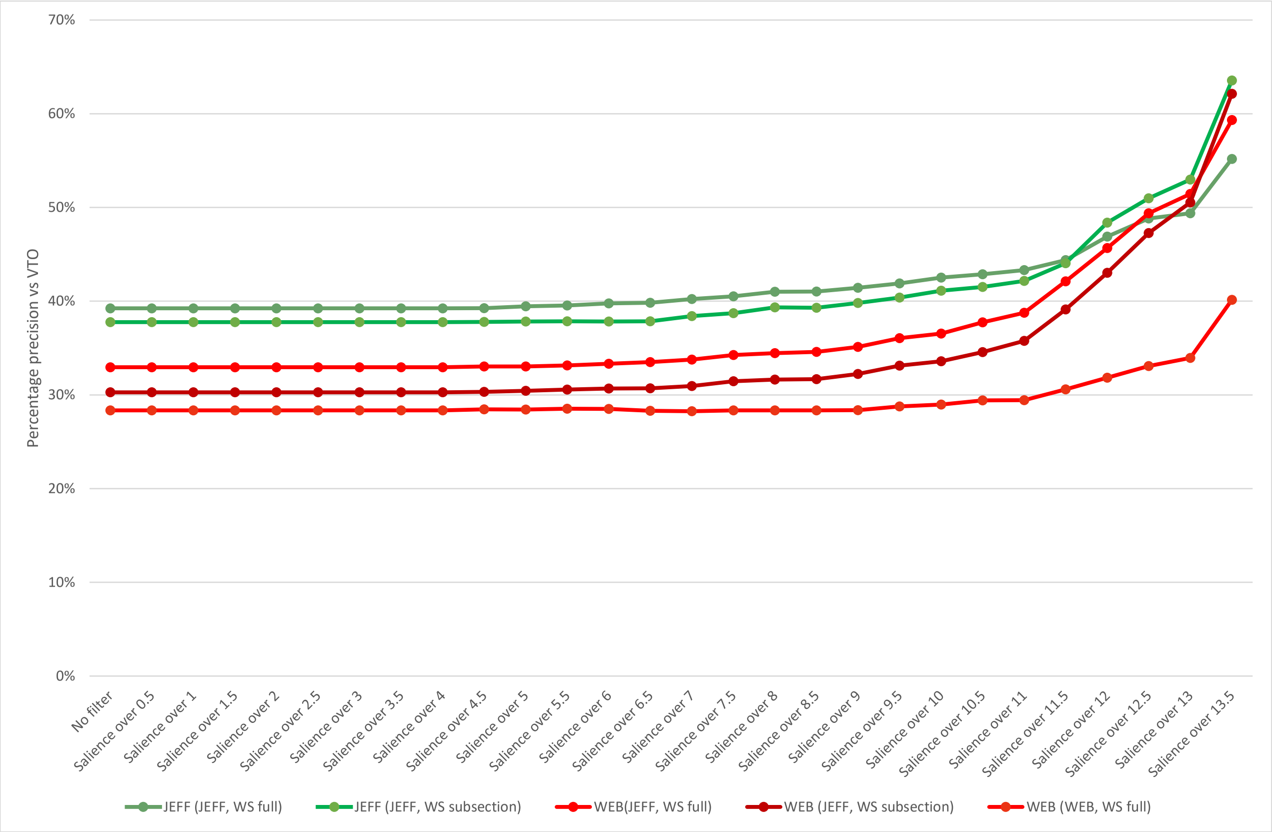


Figure 9.20: JEFF and WEB: precision vs VTO (salience filter) [in colour]

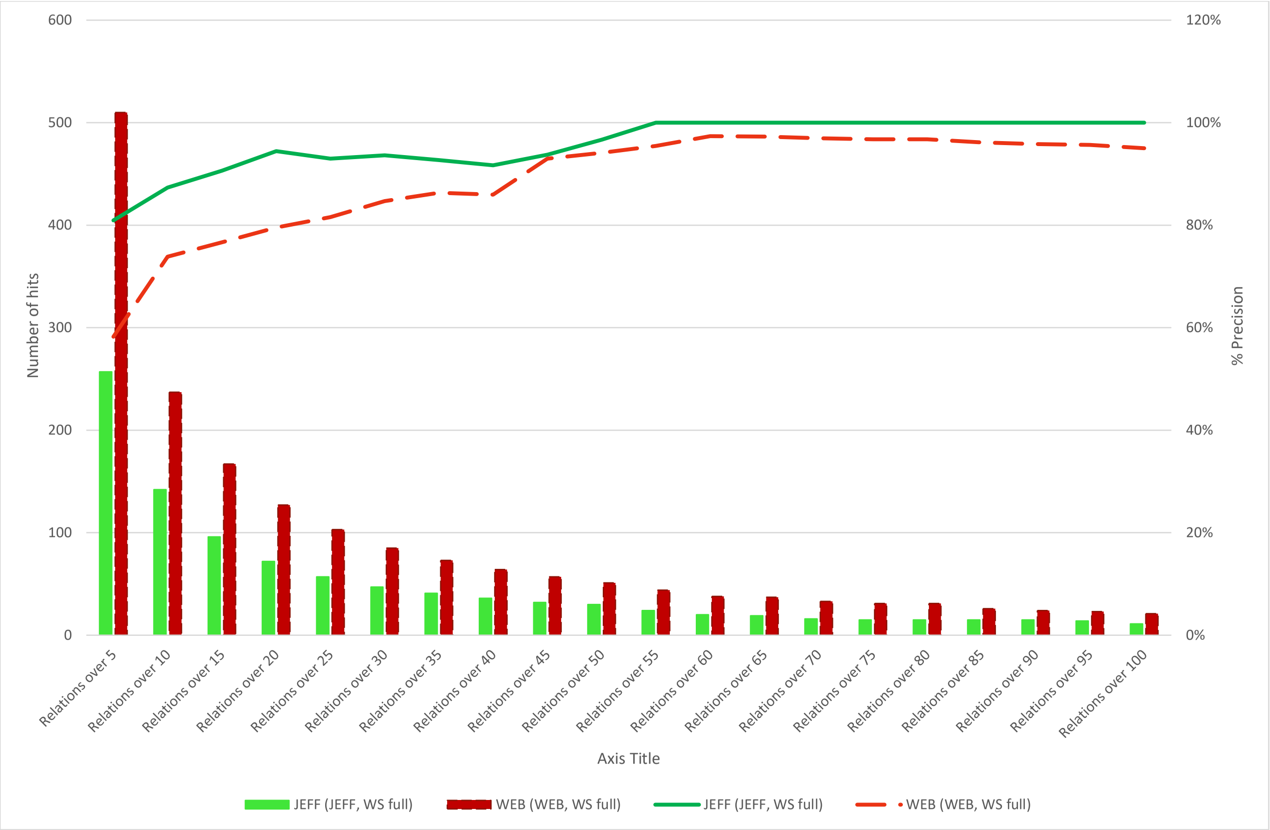


Figure 9.21: JEFF and WEB (Scenario 2): precision with rising frequency threshold [in colour]

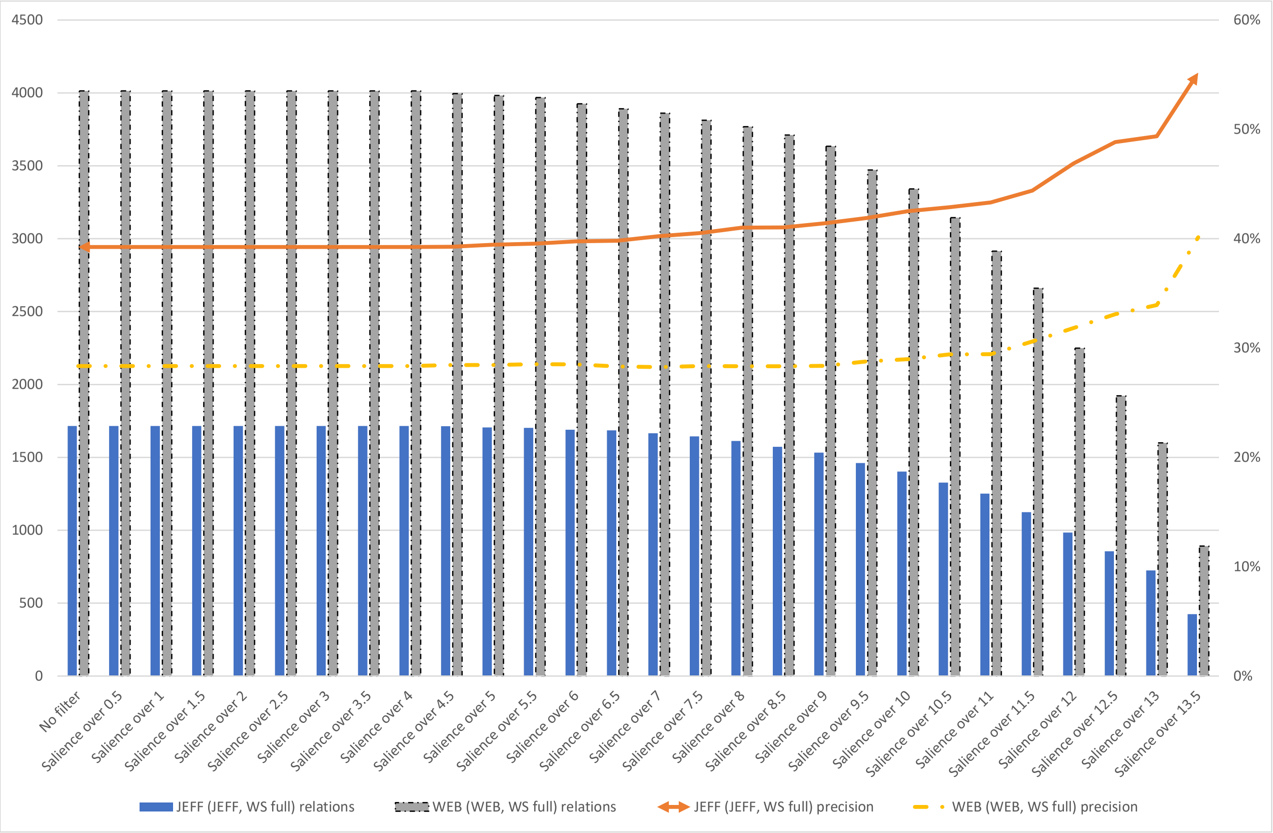


Figure 9.22: JEFF and WEB (Scenario 2): precision with rising salience threshold [in colour]

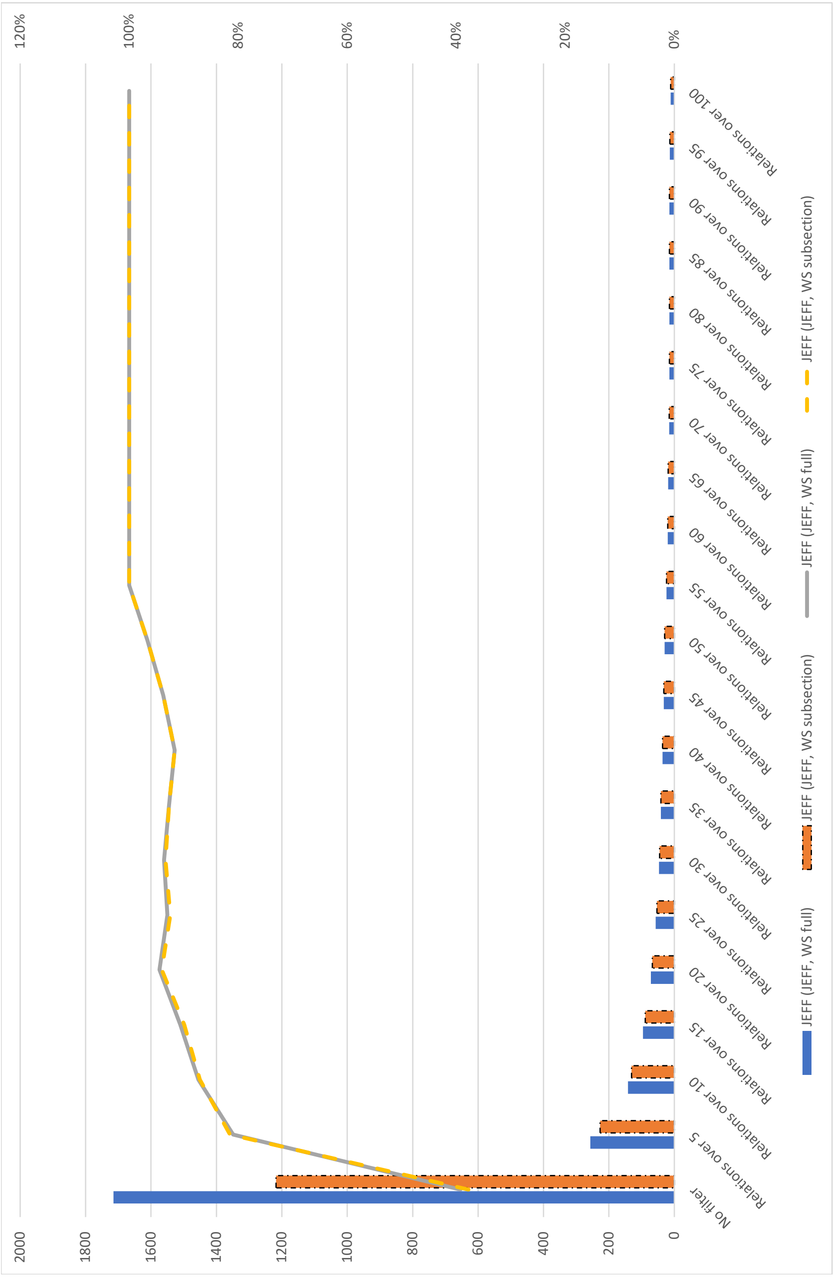


Figure 9.23: JEFF Scenarios 1 and 3: precision with rising frequency [in colour]

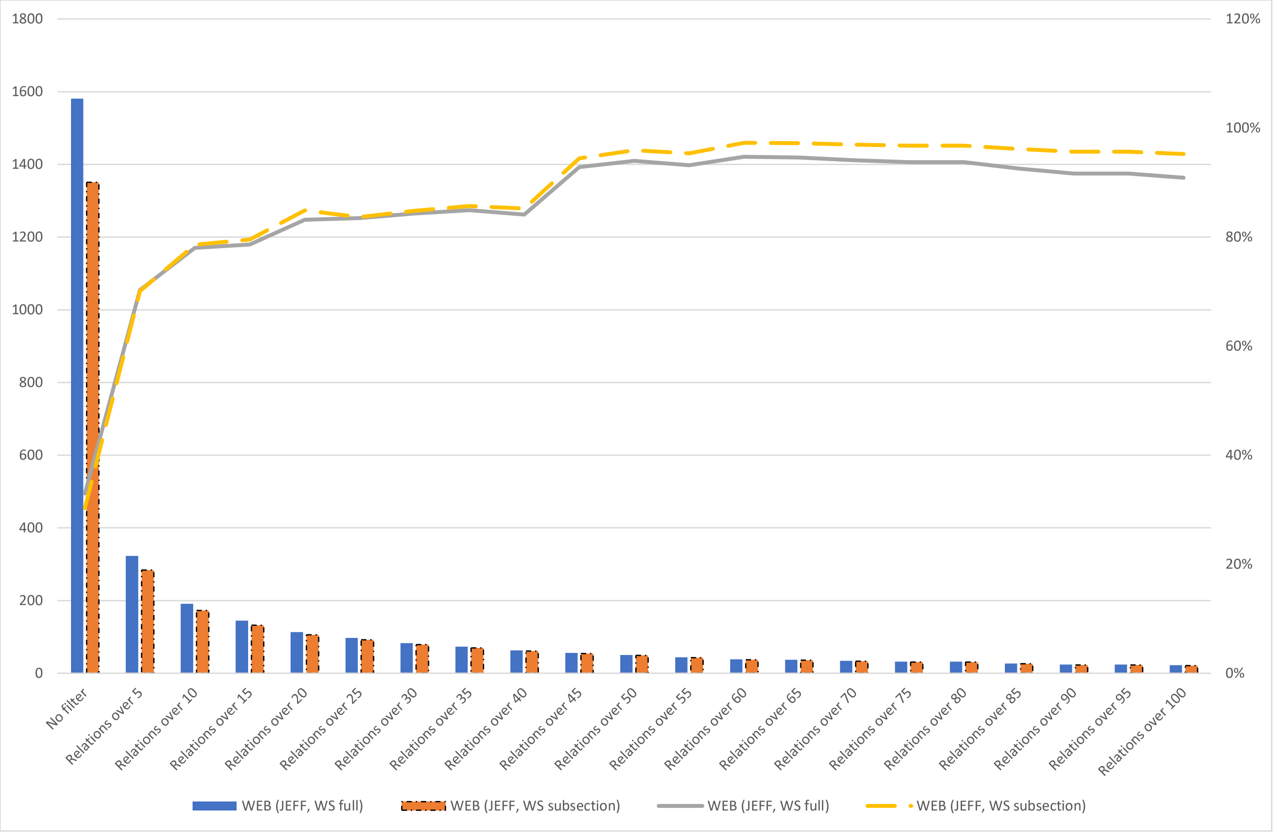


Figure 9.24: WEB Scenarios 1 and 3: precision with rising frequency [in colour]

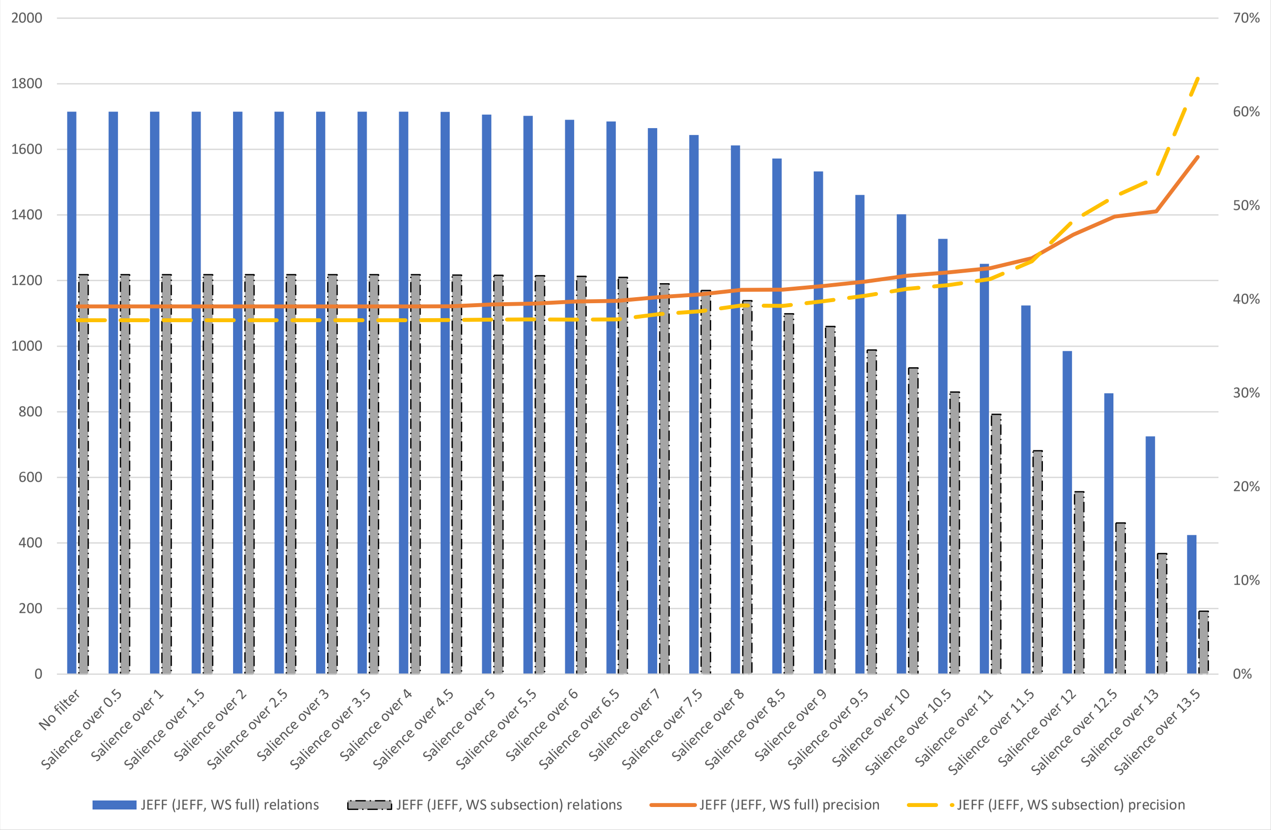


Figure 9.25: JEFF Scenarios 1 and 3: precision with rising salience [in colour]

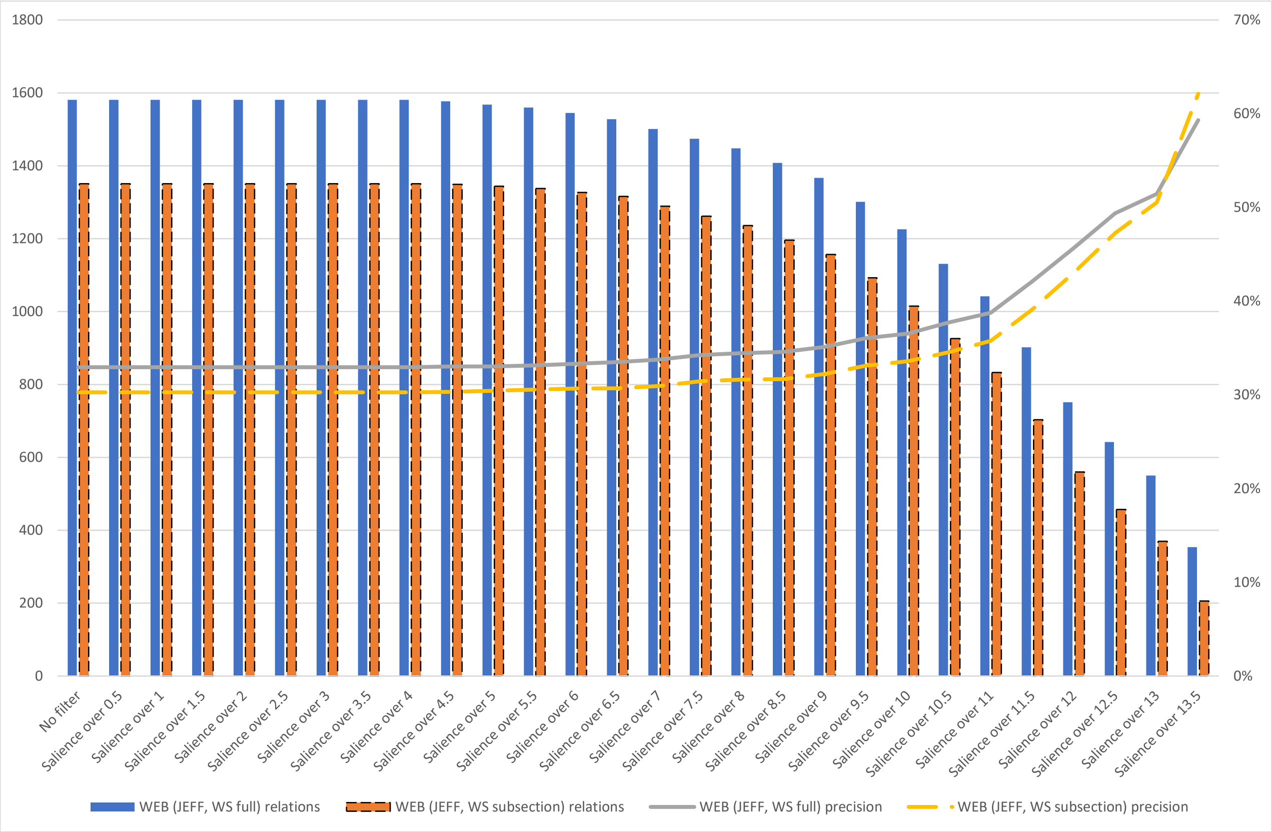


Figure 9.26: WEB Scenarios 1 and 3: precision with rising salience [in colour]

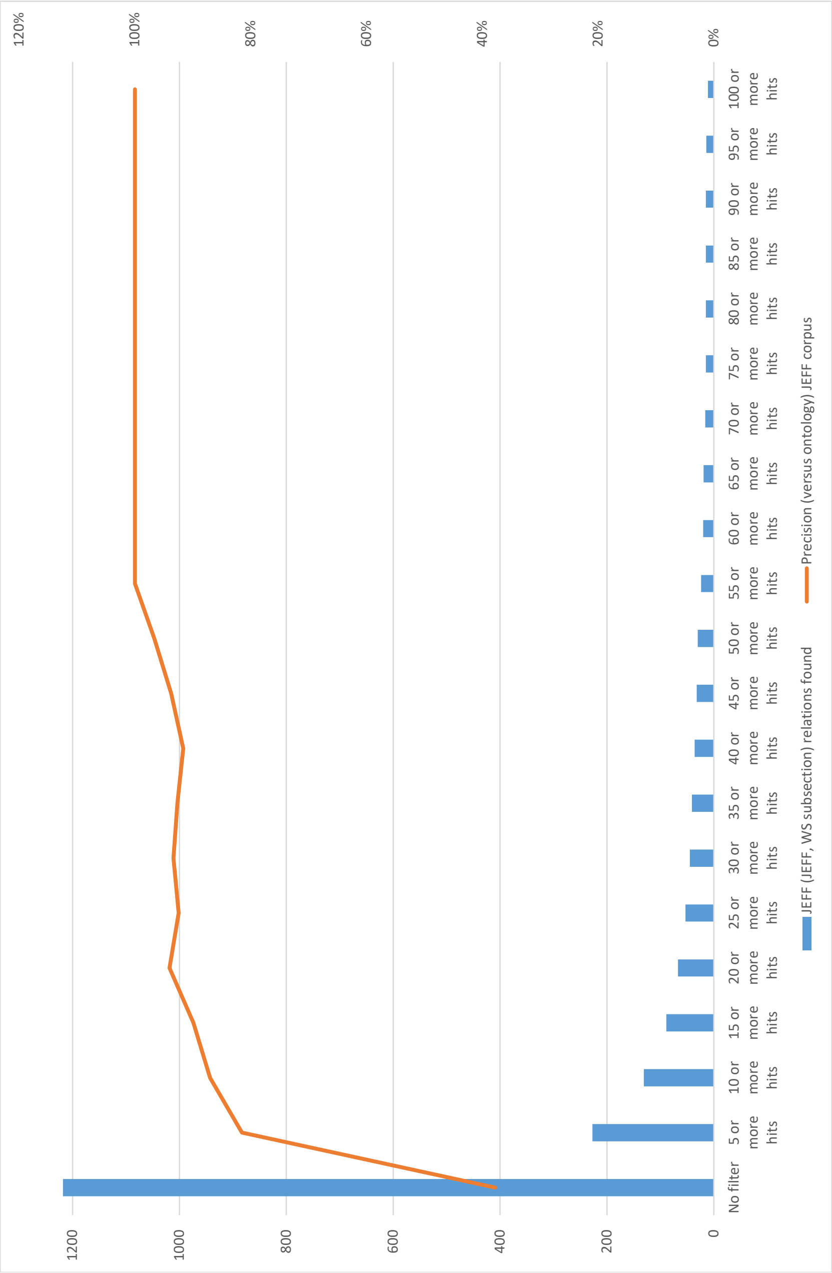


Figure 9.27: JEFF (JEFF, WS subsection) corpus : precision versus frequency of nomenclature pairs with frequency filter (Scenario 3)

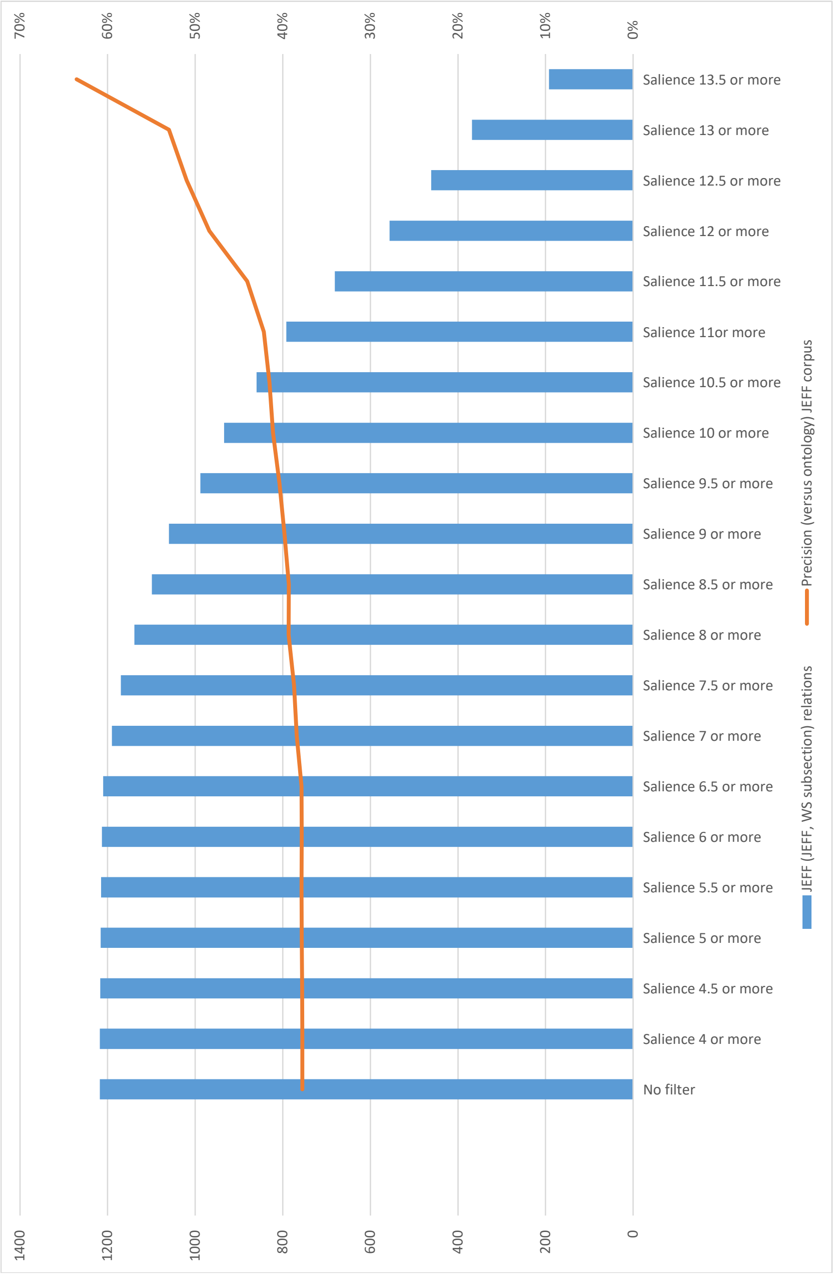


Figure 9.28: JEFF (JEFF, WS, subsection) corpus: precision versus frequency of nomenclature pairs with salience filter (Scenario 3)

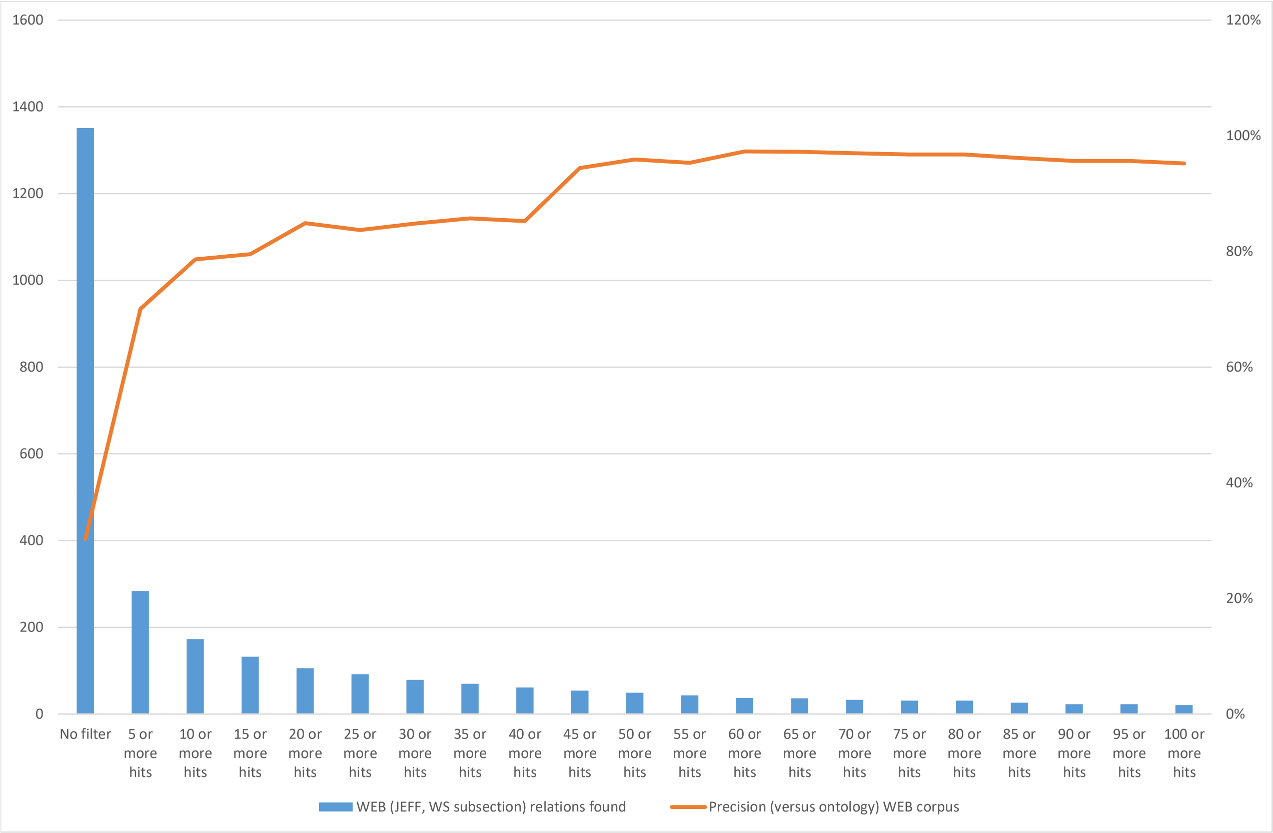


Figure 9.29: WEB corpus (JEFF, WS subsection): precision versus frequency of nomenclature pairs with frequency filter (Scenario 3)

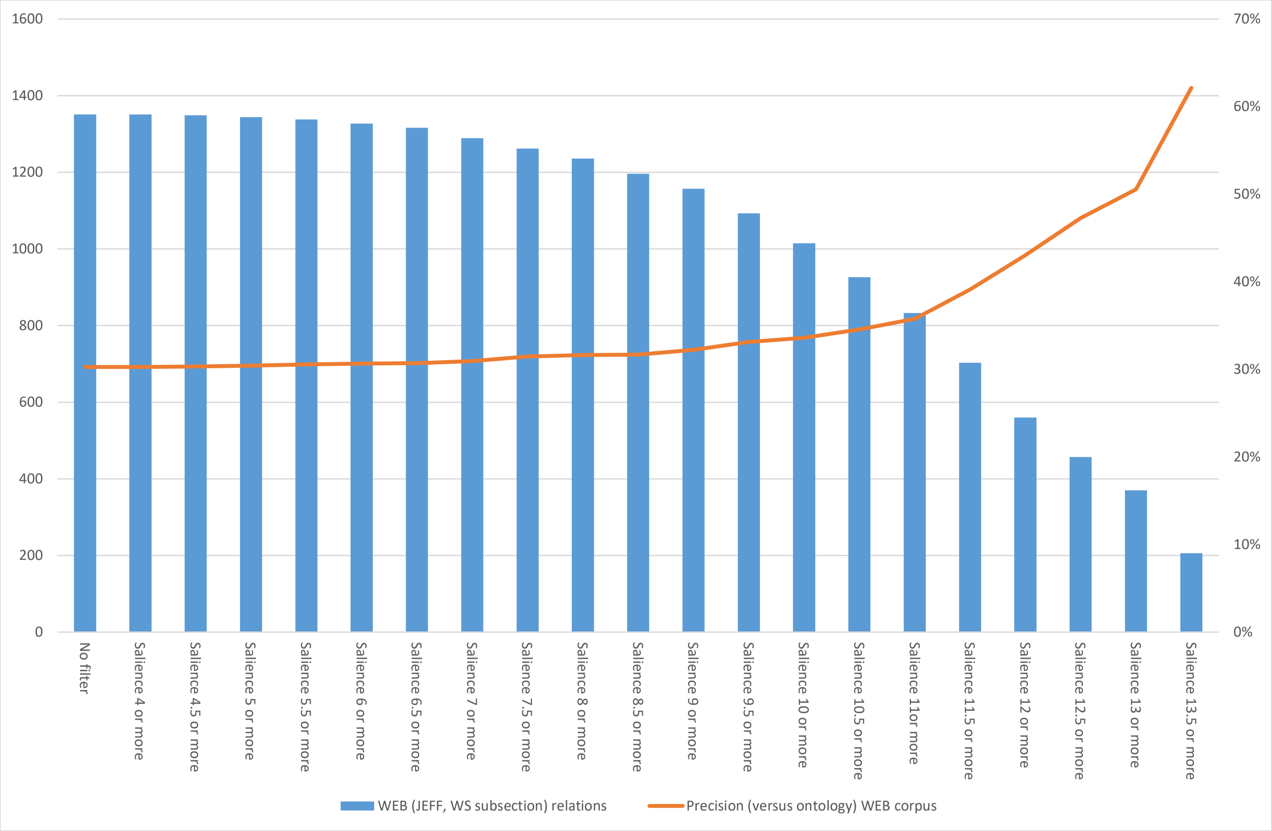


Figure 9.30: WEB (JEFF, WS, subsection) corpus: precision versus frequency of nomenclature pairs with salience filter (Scenario 3)

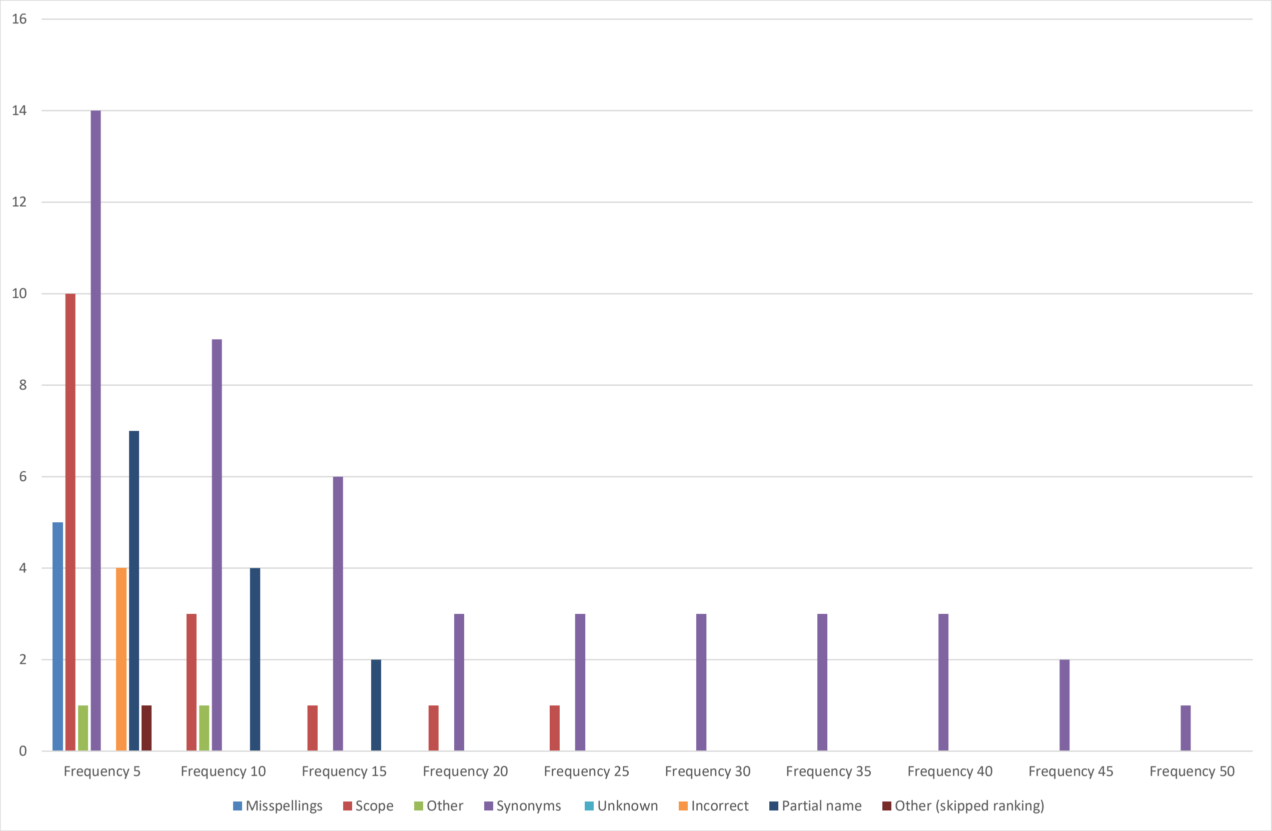


Figure 9.31: Breakdown of differences between JEFF corpus and VTO [in colour]

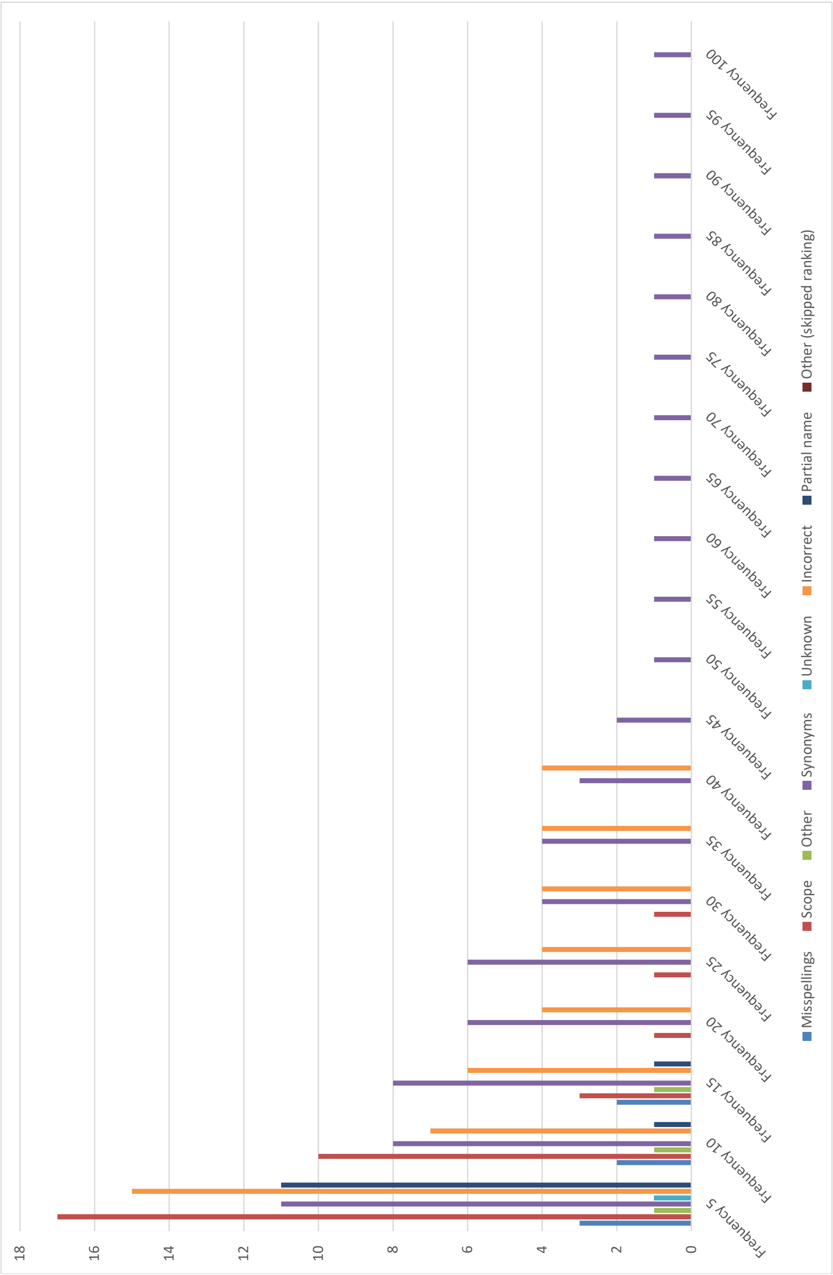


Figure 9.32: Breakdown of differences between WEB corpus and VTO [in colour]

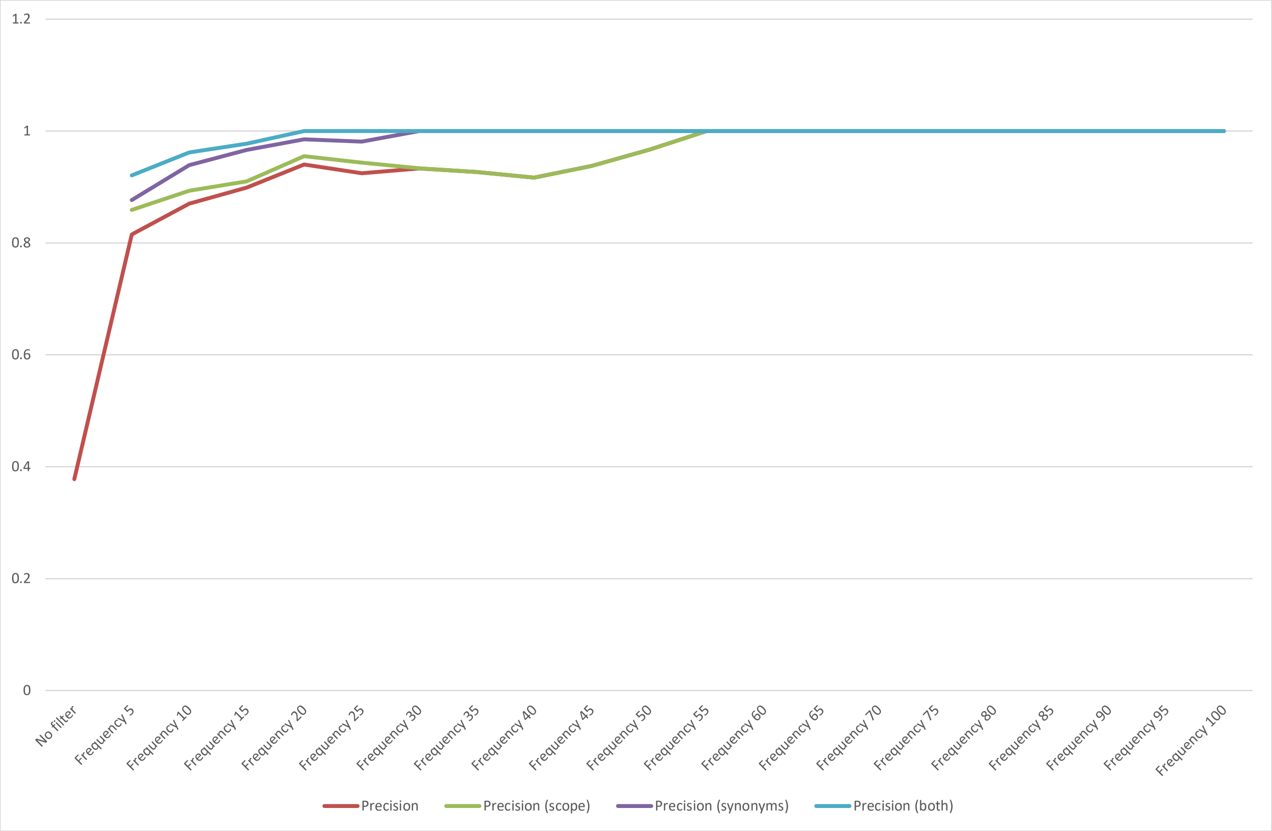


Figure 9.33: JEFF corpus: precision comparison whether including out-of-scope species or synonyms or both (adjusted precision scores)

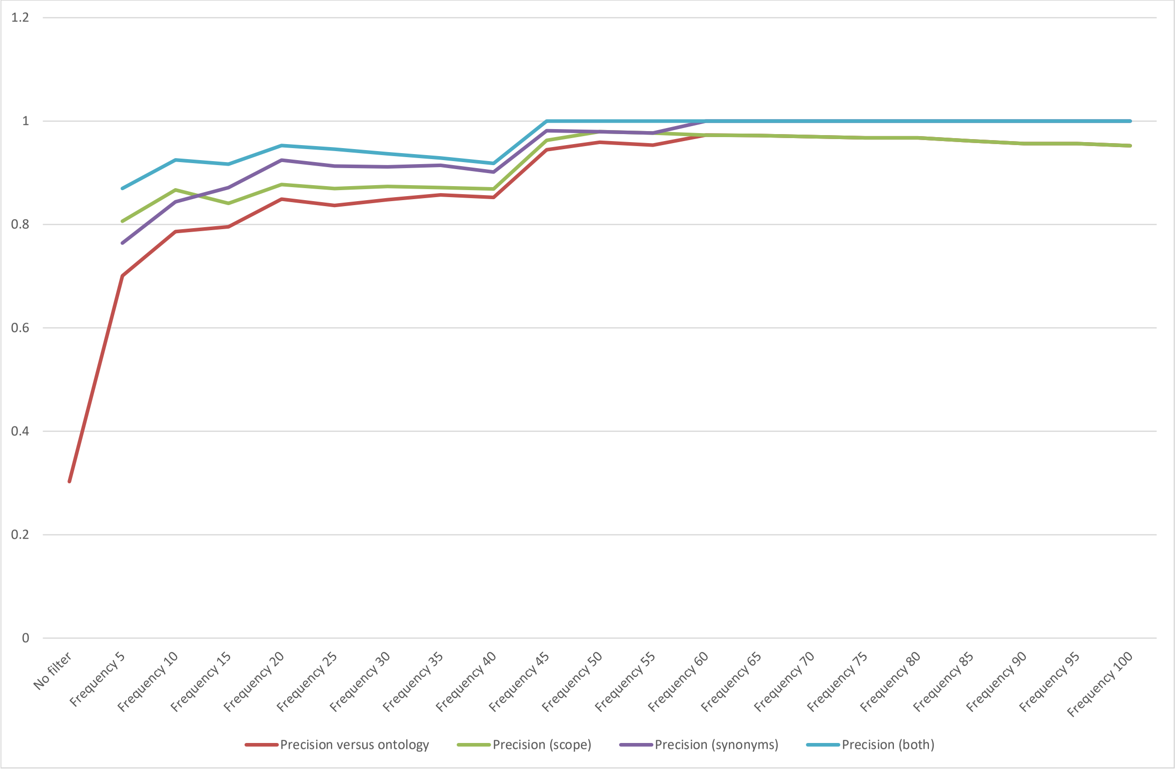


Figure 9.34: WEB corpus: precision comparison whether including out-of-scope species or synonyms or both (adjusted precision scores)

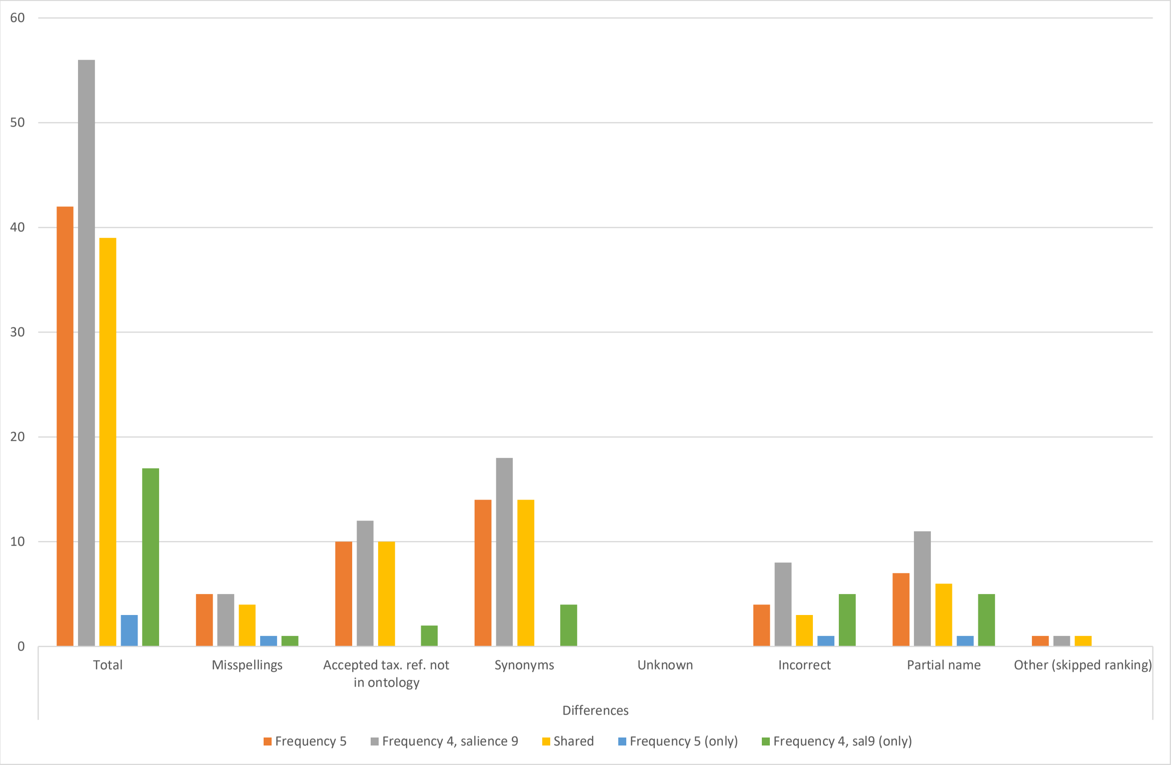


Figure 9.35: JEFF corpus: graph breakdown of differences which compare frequency 5 with frequency 4, salience 9

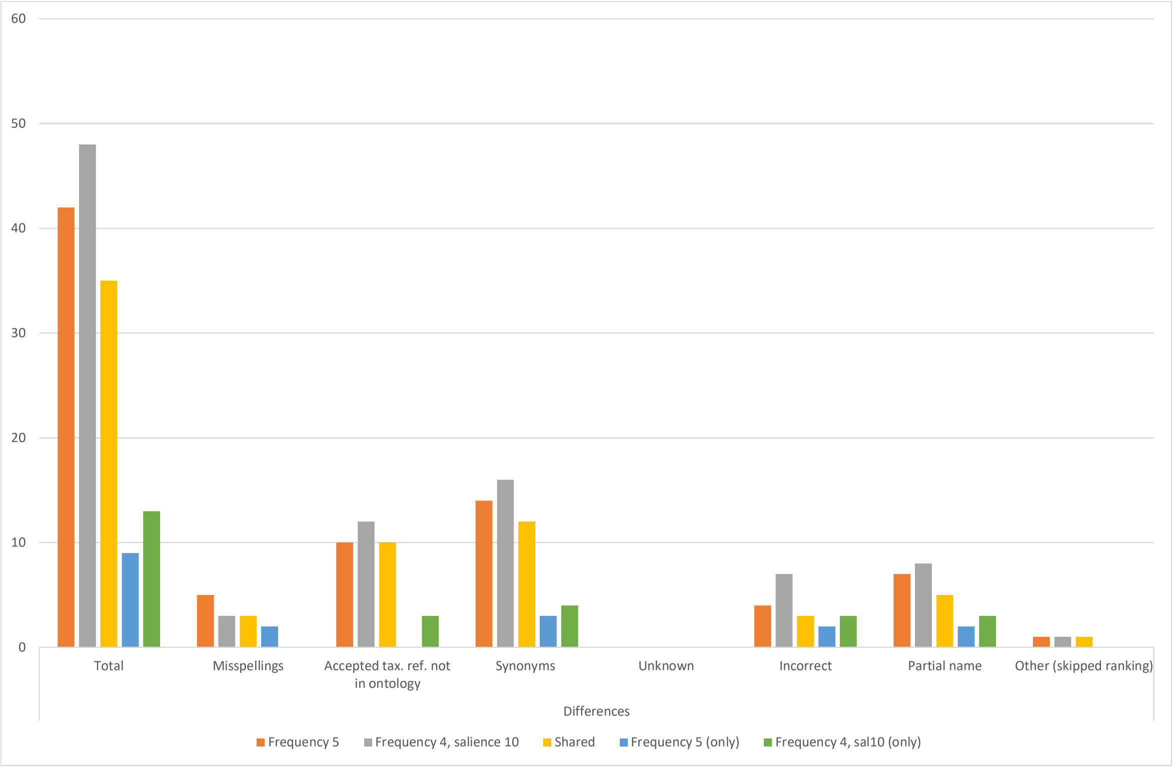


Figure 9.36: JEFF corpus: graph breakdown of differences which compare frequency 5 with frequency 4, salience 10

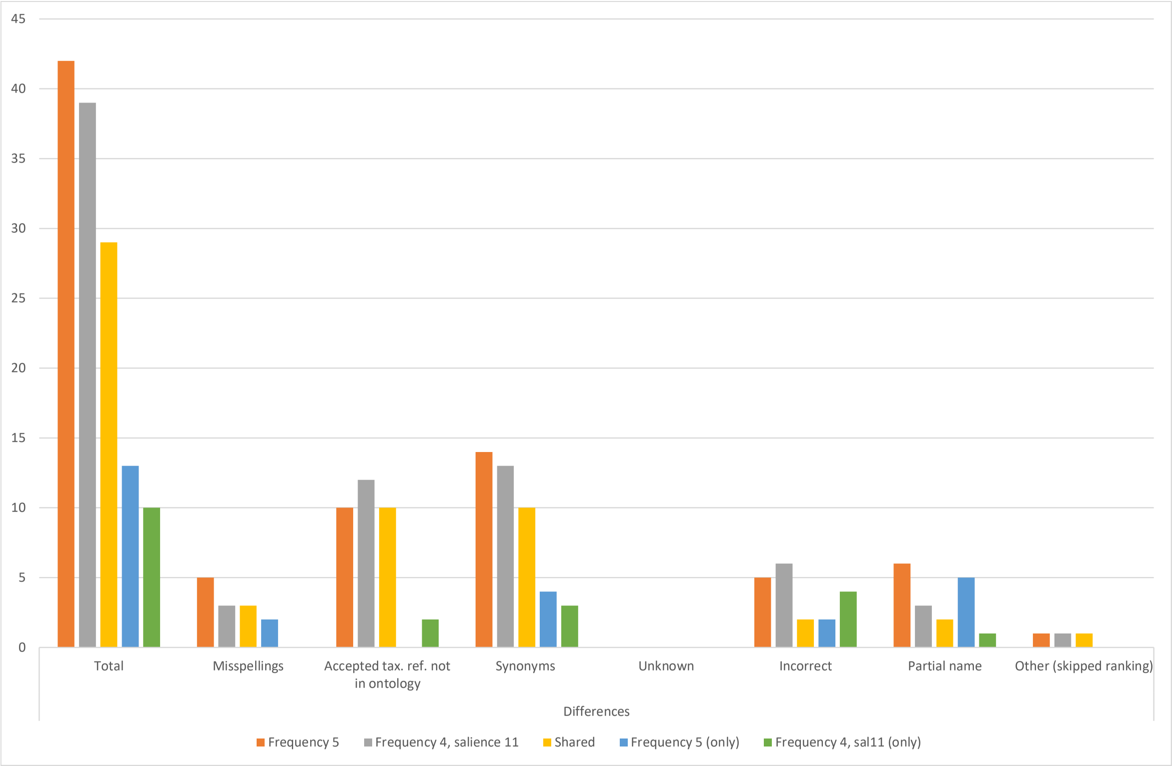


Figure 9.37: JEFF corpus: graph breakdown of differences which compare frequency 5 with frequency 4, salience 11

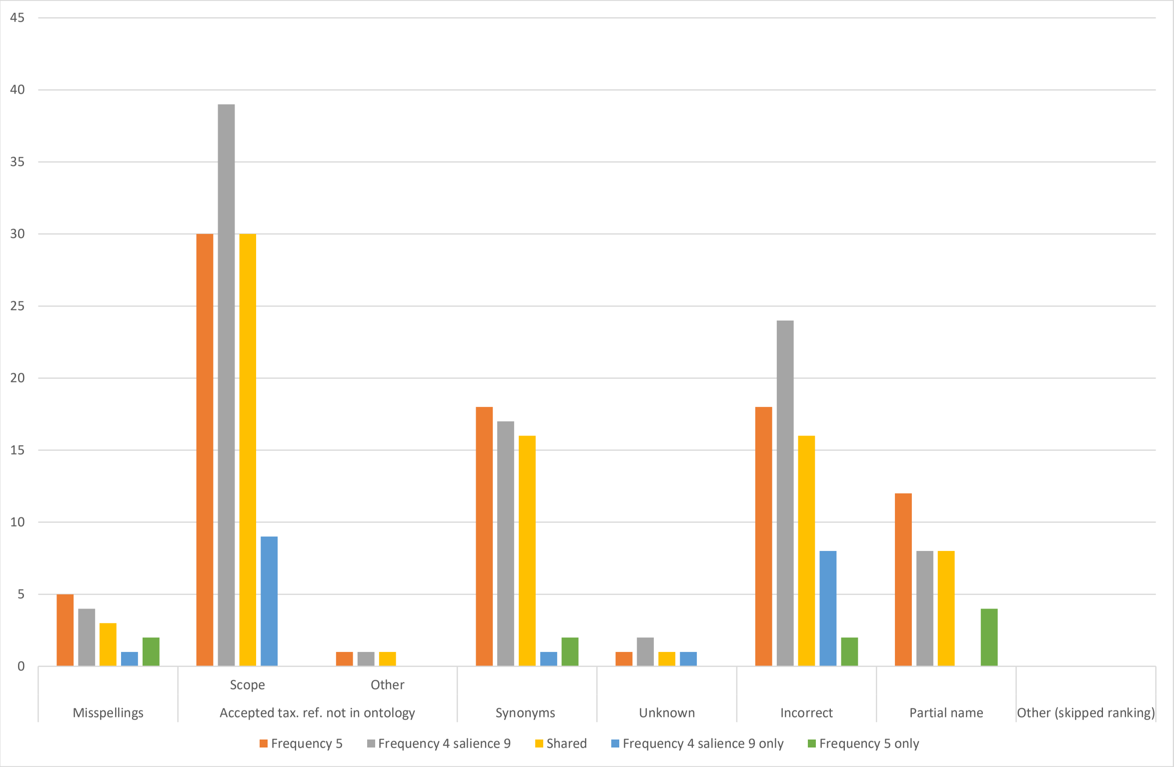


Figure 9.38: WEB corpus: graph breakdown of differences which compare frequency 5 with frequency 4, salience 9

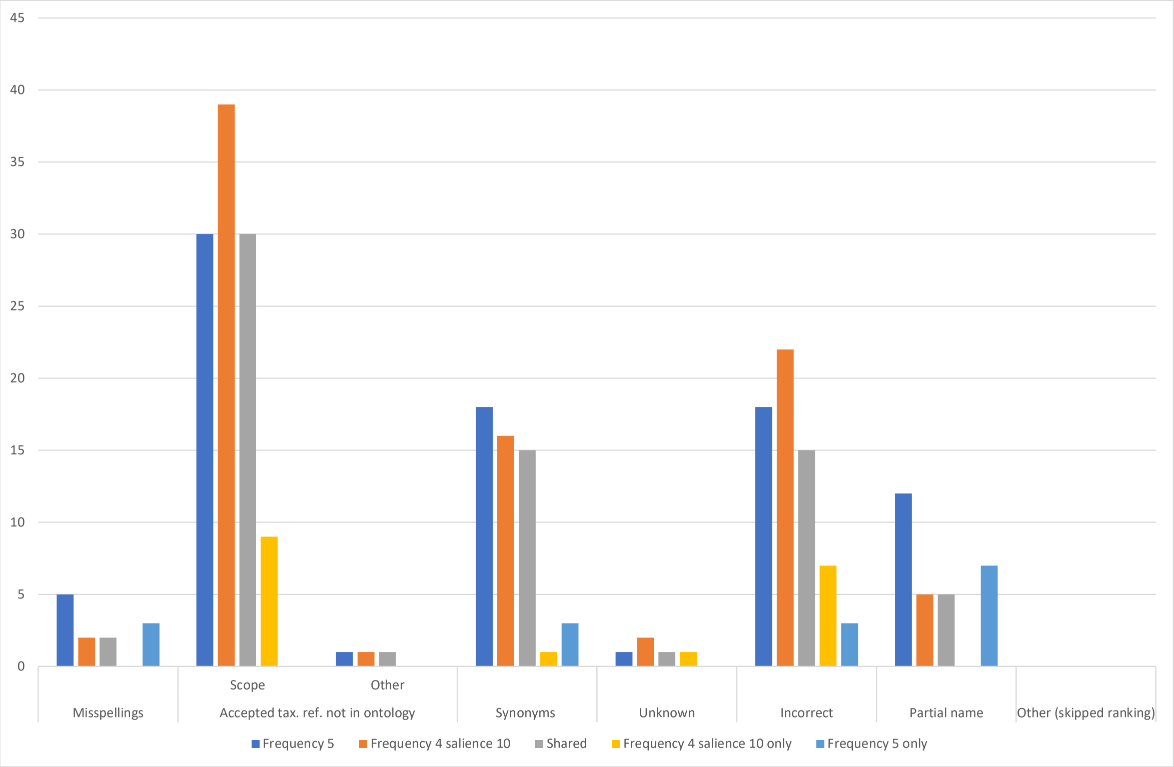


Figure 9.39: WEB corpus: graph breakdown of differences which compare frequency 5 with frequency 4, salience 10

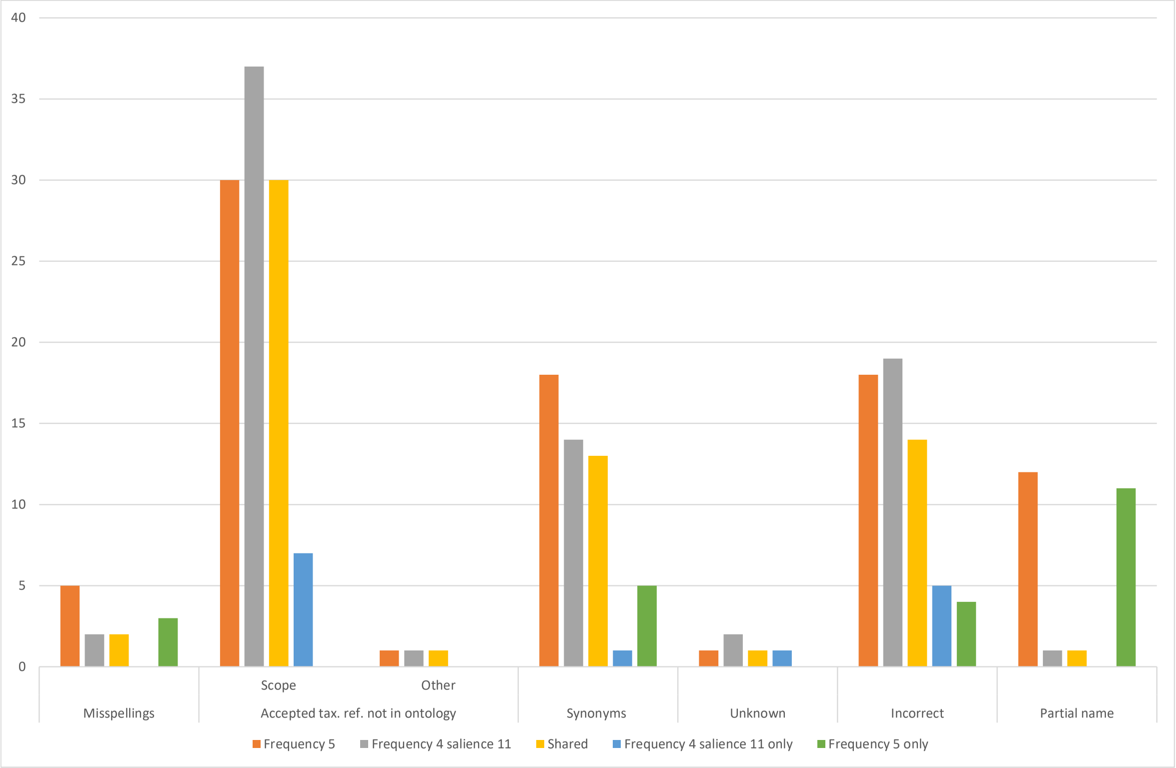


Figure 9.40: WEB corpus: graph breakdown of differences which compare frequency 5 with frequency 4, salience 11

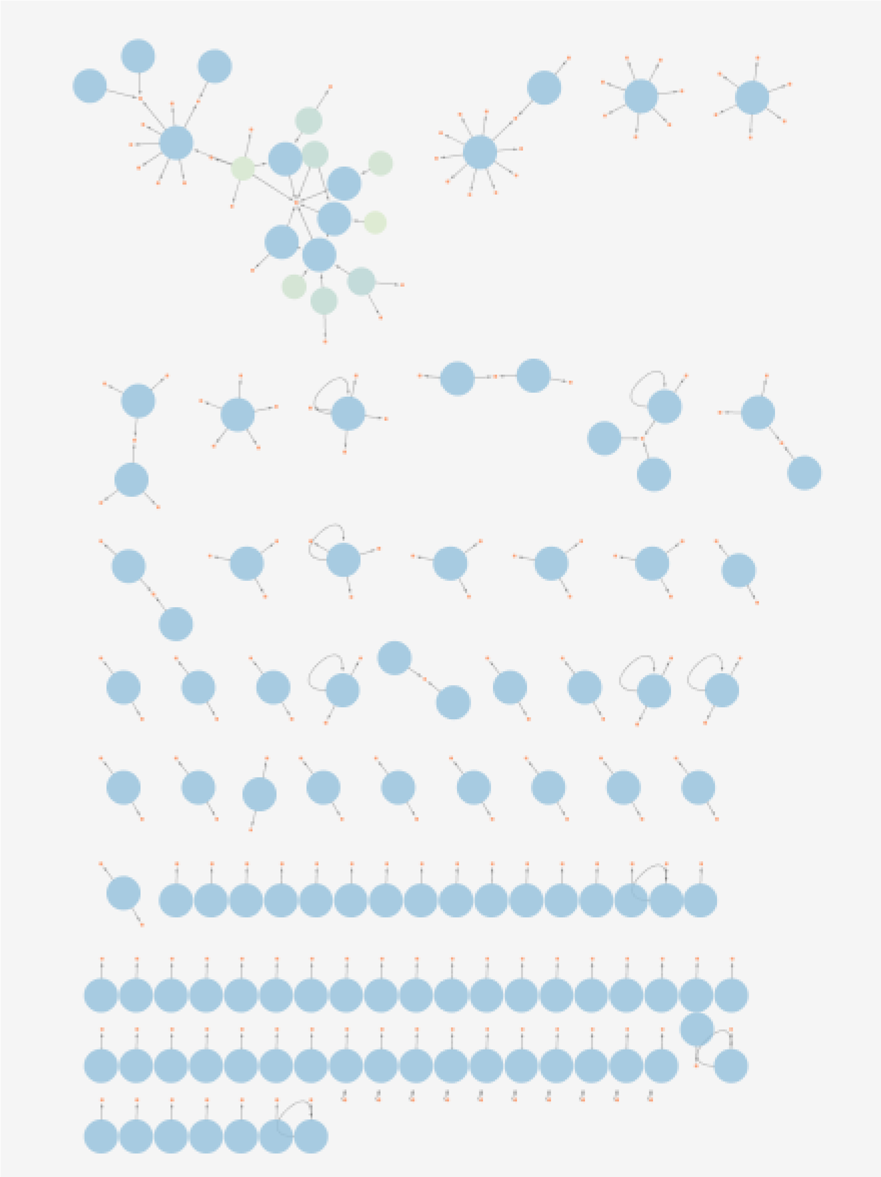


Figure 9.41: JEFF corpus: overview, relation network graph frequency 5. The large, connected section of the graph relates to the Salmonidae family, plus a number of other smaller groupings.

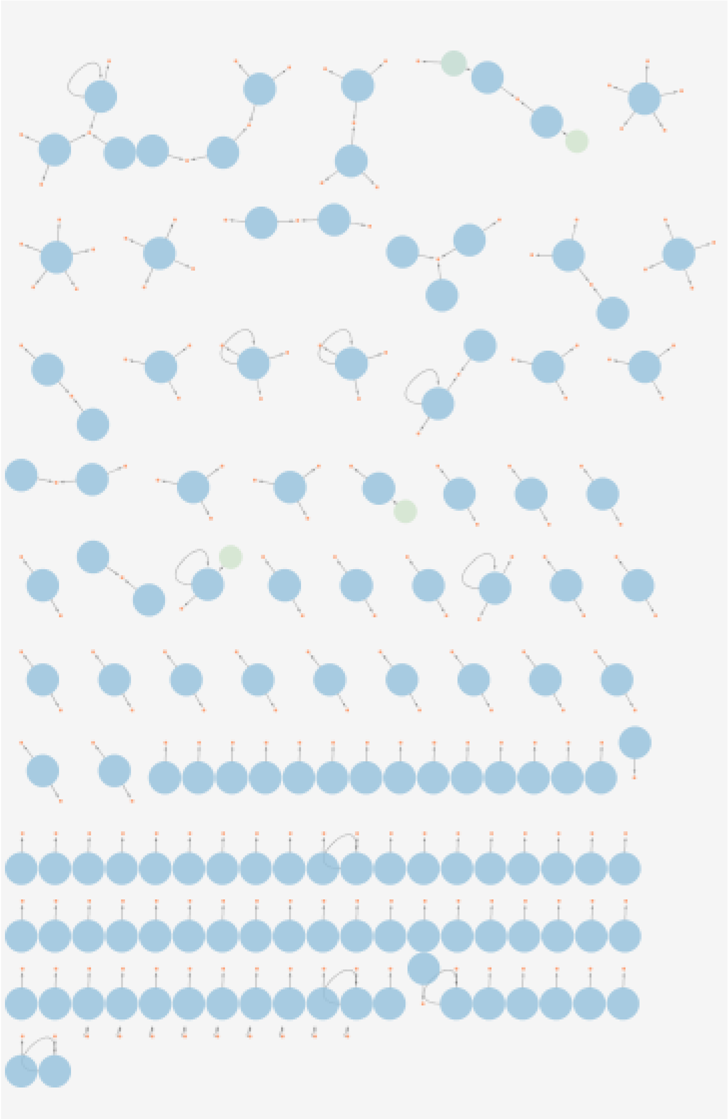


Figure 9.42: JEFF corpus: overview, relation network graph frequency 4, salience 11. In comparison with the previous graph, the graph has more, less connected groupings and the large connected section at the top of the graph is no longer there.

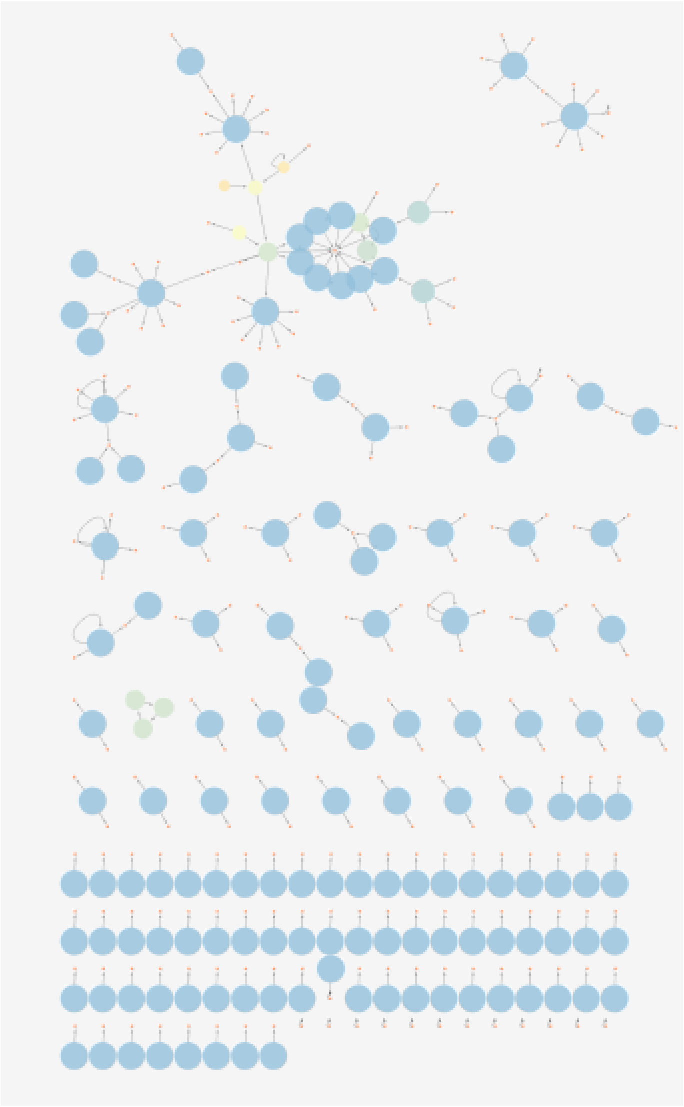


Figure 9.43: WEB corpus: overview, relation network graph frequency 5. As in the JEFF frequency 5 graph there is a large connected section of the graph at the top, again relating to the Salmonidae family.

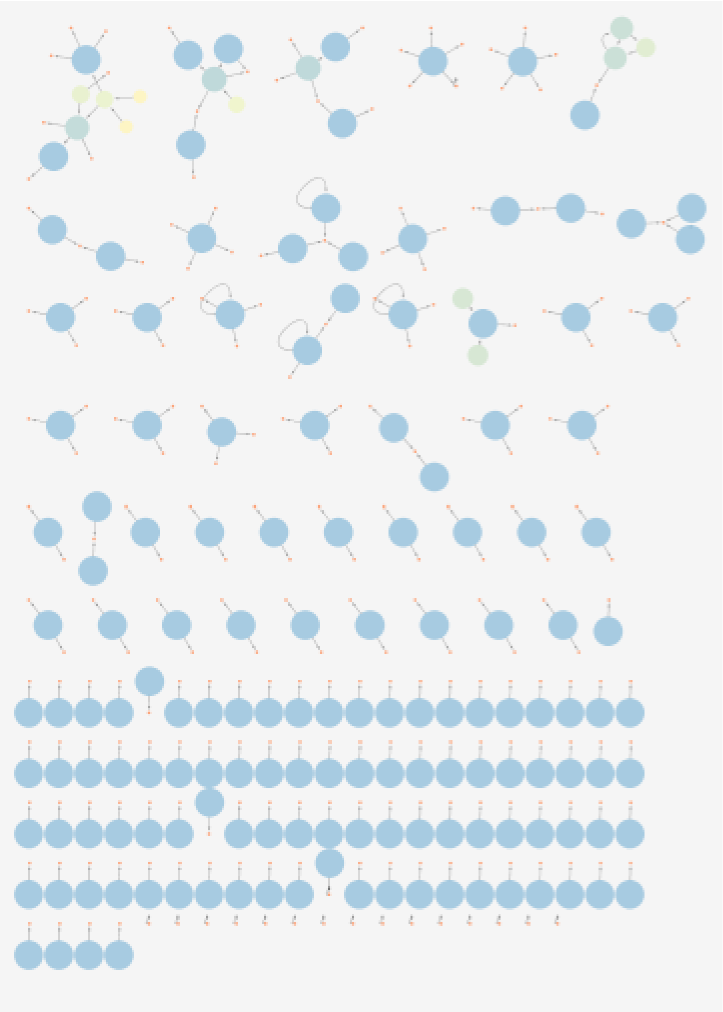


Figure 9.44: WEB corpus: overview, relation network graph frequency 4, salience 11. Again, in line with the more segmented nature of the salience filter graph, here there are multiple, smaller groupings and the large, connected section of the graph is not seen.

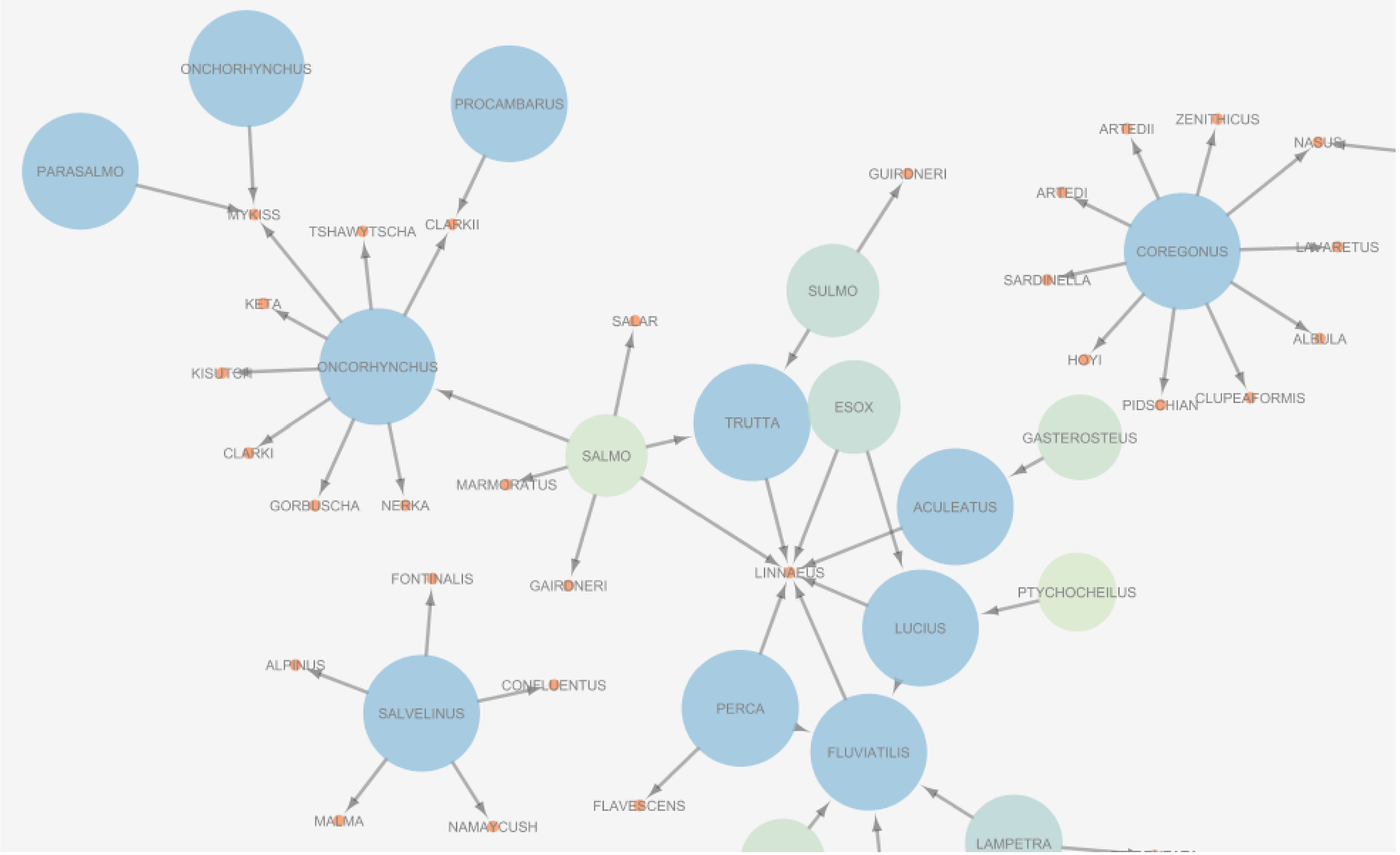


Figure 9.45: JEFF corpus: Salmonidae, relation network graph frequency 5. Many hubs can be identified in this graph, including the Oncorhynchus, Salvelinus, Coregonus and Salmo hubs. These are all surrounded by multiple species variants and linked through some species-level nodes as well as Linneaus as an authorship-level node. Misspellings such as Sulmo are also included in the graph.

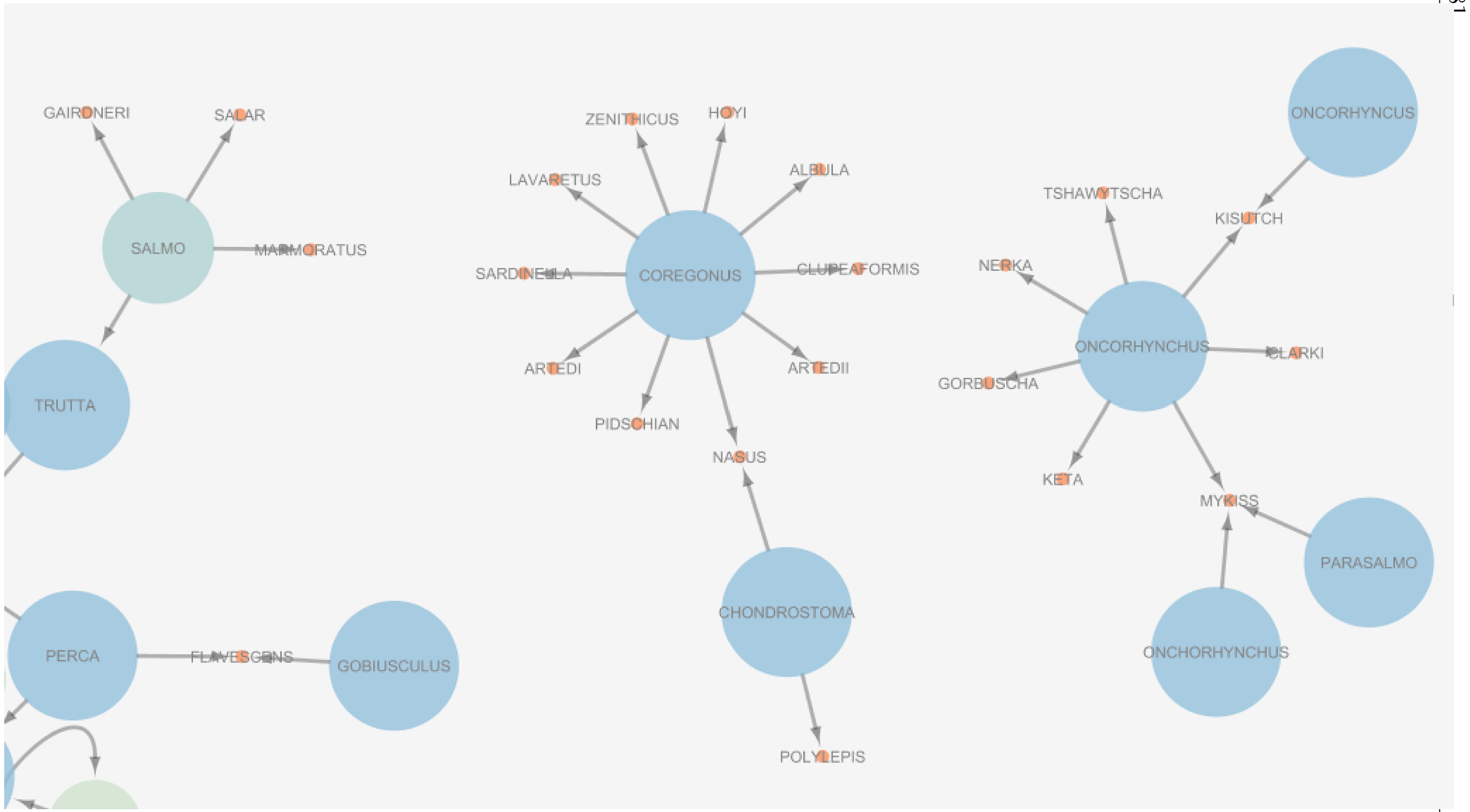


Figure 9.46: JEFF corpus: Salmonidae, relation network graph frequency 4, salience 9. The linkages between the different hubs have now reduced, including the disappearance of Linneaus as a node. Now the different genera remain as hubs but in separate forms of the graph. This graph identifies linked genus variants (i.e. Parasalmo being an alternative variant to Oncorhynchus) and misspellings (Oncorhynchus versus Oncorchyncus). There are still a number of species variants surrounding each hub.

.

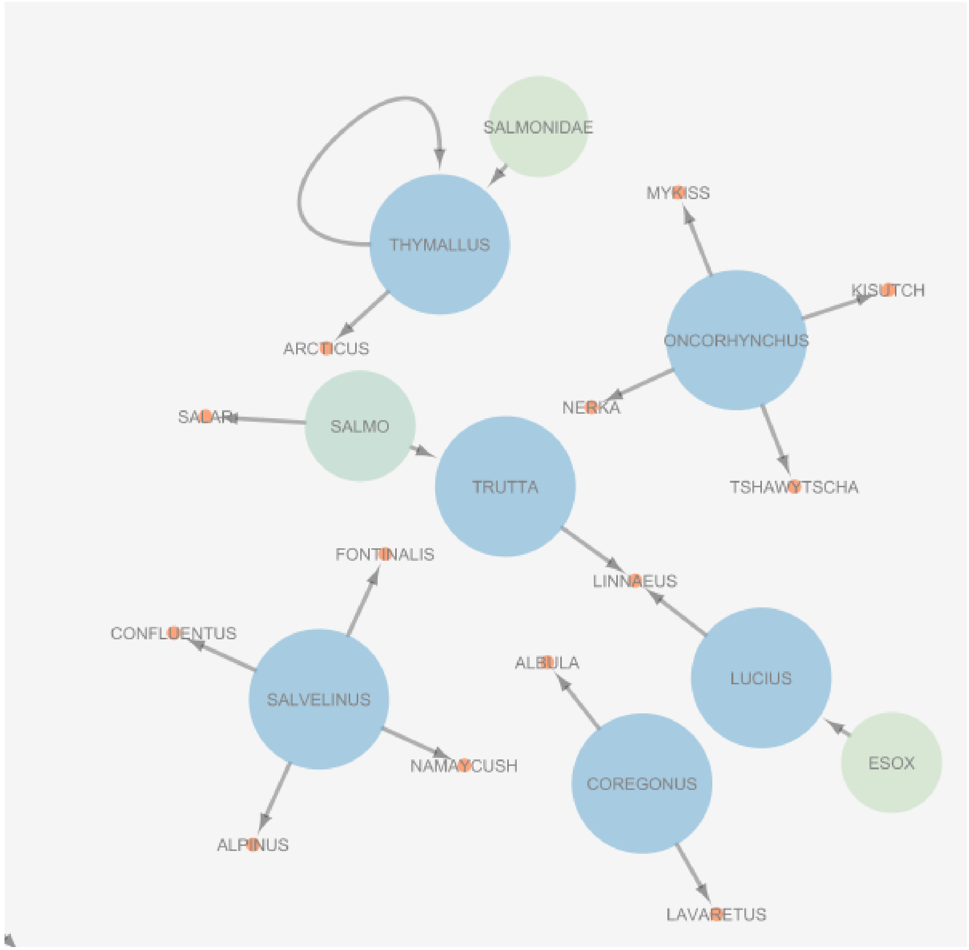


Figure 9.47: JEFF corpus: Salmonidae, relation network graph frequency 4, salience 11. A much more sparsely populated graph. While a number of genus hubs still appear, the number of different species variants have been greatly reduced (Coregonus only has two variants in comparison with ten in the filter 4, salience 9 graph).

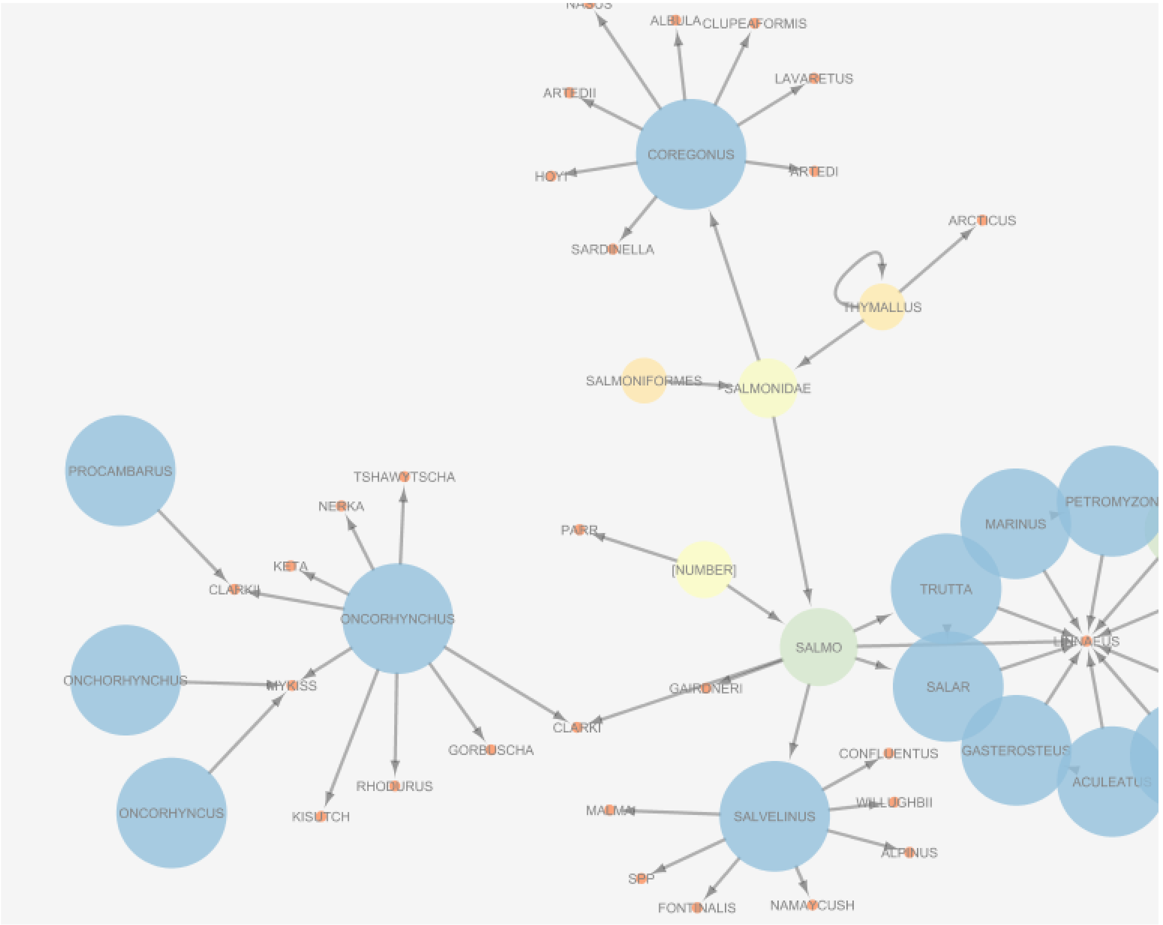


Figure 9.48: WEB corpus: Salmonidae, relation network graph frequency 5. Various genera hubs identified as in the JEFF corpus relation network graph, such as Oncorhynchus, Salvelinus, Coregonus. There is an improved hierarchy than in the JEFF corpus through Salmoniforme and Salmonidae.

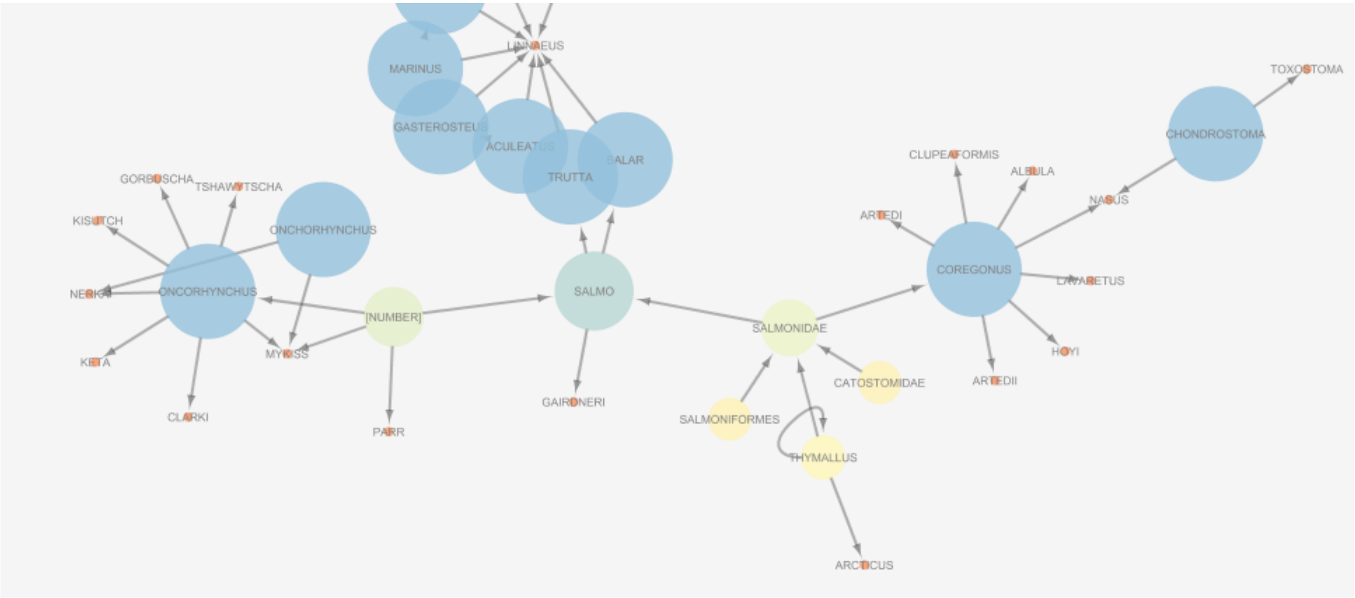


Figure 9.49: WEB corpus: Salmonidae, relation network graph frequency 4, salience 9. Here most of the links are still remaining, the hierarchy still remains through Salmoniformes and Salmonidae but some of the misspellings have disappeared.

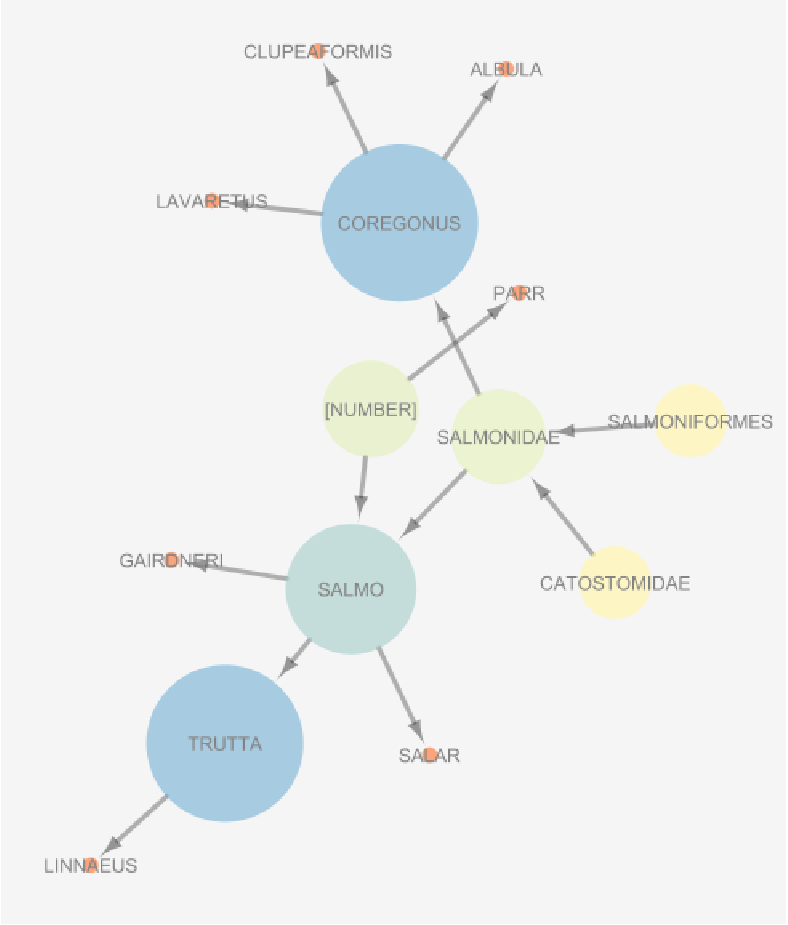


Figure 9.50: WEB corpus: Salmonidae, relation network graph frequency 4, salience 11.This final graph is very limited, the hierarchy remains but most of the genera hubs have completed disappeared.

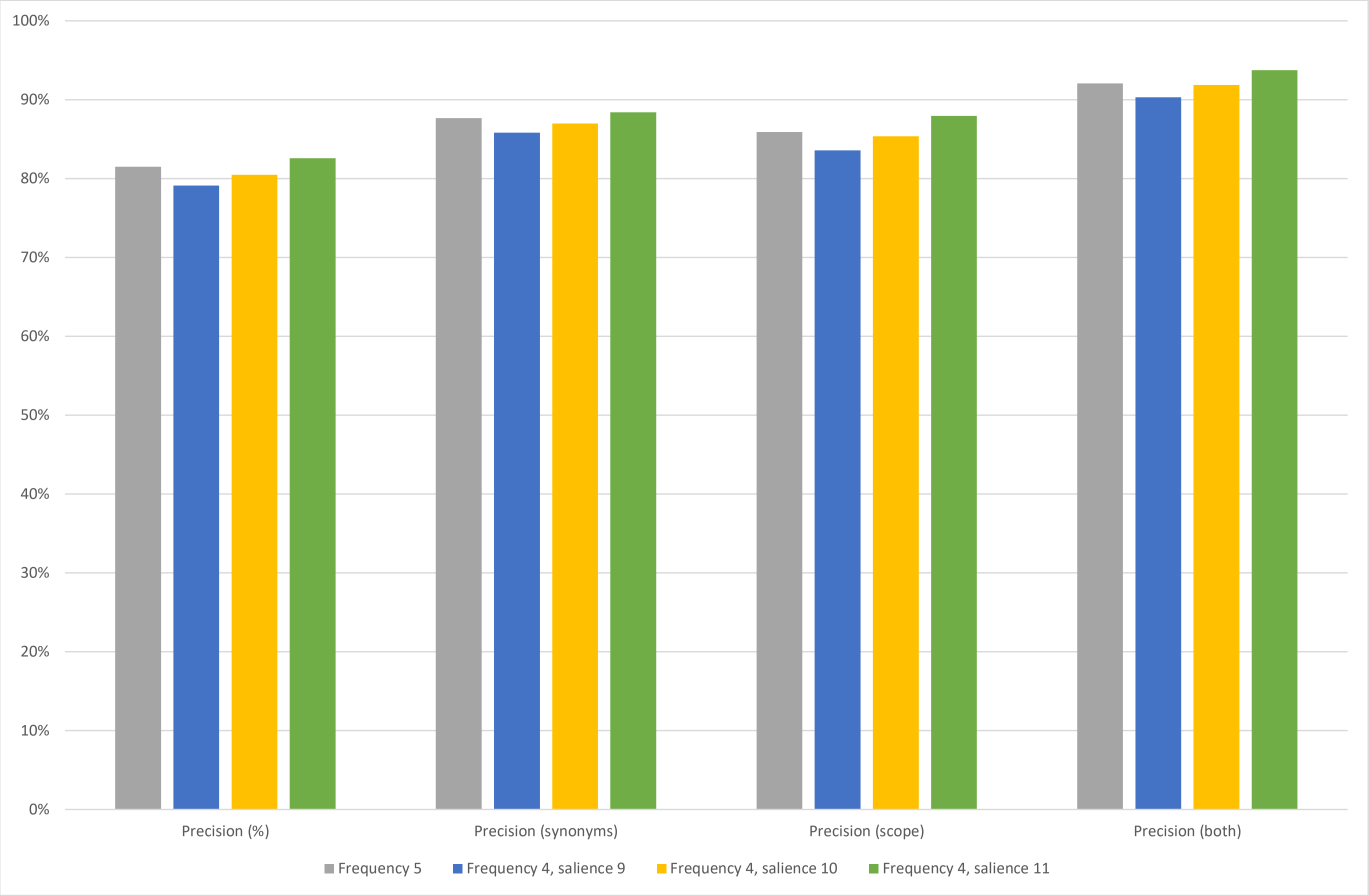


Figure 9.51: JEFF corpus: adjusted precision score comparison weighted for synonyms and scope

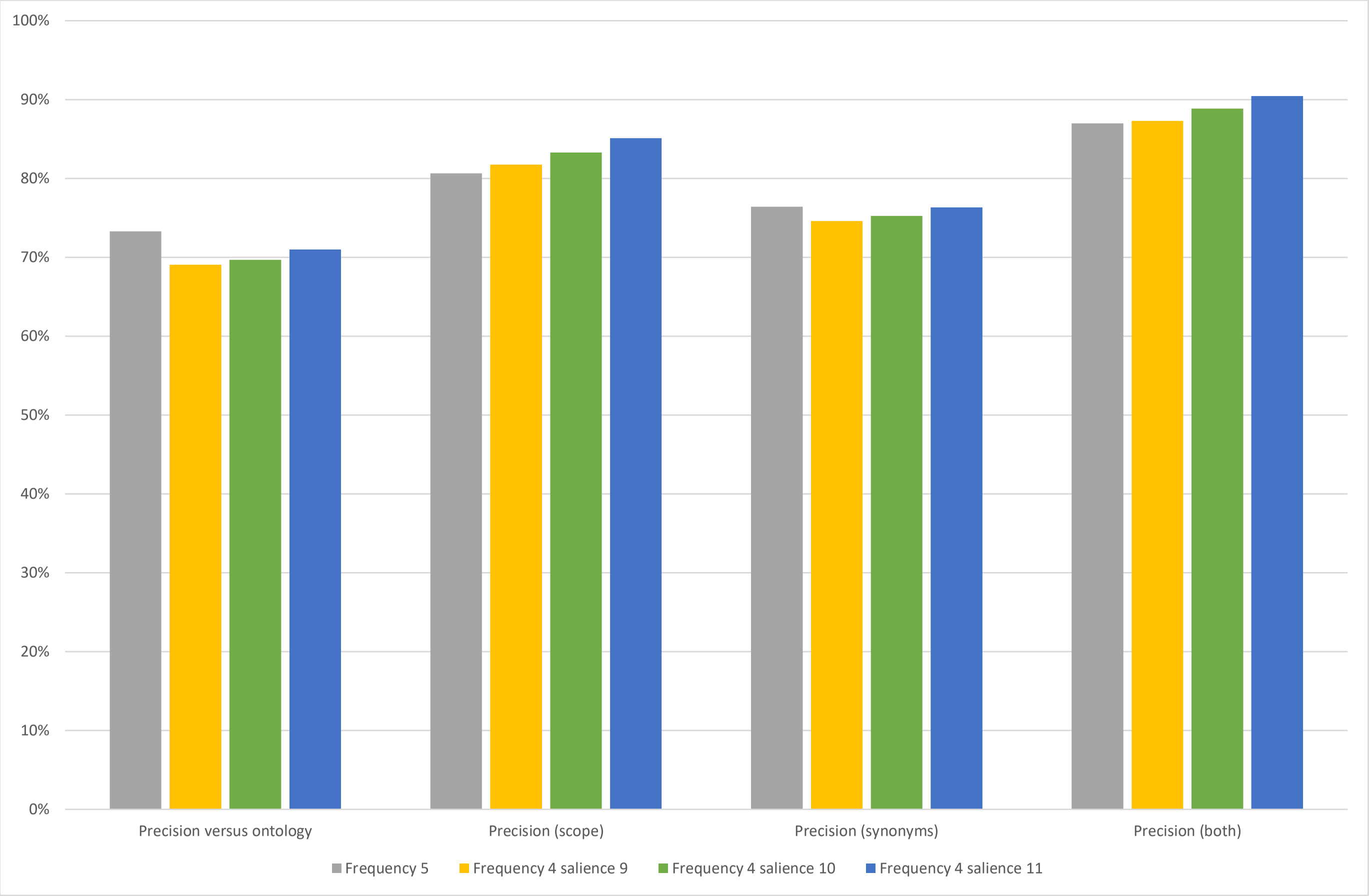


Figure 9.52: WEB corpus: adjusted precision score comparison weighted for synonyms and scope

9.3 Chapter 6

The graphs in this section are separated according to the Nomenclature Profile Study included in the final results chapter, relating to the application of the method. They consist of a number of dispersion and frequency graphs before moving onto study the relation network graphs and how they have been used to profile nomenclature usage in the test corpora.

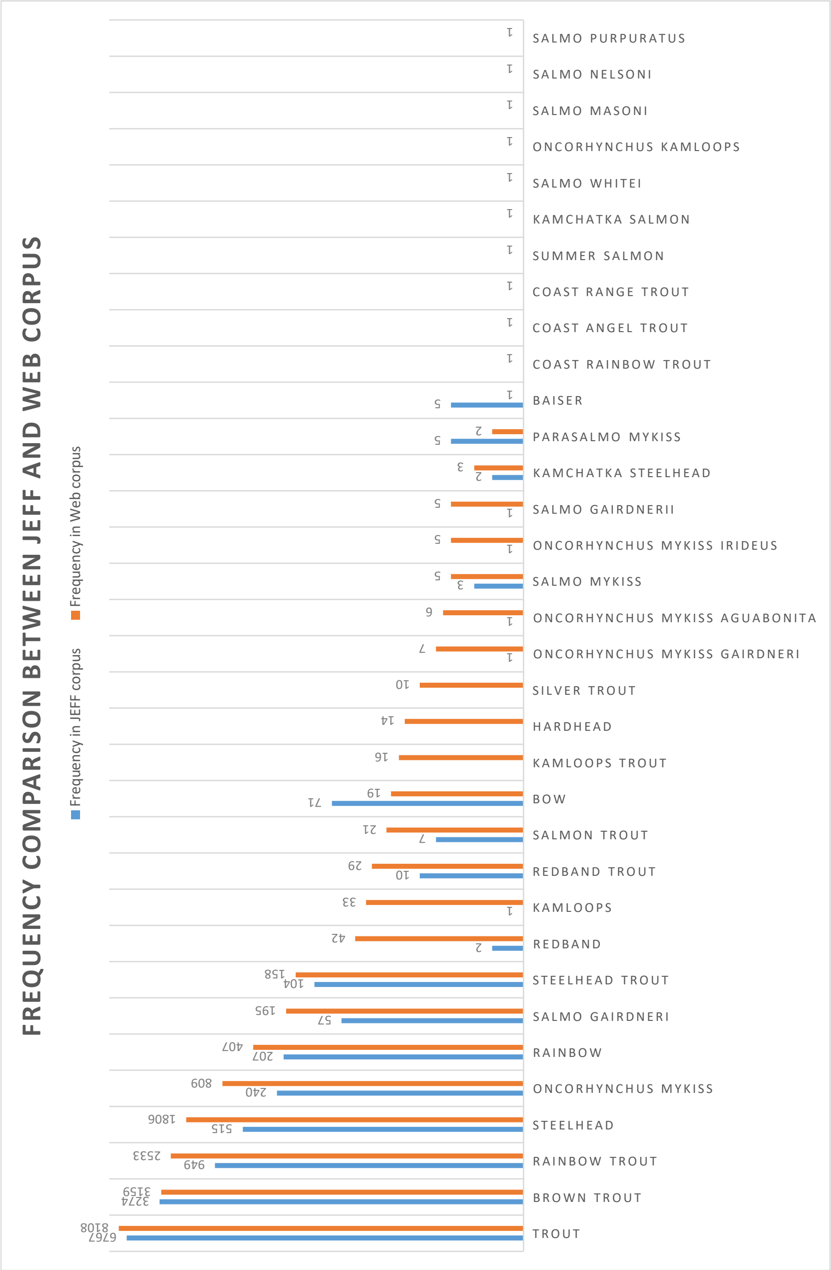


Figure 9.53: Graph showing comparison between JEFF and WEB corpus and name variant frequencies (logarithmic scale)

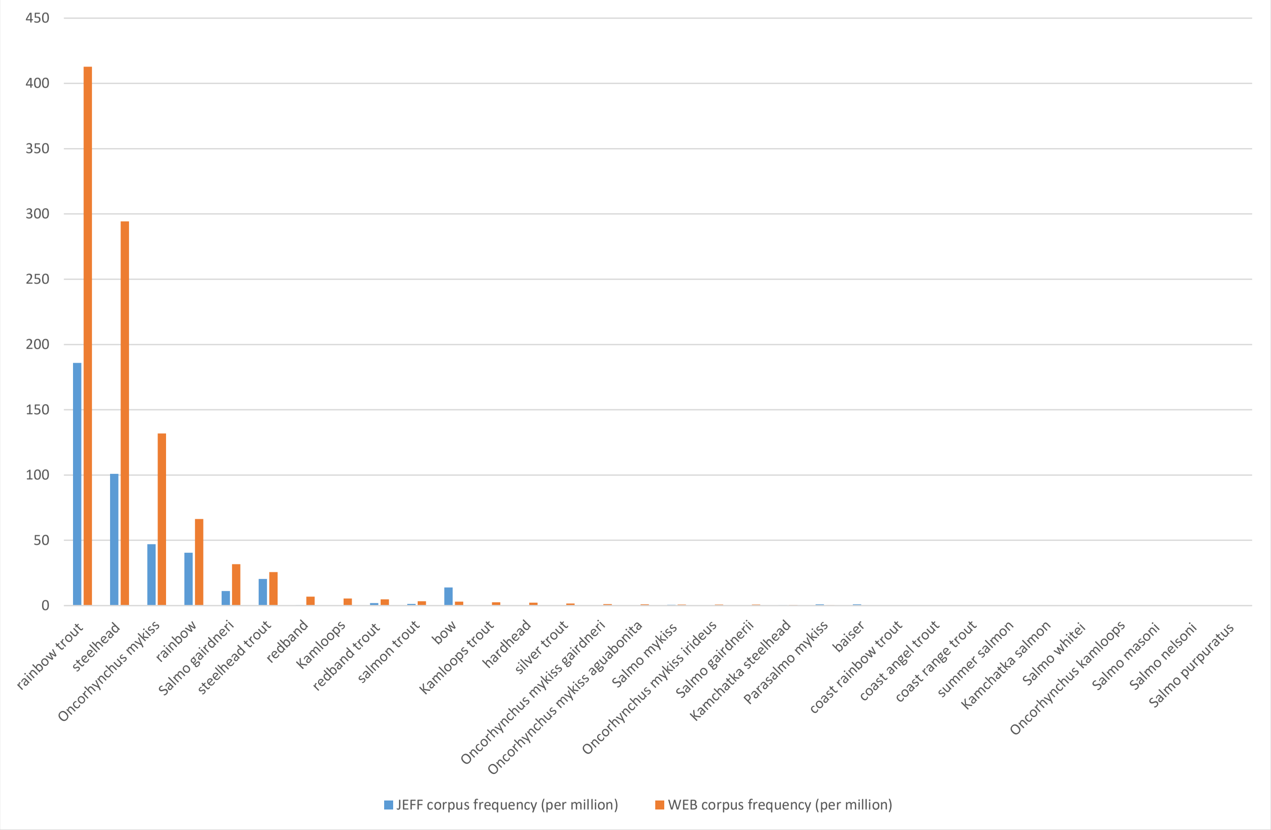


Figure 9.54: Graph showing comparison between JEFF and WEB corpus and name variant frequencies (hits per million)

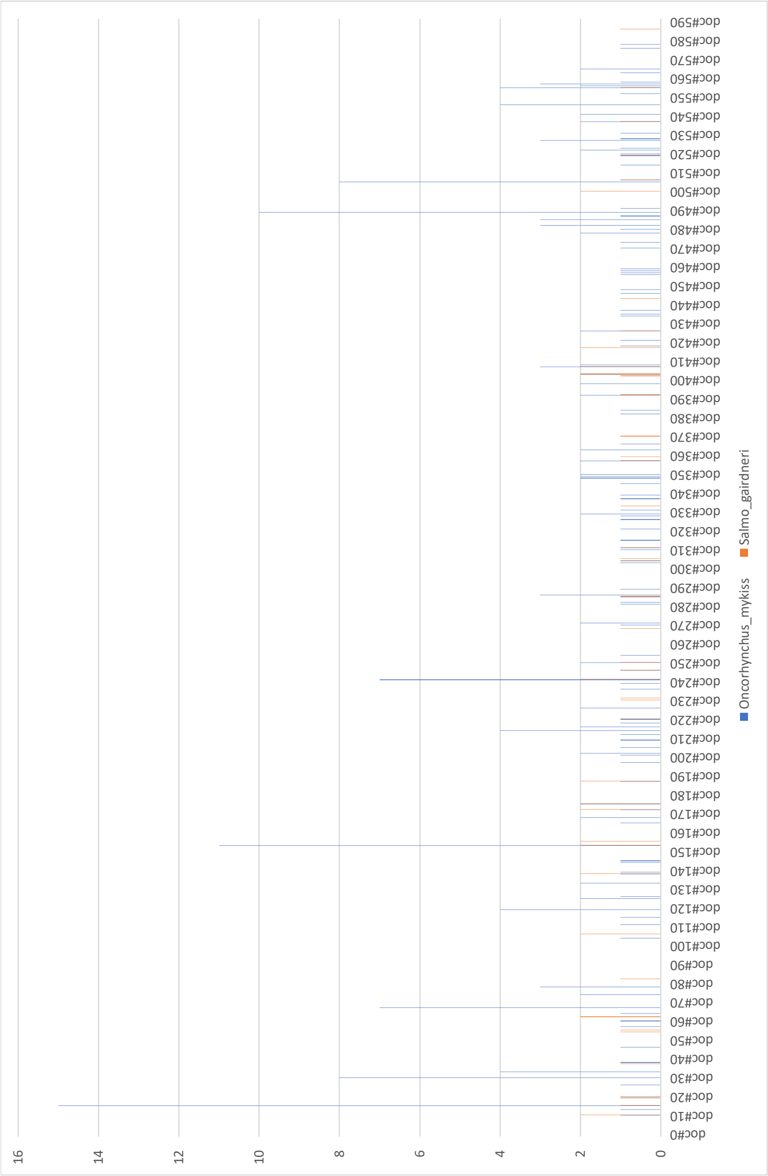


Figure 9.55: JEFF corpus: dispersion for Oncorhynchus mykiss versus Salmo gairdneri [in colour]

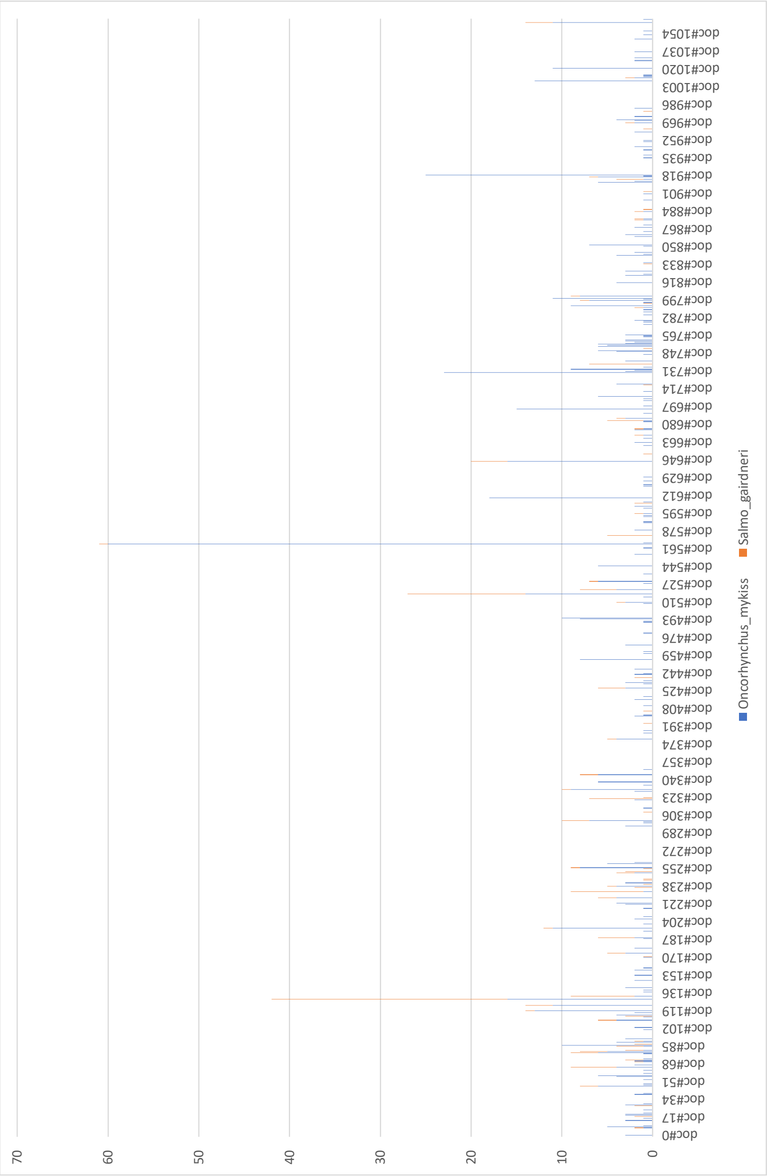


Figure 9.56: WEB corpus: dispersion for Oncorhynchus mykiss versus Salmo gairdneri [in colour]

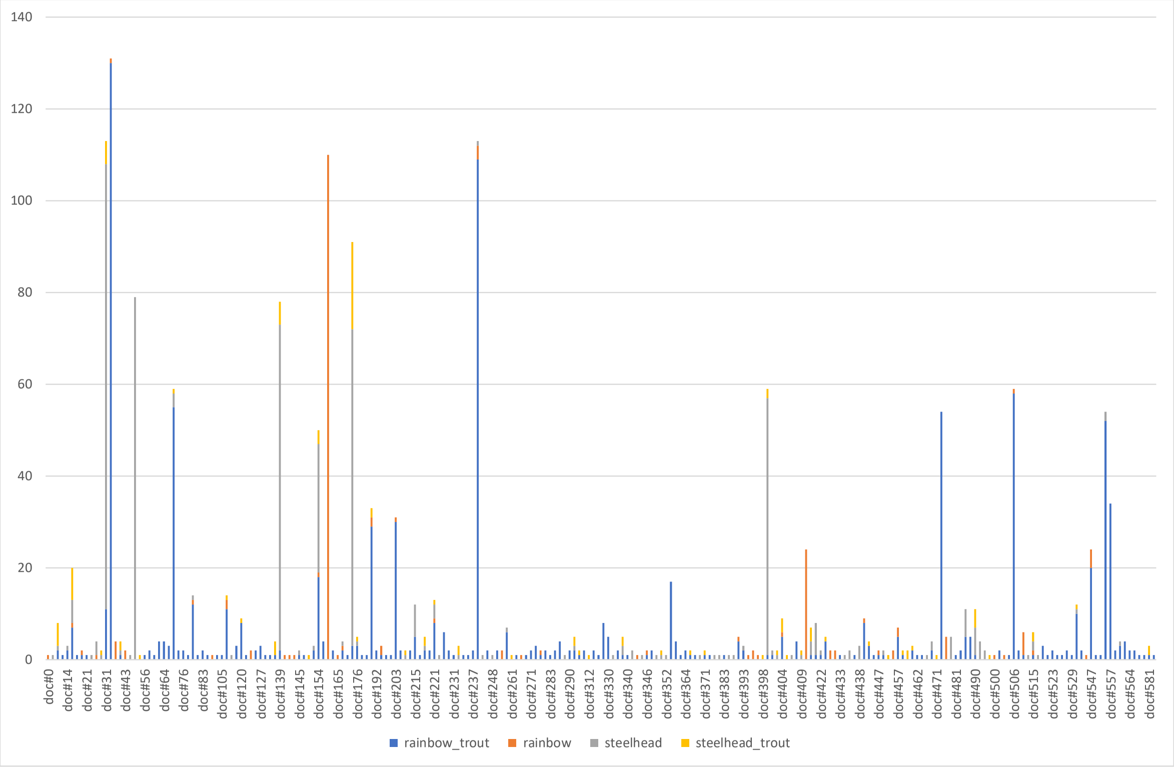


Figure 9.57: JEFF corpus: dispersion comparison steelhead and rainbow trout variants [in colour]

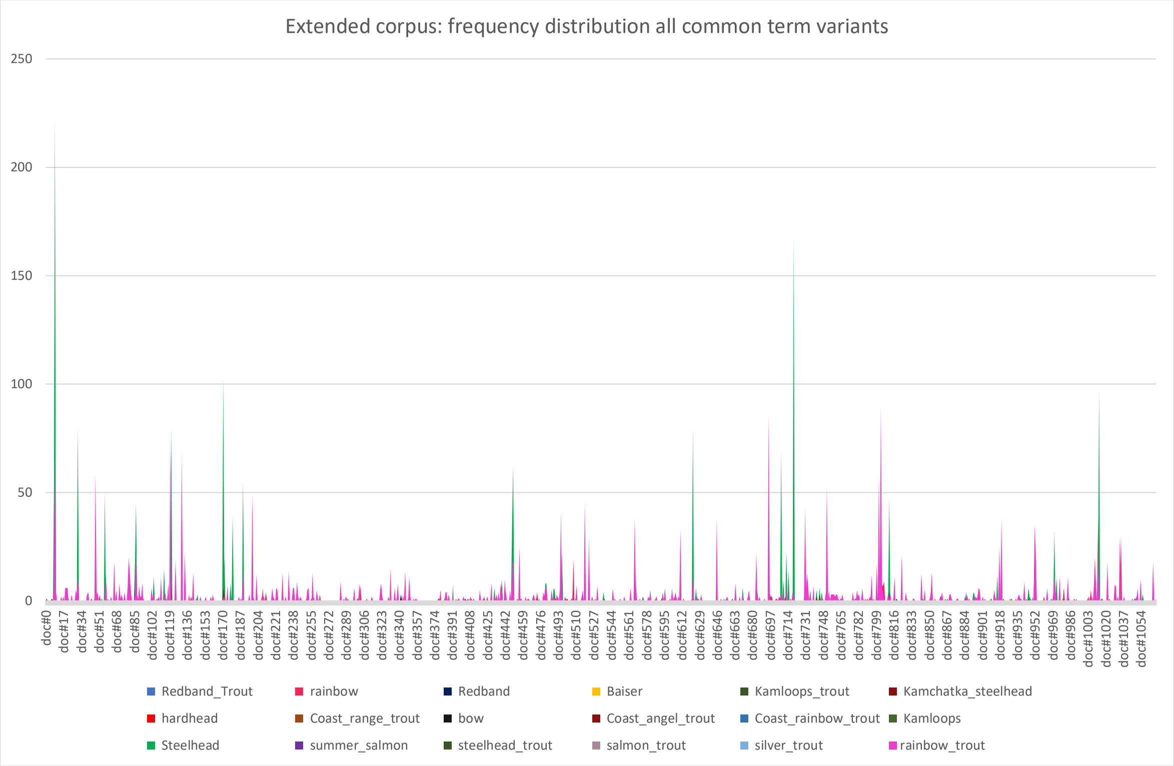


Figure 9.58: WEB corpus: dispersion for steelhead and rainbow trout variants [in colour]

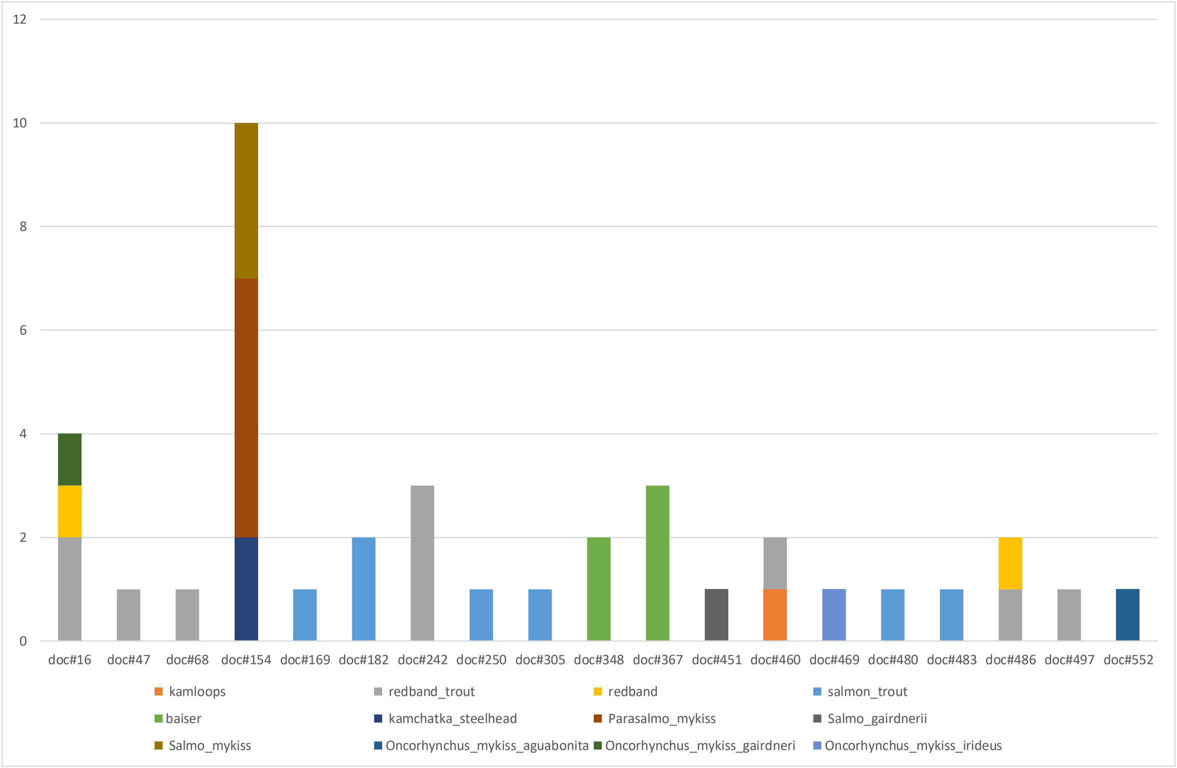


Figure 9.59: JEFF corpus: dispersion for infrequent common and scientific nomenclature variants [in colour]

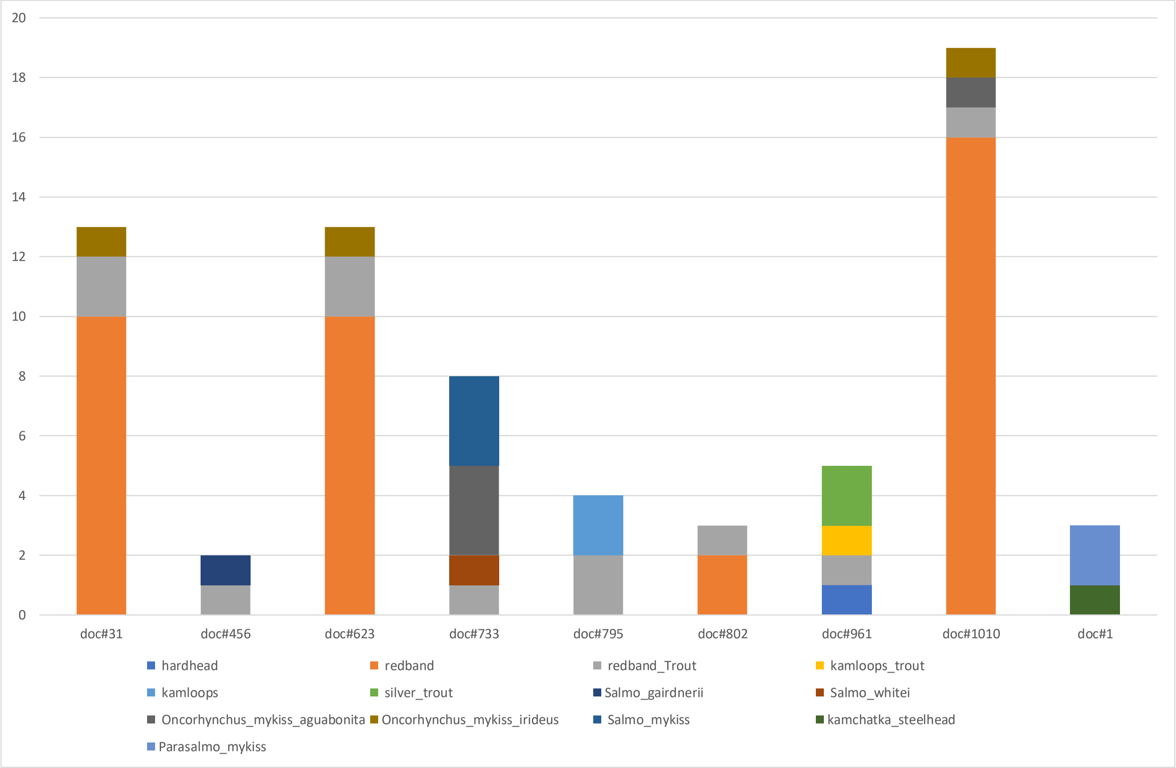


Figure 9.60: WEB corpus: dispersion for infrequent common and scientific nomenclature in which multiple terms are used in the same document [in colour]

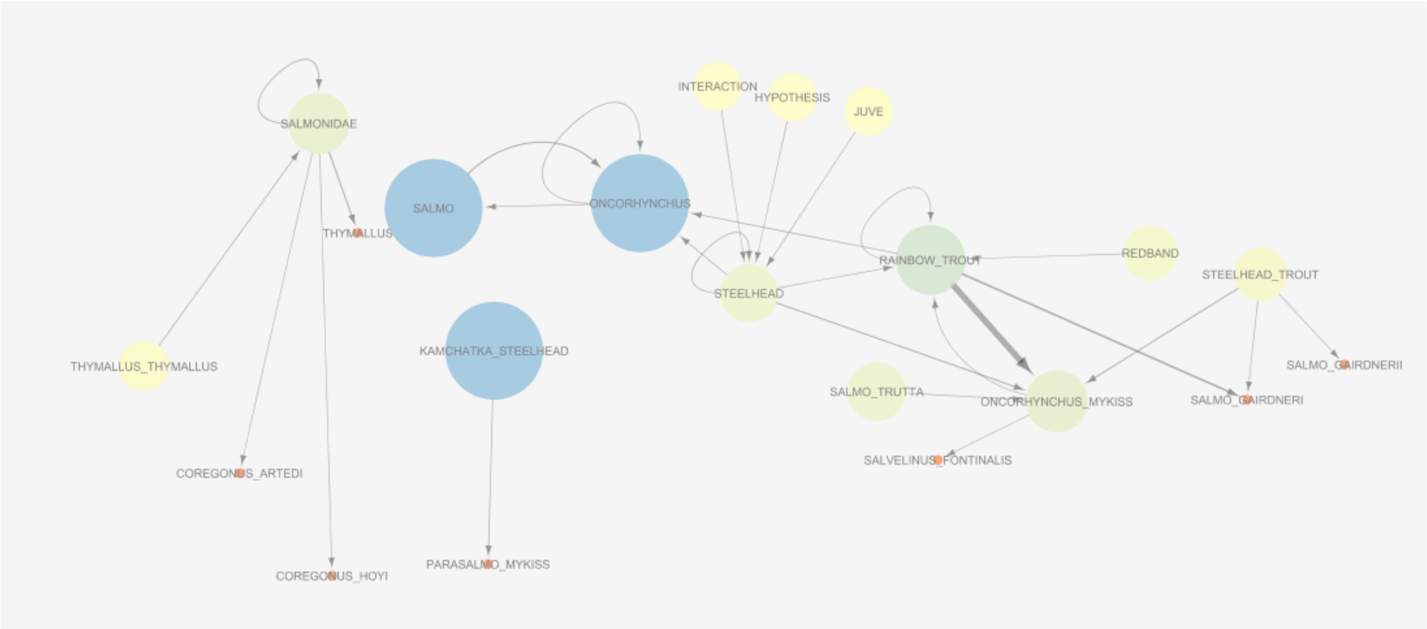


Figure 9.61: JEFF corpus for Oncorhynchus mykiss, filtered to exclude trout and brown trout [in colour]. Here the links between Oncorhynchus mykiss, rainbow trout and steelhead trout can be seen, as well as the links between Salmo gairdneri and both the previously mentioned common names, and Salmo gairdnerii only with steelhead trout. Parasalmo mykiss is seen in a separate part of the graph only linked to Kamchatka steelhead.

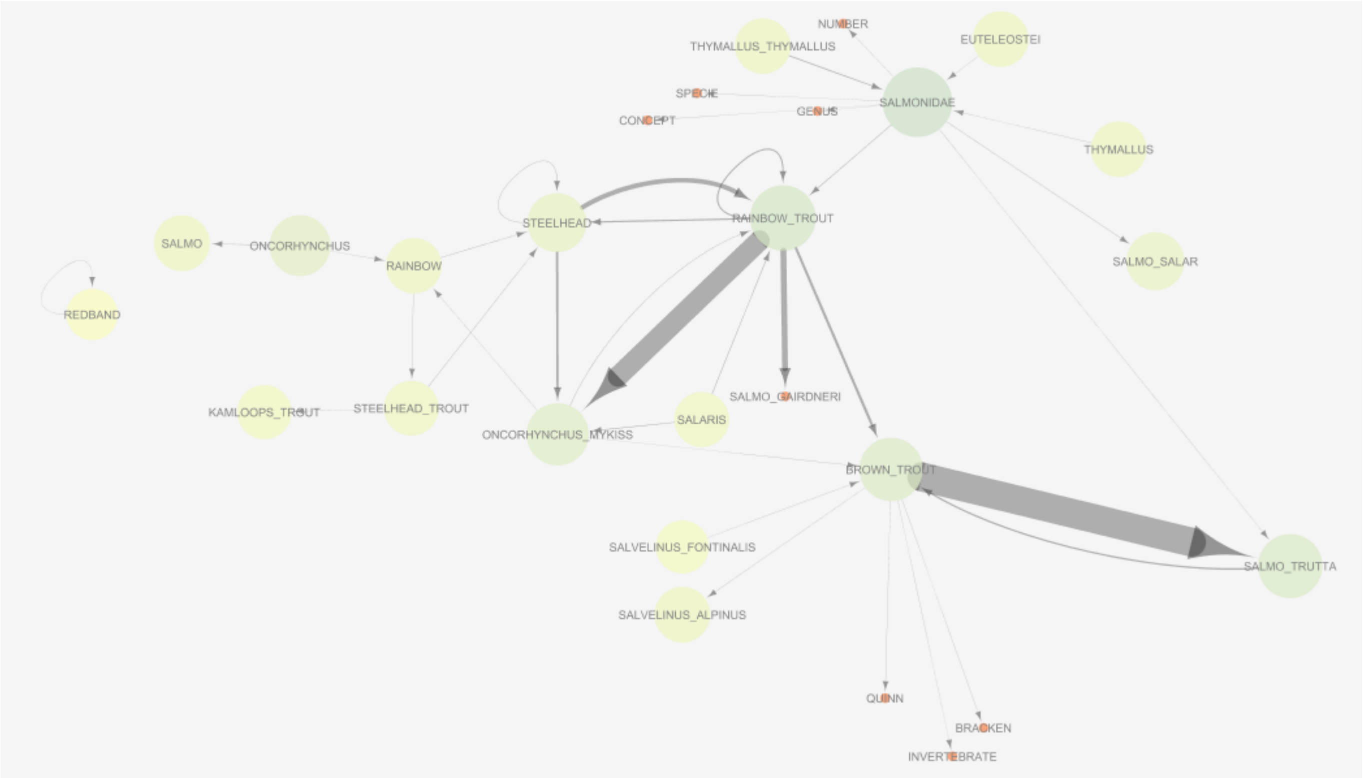


Figure 9.62: WEB corpus, filtered for relations of two or more hits and trout relations removed. Here the strong link between rainbow trout and Oncorhynchus mykiss is seen, as well as a link between steelhead and the accepted name, although not as frequent. Links between the variants rainbow, steelhead and steelhead trout also identified. Only rainbow trout is linked to Salmo gairdneri and the strong links between Salmo trutta and brown trout are identified.

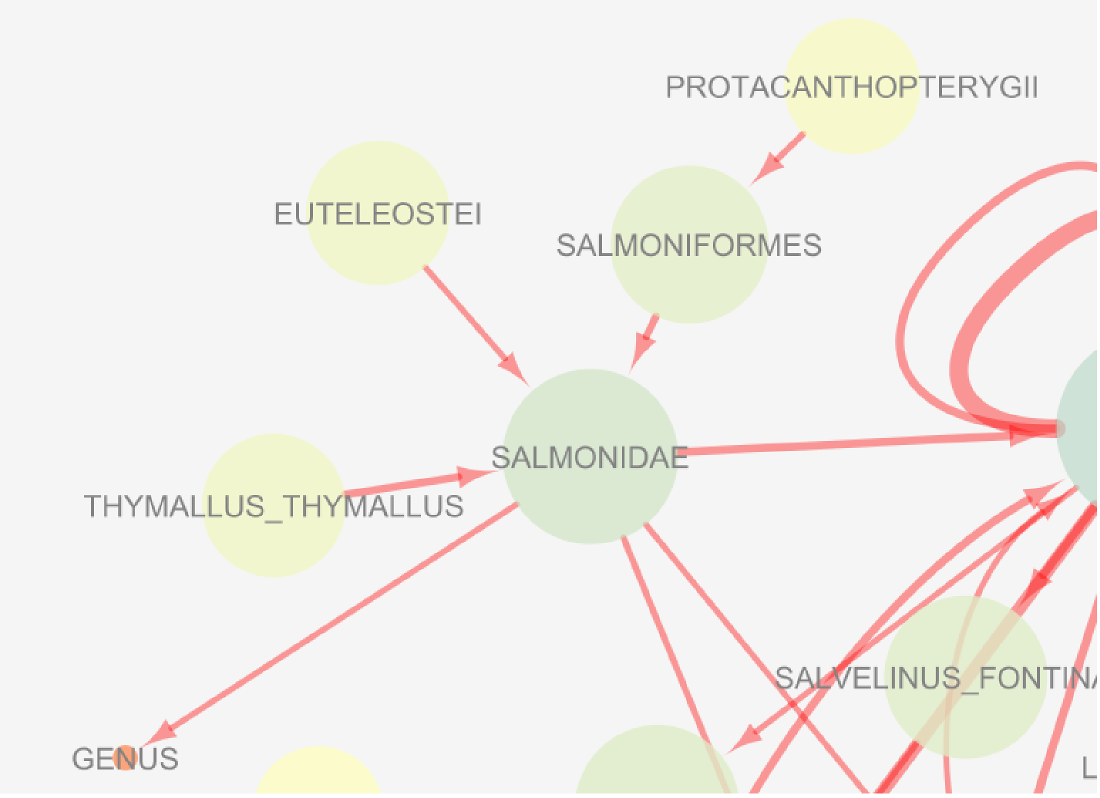


Figure 9.63: WEB corpus for Oncorhynchus mykiss, no lower_case hierarchy. Here the increased hierarchy is identified in comparison with the lower-cased corpus, with Protacanthopterygii.

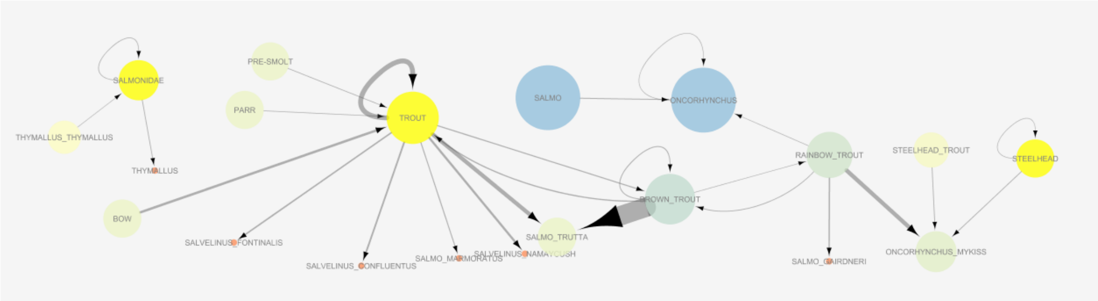


Figure 9.64: JEFF corpus for Oncorhynchus mykiss: filtered for relations of two or more hits, salience 9.5 and with hubs highlighted in yellow [in colour]

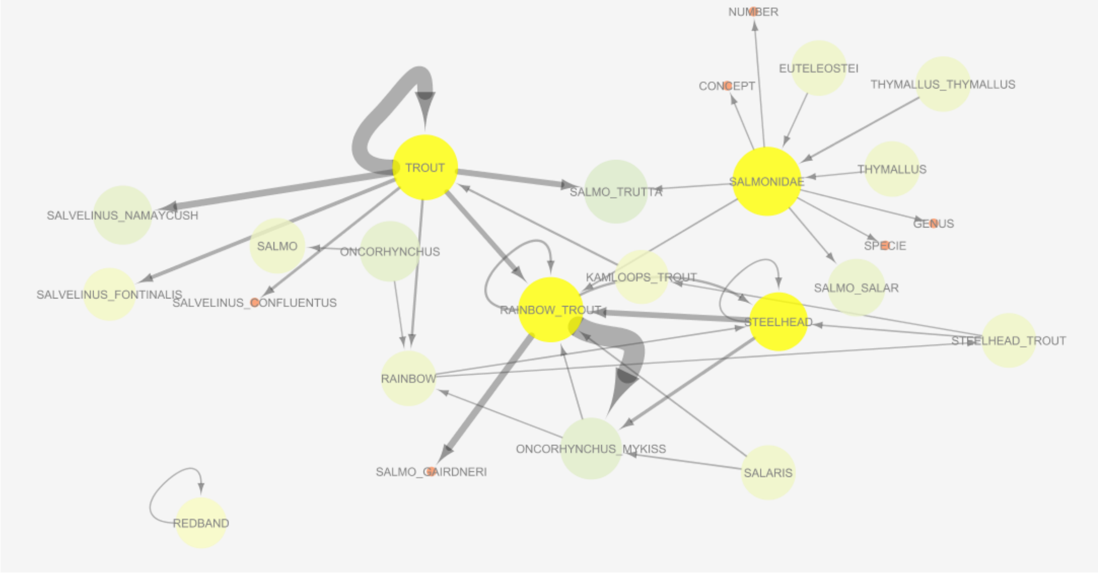
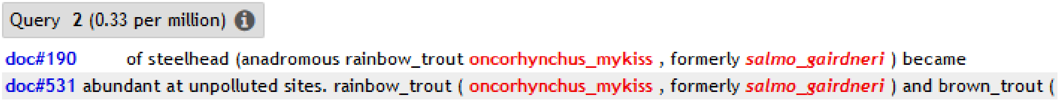


Figure 9.65: WEB corpus for Oncorhynchus mykiss: filtered for relations of two or more hits, salience 9.5 and with hubs highlighted in yellow [in colour]



Query 2 (0.33 per million)

doc#190 of steelhead (anadromous rainbow_trout oncorhynchus_mykiss , formerly *salmo_gairdneri*) became
doc#531 abundant at unpolluted sites. rainbow_trout (oncorhynchus_mykiss , formerly *salmo_gairdneri*) and brown_trout (

Figure 9.66: Concordance lines mentioning Salmo gairdneri as former term for Oncorhynchus mykiss

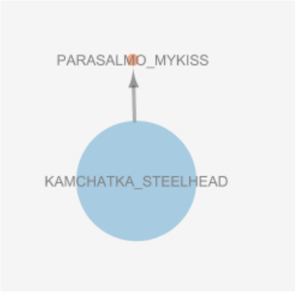


Figure 9.67: JEFF corpus: Oncorhynchus mykiss and Parasalmo mykiss relation

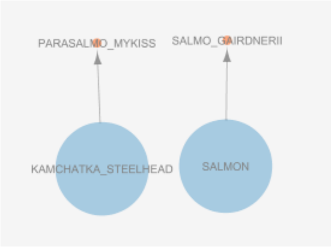
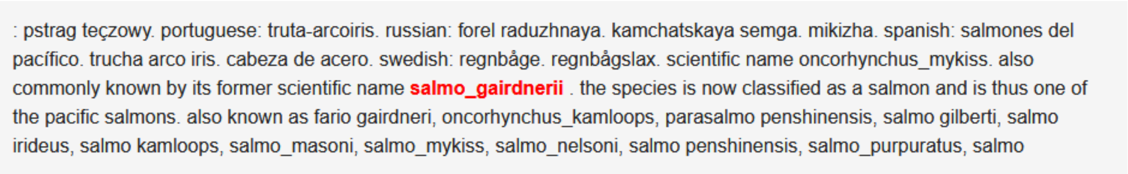
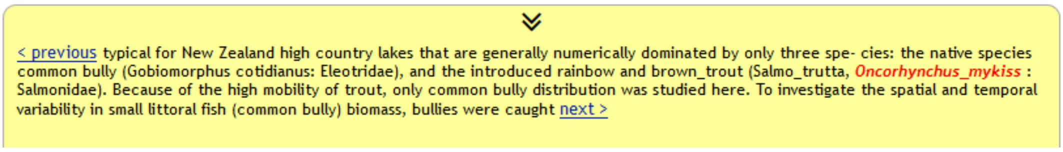


Figure 9.68: WEB corpus: Oncorhynchus mykiss, Parasalmo mykiss and Salmo gairdnerii



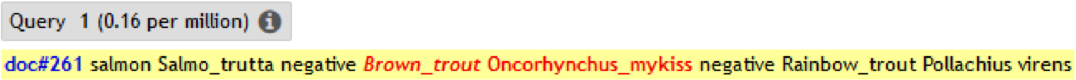
: pstrag tęczowy. portuguese: truta-arcoiris. russian: forel raduzhnaya. kamchatskaya semga. mikizha. spanish: salmones del pacífico. trucha arco iris. cabeza de acero. swedish: regnbåge. regnbågslax. scientific name oncorhynchus_mykiss. also commonly known by its former scientific name **salmo_gairdnerii** . the species is now classified as a salmon and is thus one of the pacific salmons. also known as fario gairdneri, oncorhynchus_kamloops, parasalmo penshinensis, salmo gilberti, salmo irideus, salmo kamloops, salmo_masoni, salmo_mykiss, salmo_nelsoni, salmo penshinensis, salmo_purpuratus, salmo

Figure 9.69: WEB corpus: Oncorhynchus mykiss and Salmo gairdnerii concordance



< previous typical for New Zealand high country lakes that are generally numerically dominated by only three spe- cies: the native species common bully (Gobiomorphus cotidianus: Eleotridae), and the introduced rainbow and brown_trout (Salmo_trutta, ***Oncorhynchus_mykiss*** : Salmonidae). Because of the high mobility of trout, only common bully distribution was studied here. To investigate the spatial and temporal variability in small littoral fish (common bully) biomass, bullies were caught next >

Figure 9.70: Concordance showing incorrect linking between brown trout and Oncorhynchus mykiss in JEFF corpus



Query 1 (0.16 per million)

doc#261 salmon Salmo_trutta negative *Brown_trout* Oncorhynchus_mykiss negative Rainbow_trout Pollachius virens

Figure 9.71: Concordance showing incorrect linking between brown trout and Oncorhynchus mykiss in WEB corpus

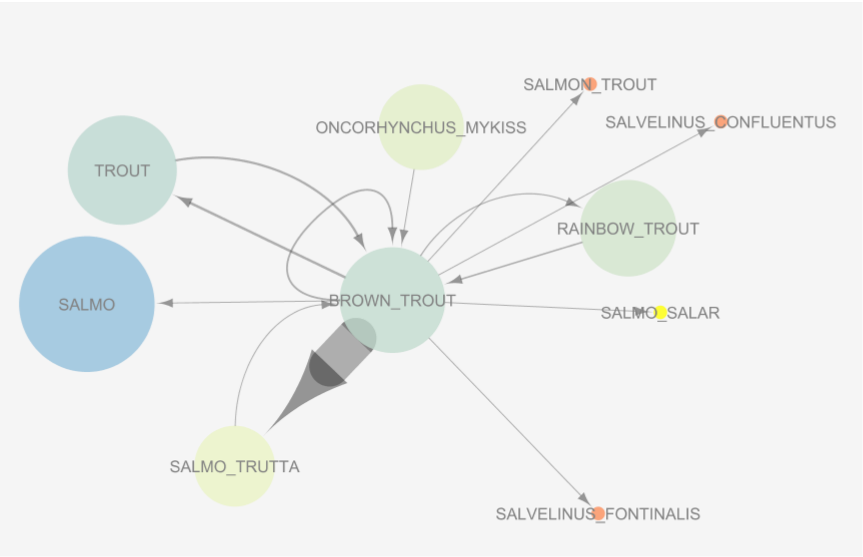


Figure 9.72: JEFF corpus: incorrect linking between brown trout and Oncorhynchus mykiss

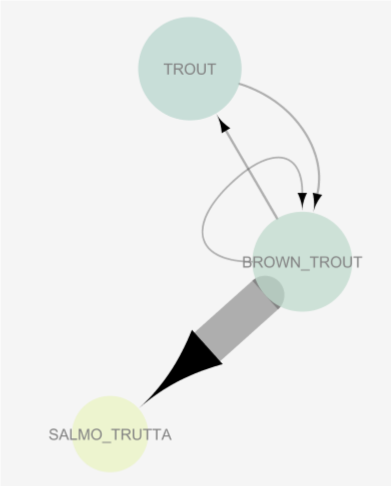


Figure 9.73: JEFF corpus: brown trout profile filtered for 5 or more hits

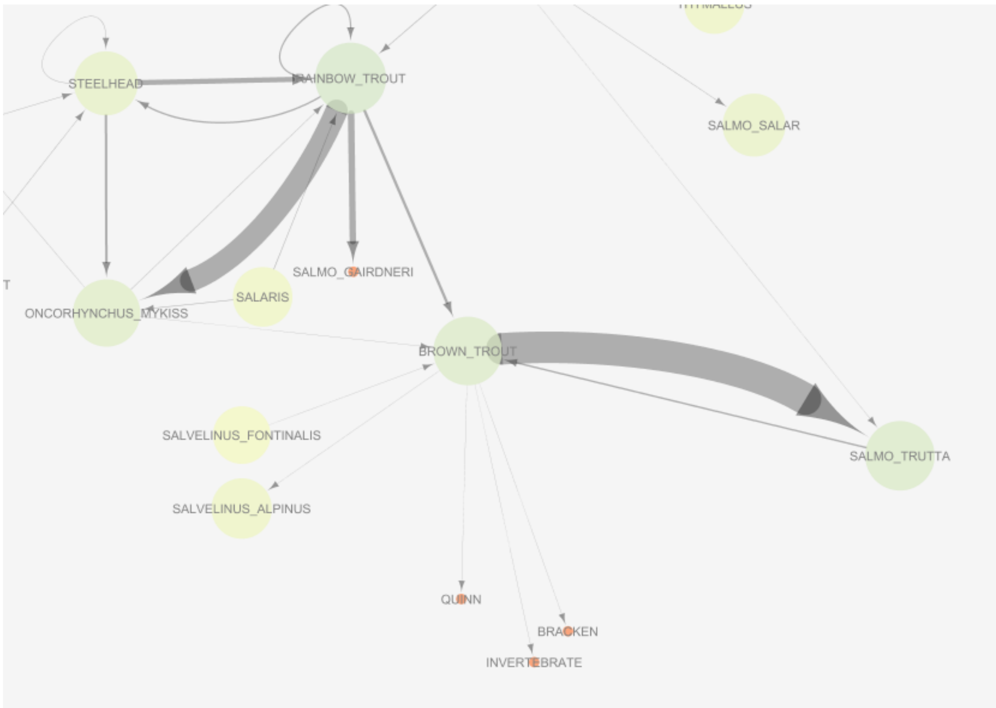


Figure 9.74: WEB corpus: relative strength of relation between brown trout and Oncorhynchus mykiss and Salmo trutta

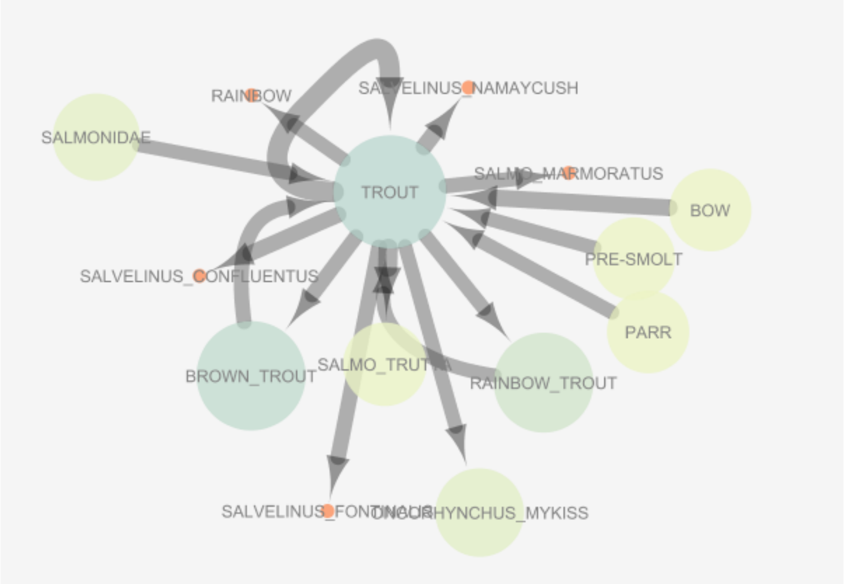


Figure 9.75: JEFF corpus: trout profile as hub

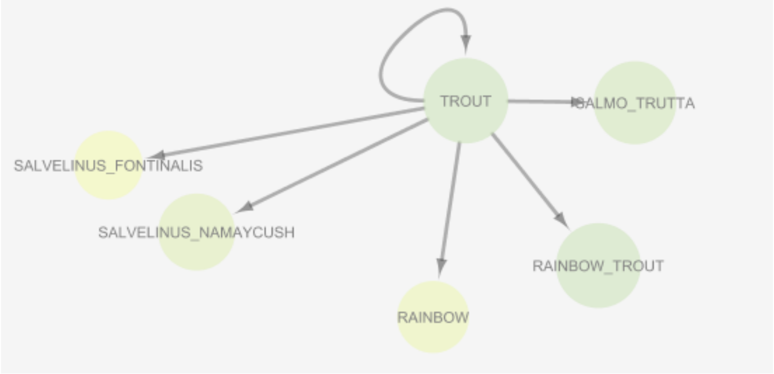


Figure 9.76: WEB corpus: trout profile as hub

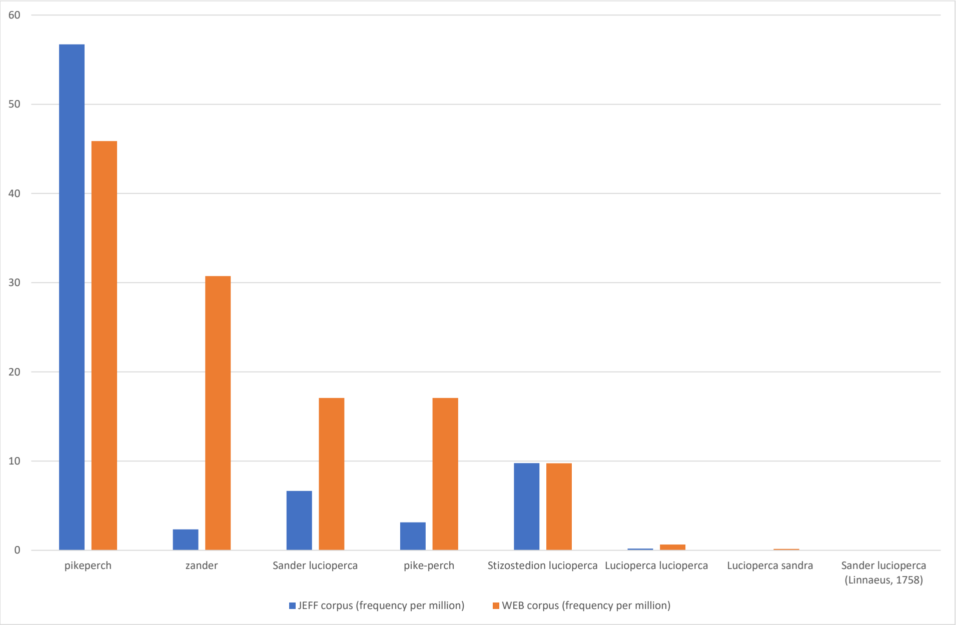


Figure 9.77: Sander lucioperca: comparison of variants across JEFF and WEB corpora (hits per million)

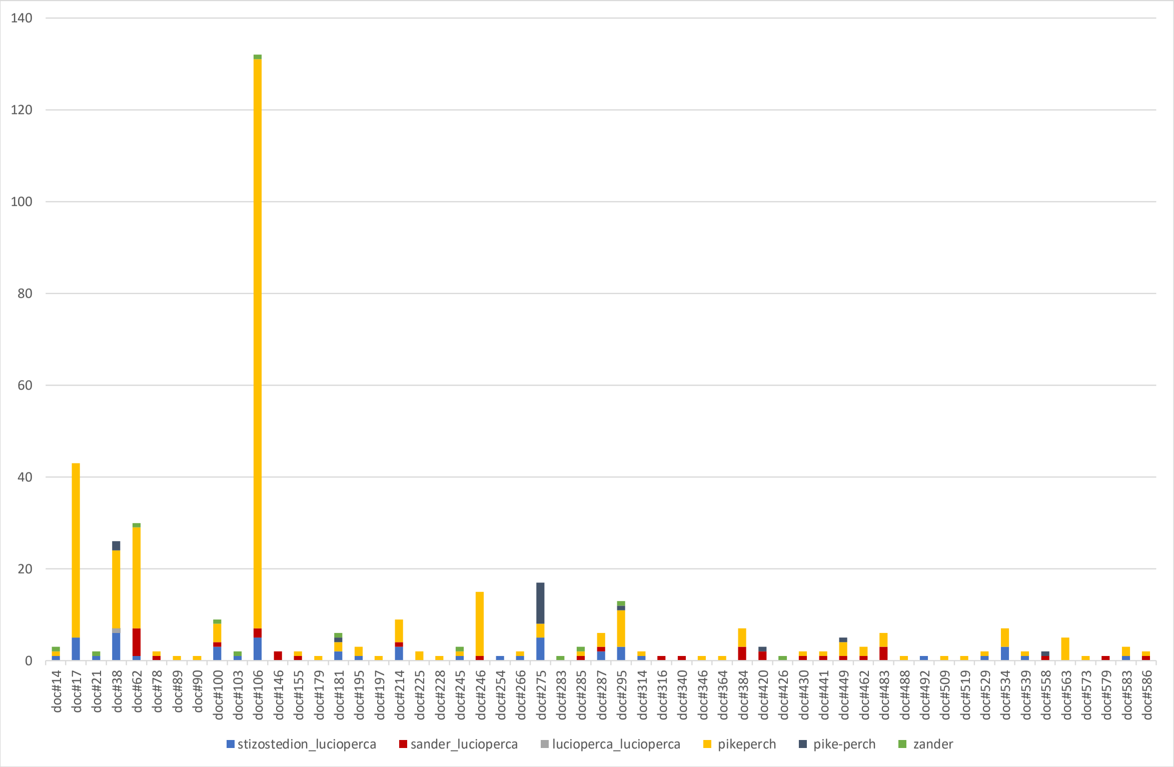


Figure 9.78: Graph showing JEFF corpus dispersion graph for all scientific and common variants of Sander lucioperca.

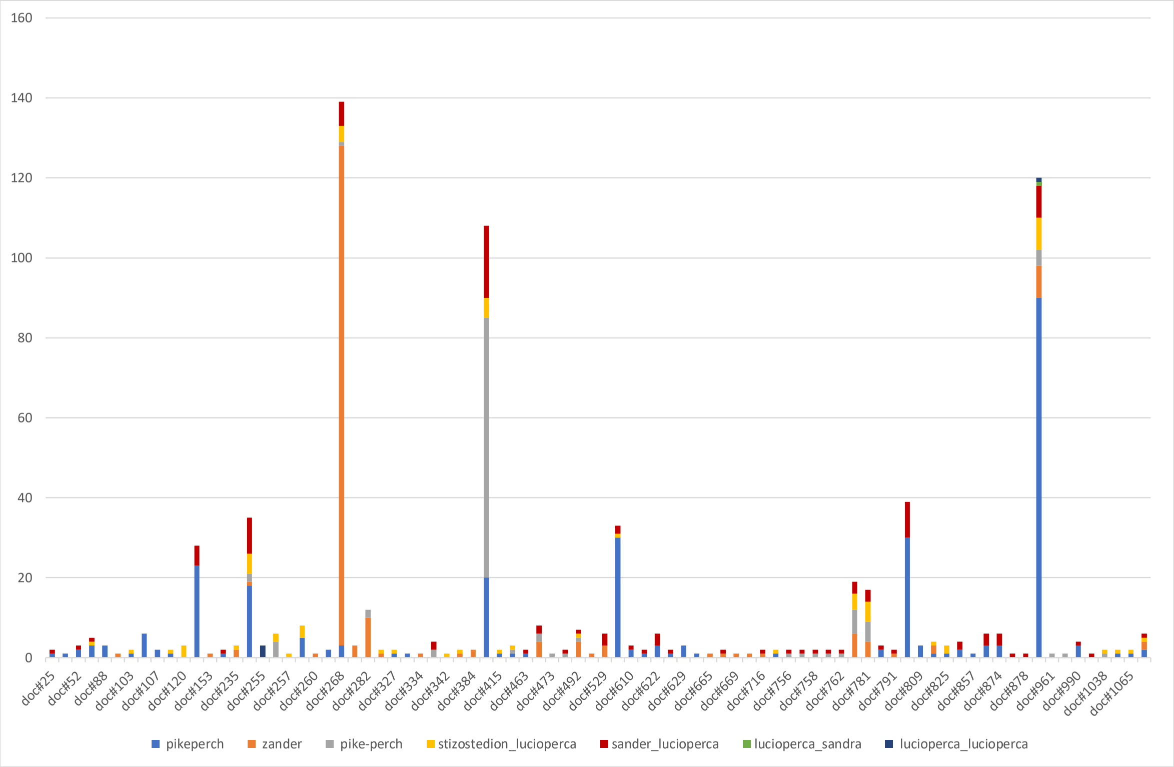


Figure 9.79: Graph showing WEB corpus dispersion graph for all scientific and common variants of Sander lucioperca.

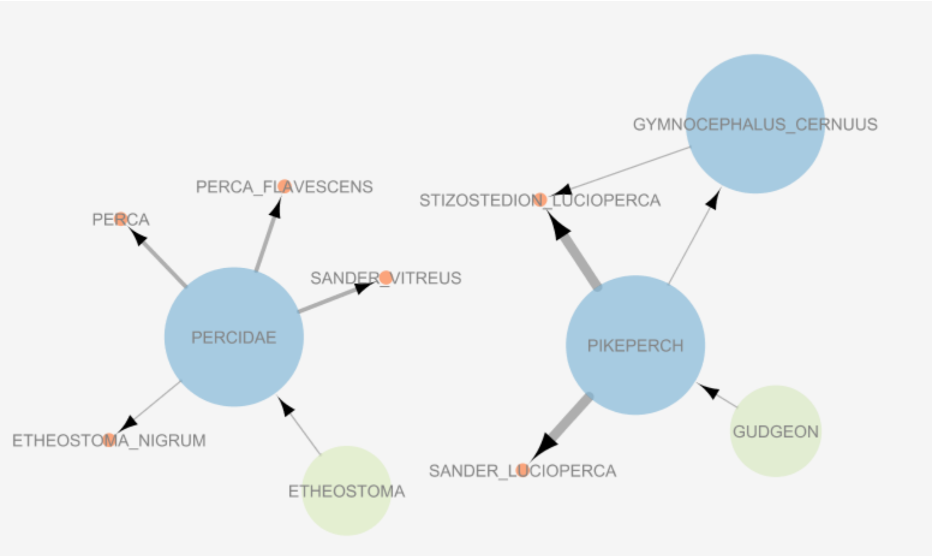
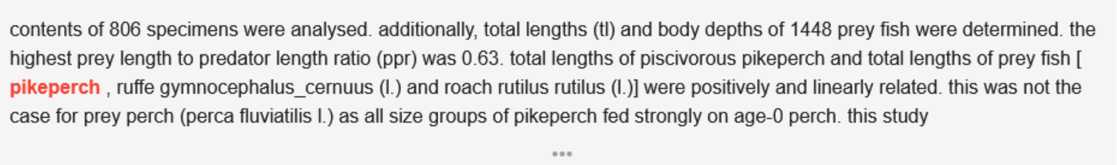


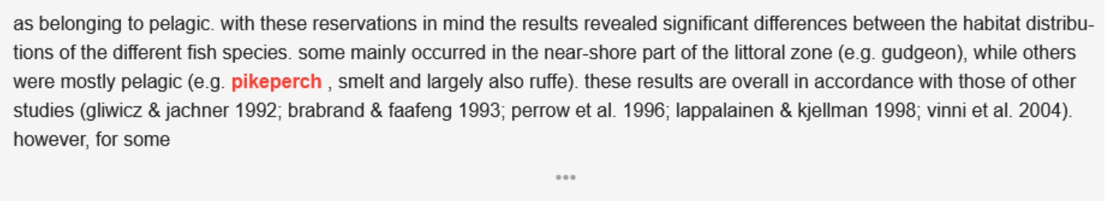
Figure 9.80: JEFF corpus: relation network graph for Sander lucioperca. Pikeperch forms a hub node which collates both the variants of Sander lucioperca acknowledged, with higher numbers of for each of these than the other linked nodes. Percidae forms another hub with other Percidae species.



contents of 806 specimens were analysed. additionally, total lengths (tl) and body depths of 1448 prey fish were determined. the highest prey length to predator length ratio (ppr) was 0.63. total lengths of piscivorous pikeperch and total lengths of prey fish [**pikeperch** , ruffe gymnocephalus_cernuus (l.) and roach rutilus rutilus (l.)] were positively and linearly related. this was not the case for prey perch (perca fluviatilis l.) as all size groups of pikeperch fed strongly on age-0 perch. this study

•••

Figure 9.81: JEFF corpus: Gymnocephalus cernuus and pikeperch concordance



as belonging to pelagic. with these reservations in mind the results revealed significant differences between the habitat distributions of the different fish species. some mainly occurred in the near-shore part of the littoral zone (e.g. gudgeon), while others were mostly pelagic (e.g. **pikeperch** , smelt and largely also ruffe). these results are overall in accordance with those of other studies (gliwicz & jachner 1992; brabrand & faafeng 1993; perrow et al. 1996; lappalainen & kjellman 1998; vinni et al. 2004). however, for some

•••

Figure 9.82: JEFF corpus: Gudgeon and pikeperch concordance

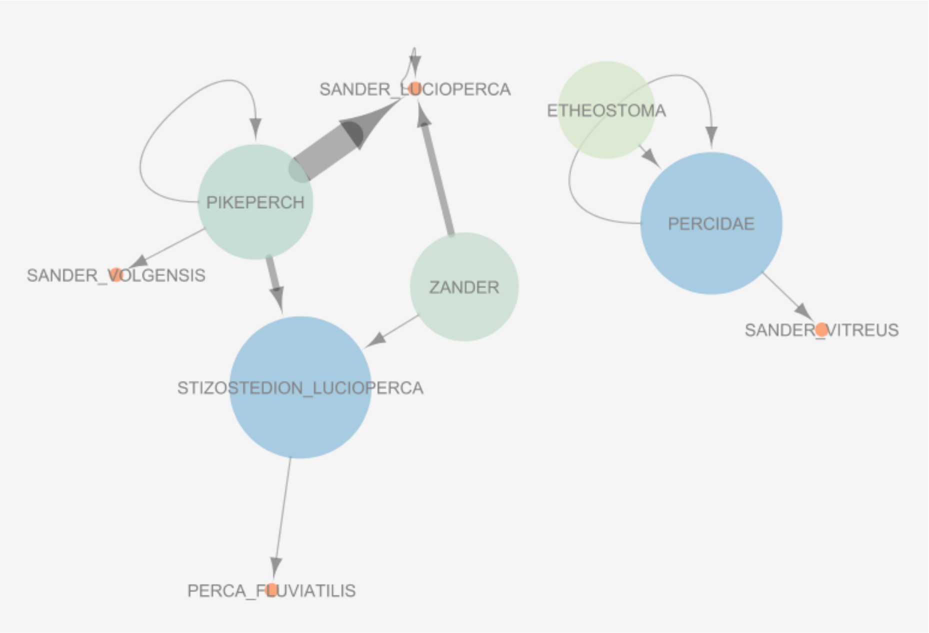
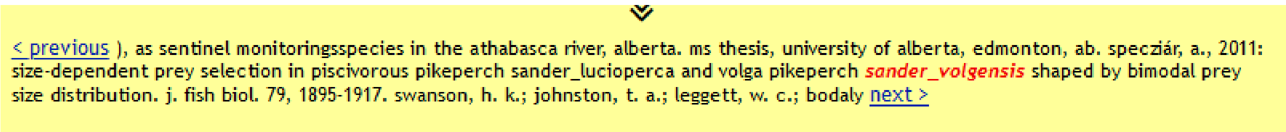


Figure 9.83: WEB corpus: relation network graph for Sander lucioperca. Here pikeperch is more strongly linked to Sander lucioperca than Stizostedion lucioperca, and zander is also linked to both. Here Sander volgensis is also linked to pikeperch as discussed.



< previous), as sentinel monitoringsspecies in the athabasca river, alberta. ms thesis, university of alberta, edmonton, ab. specziár, a., 2011: size-dependent prey selection in piscivorous pikeperch sander_lucioperca and volga pikeperch ***sander_volgensis*** shaped by bimodal prey size distribution. j. fish biol. 79, 1895-1917. swanson, h. k.; johnston, t. a.; leggett, w. c.; bodaly next >

Figure 9.84: WEB corpus: Sander volgensis and pikeperch concordance

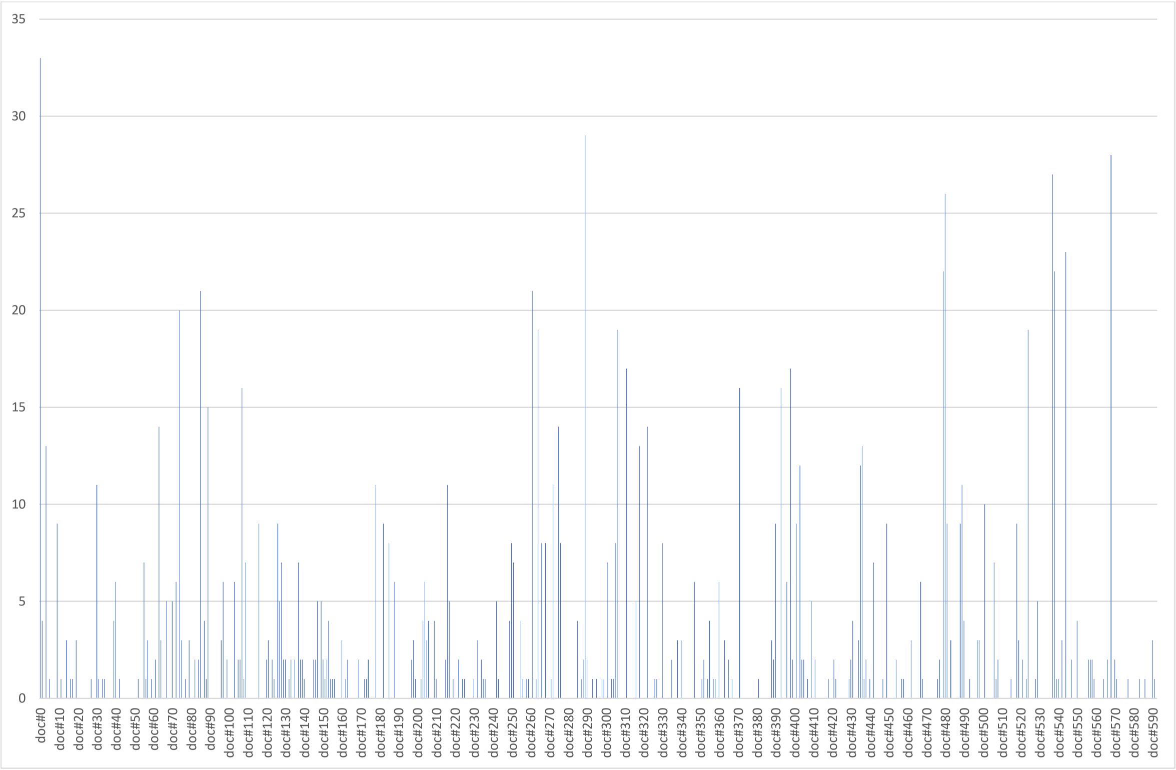


Figure 9.85: Graph showing JEFF corpus dispersion graph for all scientific nomenclature variants of Salmo trutta [in colour]

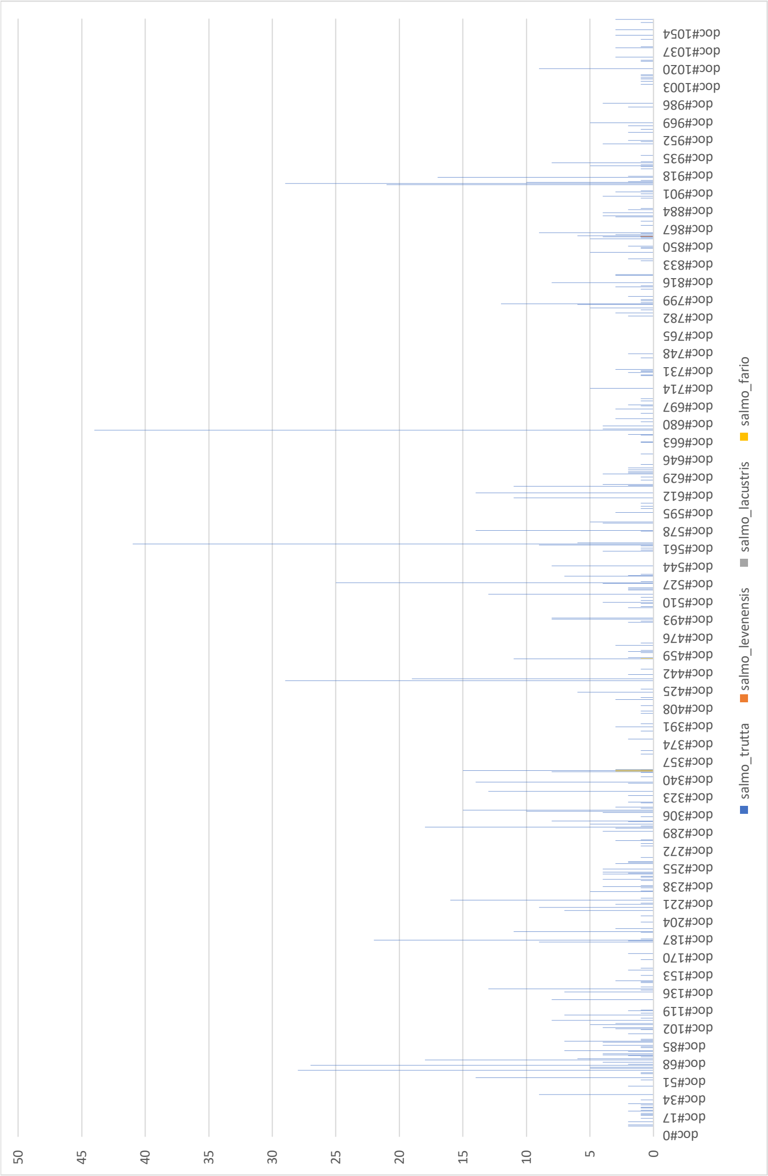


Figure 9.86: Graph showing WEB corpus dispersion graph for all scientific nomenclature variants of Salmo trutta [in colour]

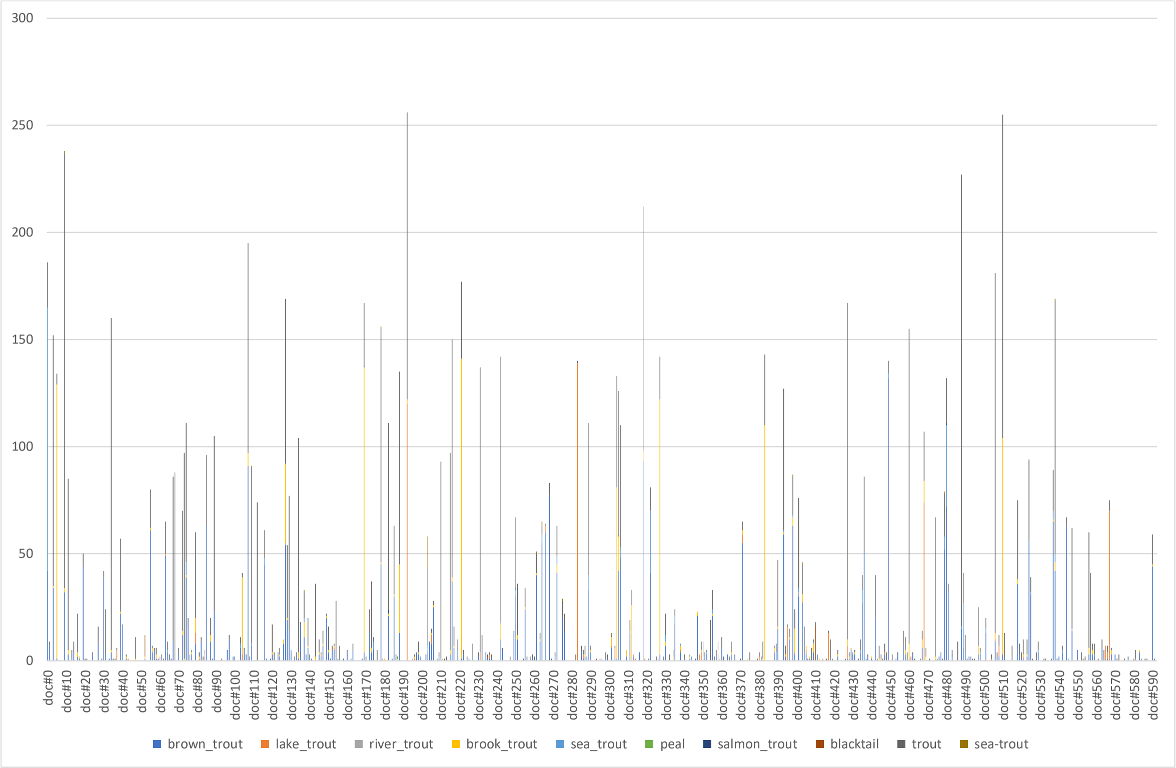


Figure 9.87: Graph showing JEFF corpus dispersion graph for all vernacular variants of Salmo trutta [in colour]

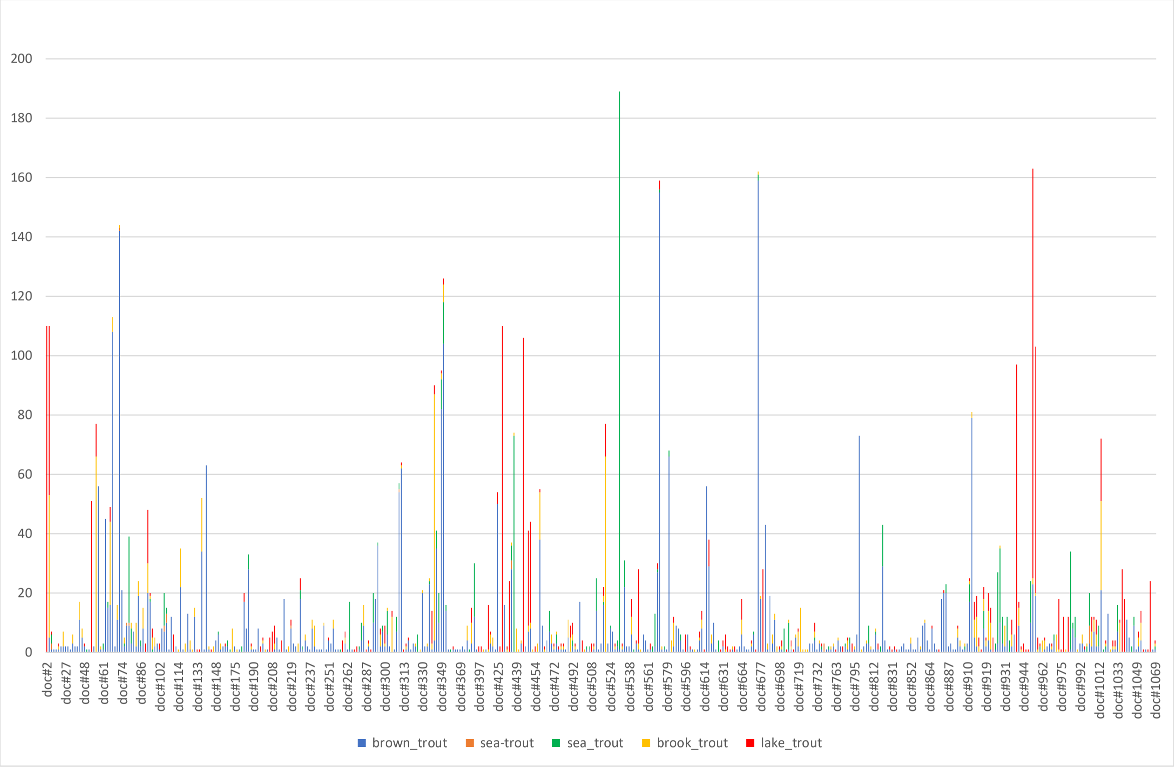


Figure 9.88: Graph showing WEB corpus dispersion graph for all vernacular variants of Salmo trutta [in colour]

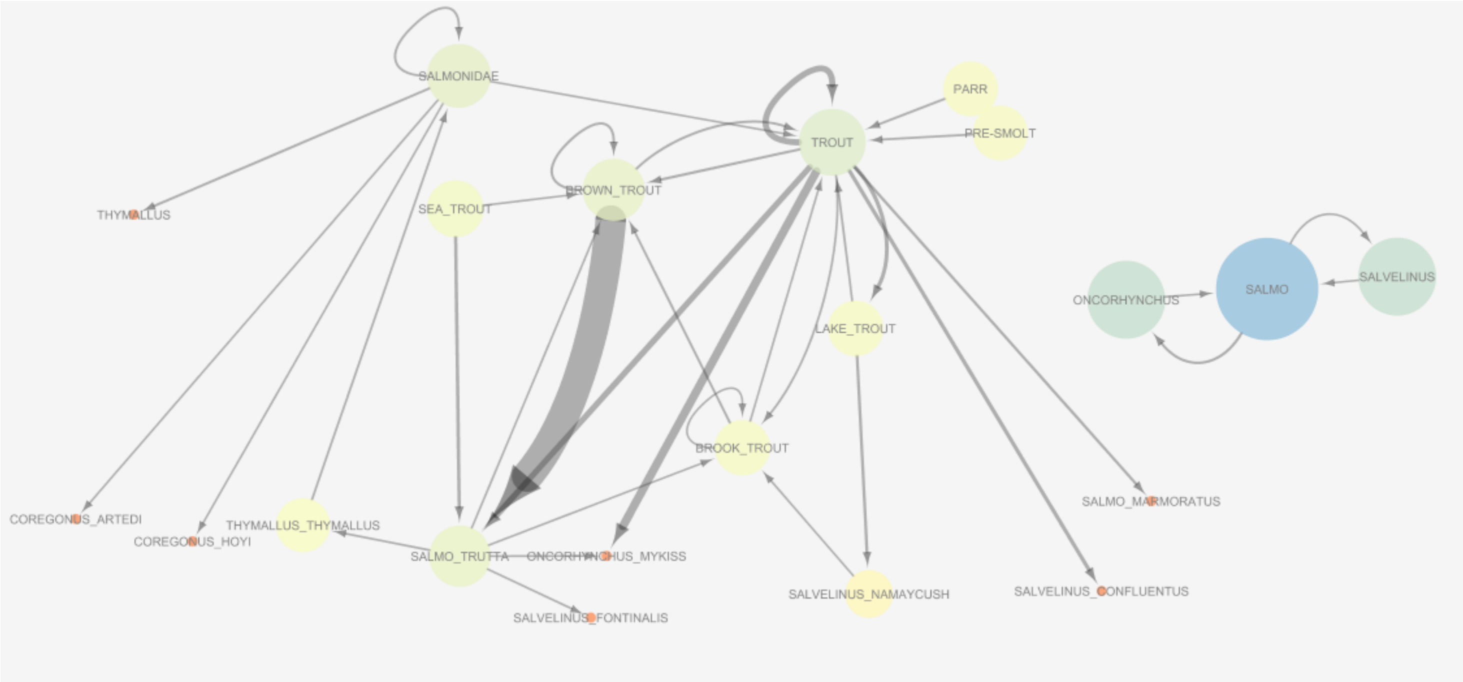


Figure 9.89: JEFF corpus: relation network graph for Salmo trutta, salience 9.5. Here you can see that brown trout is the most strongly linked term to Salmo trutta, whereas sea trout is also linked to a lesser extent with Salmo trutta and brown trout. Other common names such as brook trout has links to many different terms, including Salvelinus namaycush and Salmo trutta. Lake trout is also linked to Salvelinus namaycush.

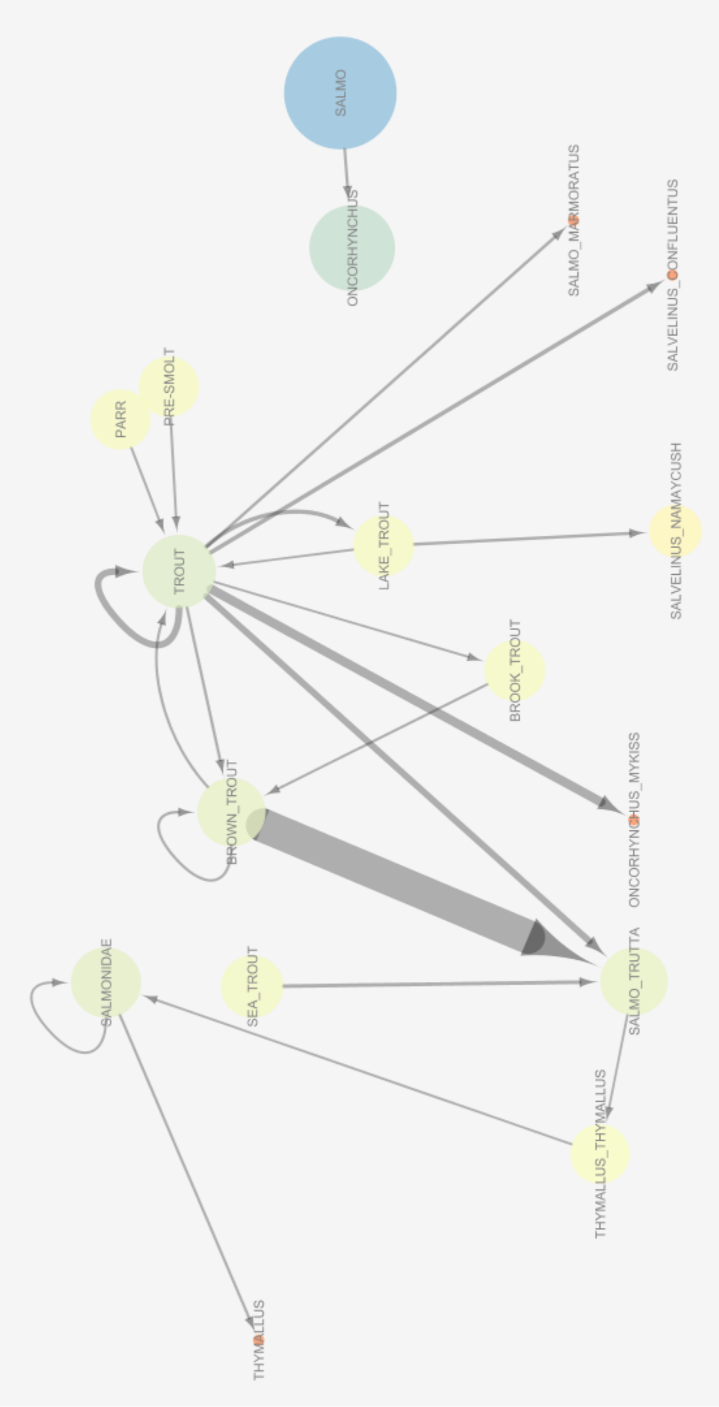


Figure 9.90: JEFF corpus: relation network graph for Salmo trutta, salience 9.5, hits 2 or over. The link between brook trout and Salvelinus namaycush is now gone, with only the lake trout link remaining, showing the usefulness of filtering in clearing the picture.

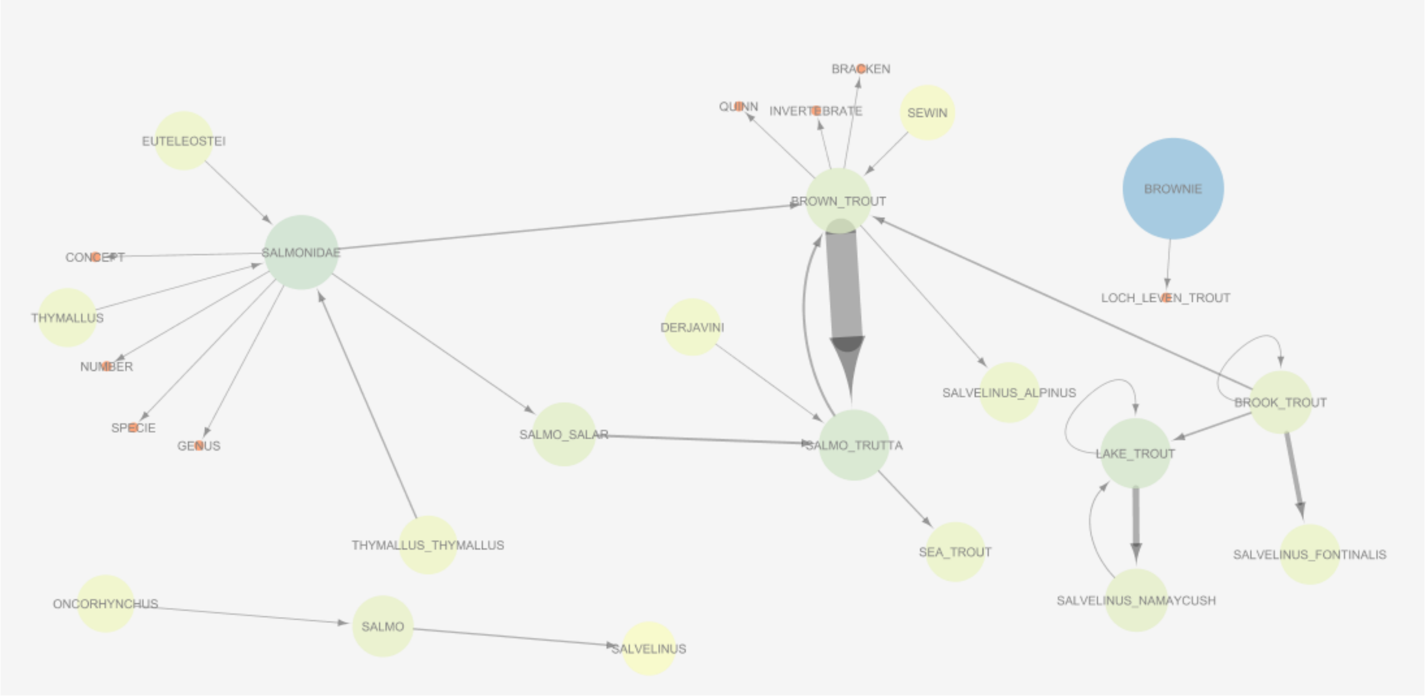
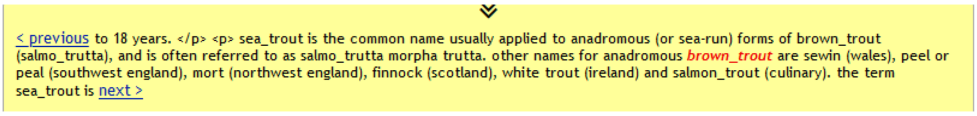


Figure 9.91: WEB corpus: relation network graph for Salmo trutta, salience 11, hits 2 or over. Brown trout again very clearly linked to Salmo trutta, again with sea trout too, but less so. In this graph the lake trout is linked again to Salvelinus namaycush whereas brook trout is linked to Salvelinus fontinalis.



< previous to 18 years. </p> <p> sea_trout is the common name usually applied to anadromous (or sea-run) forms of brown_trout (salmo_trutta), and is often referred to as salmo_trutta morpha trutta. other names for anadromous *brown_trout* are sewin (wales), peel or peal (southwest england), mort (northwest england), finnock (scotland), white trout (ireland) and salmon_trout (culinary). the term sea_trout is next >

Figure 9.92: WEB corpus: concordance for sewin with brown trout

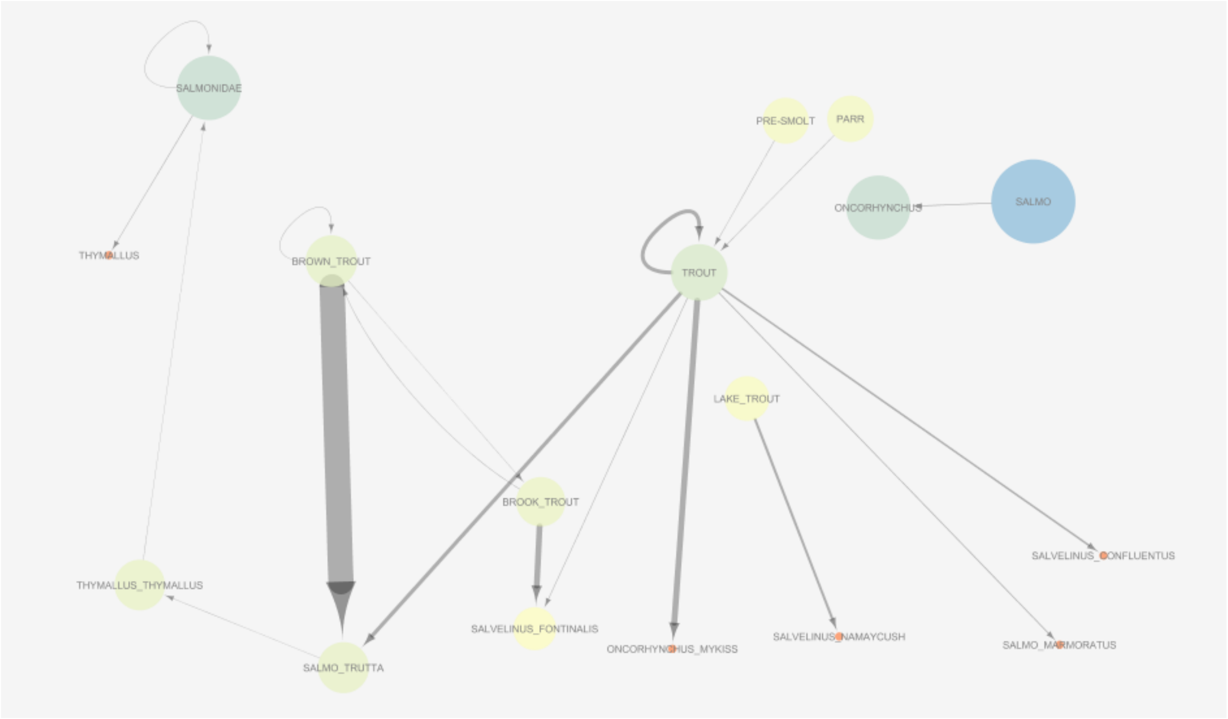


Figure 9.93: JEFF corpus: relation network graph for Salmo trutta, not lower-case, 2 or more hits, salience 9.5. In this graph the link between brook trout and salvelinus fontinalis is identified.

9.4 Chapter 7

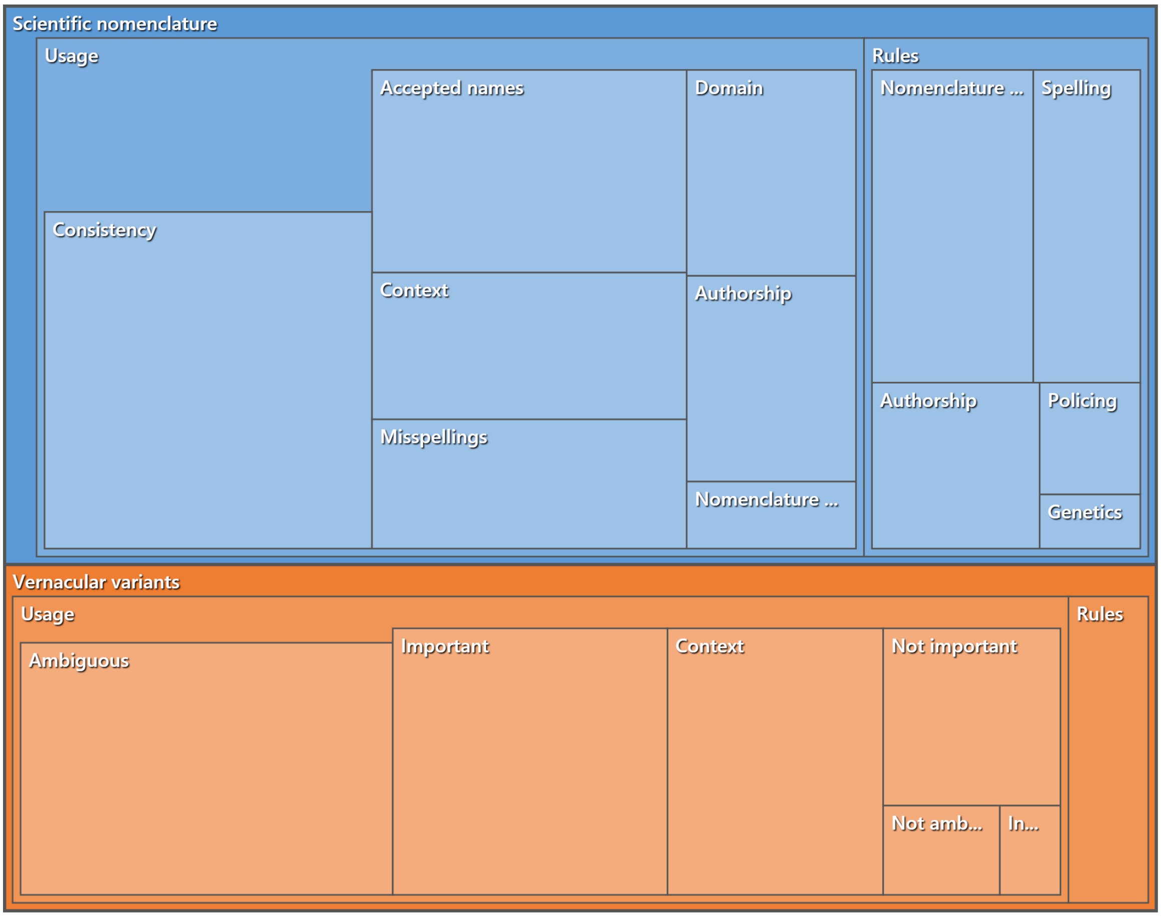


Figure 9.94: Comparison between topics and discussion points between Scientific Nomenclature and Vernacular Variants. Here the slightly higher emphasis on scientific nomenclature is identified and the division of different discussion topics within each area.

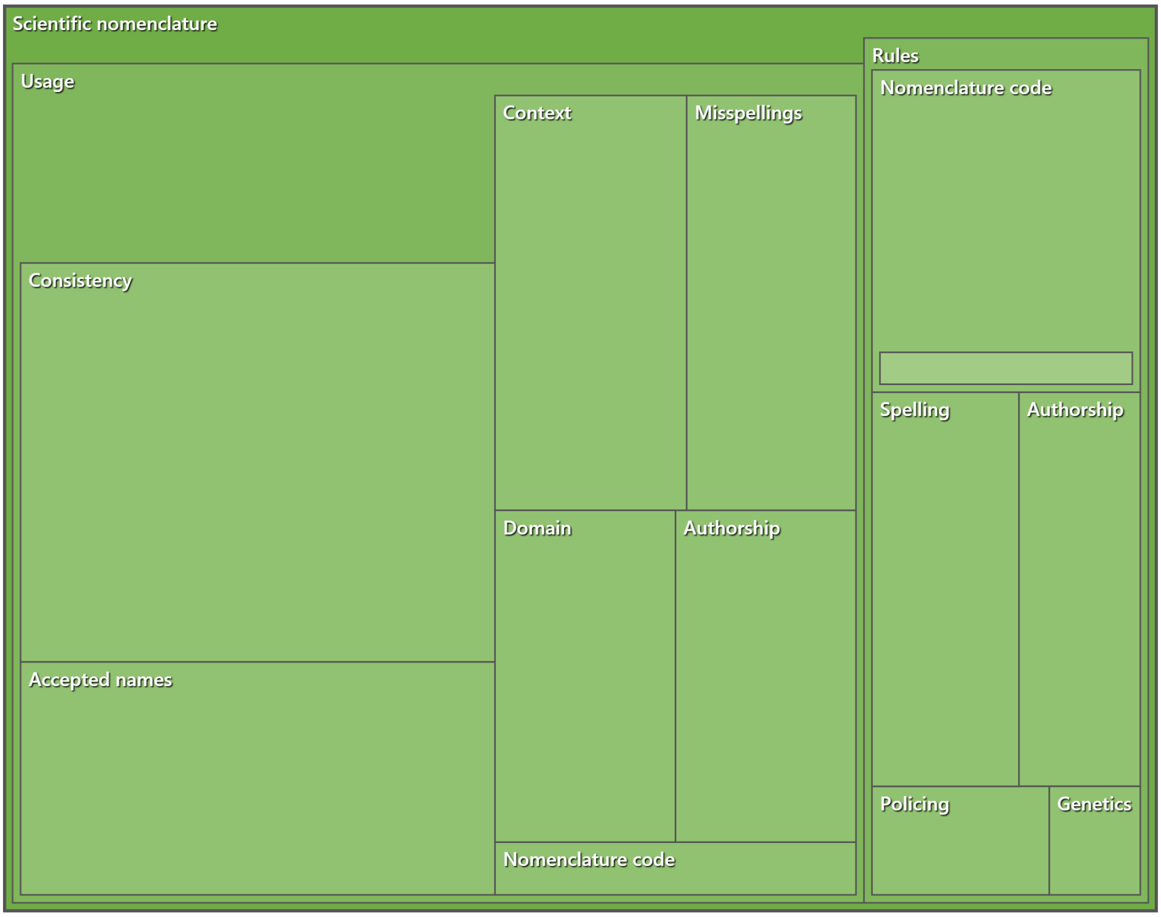


Figure 9.95: Scientific nomenclature: split between rules and usage. Here it is possible to see that usage dominated the conversation slightly but there was a wide variety of discussion on each part.

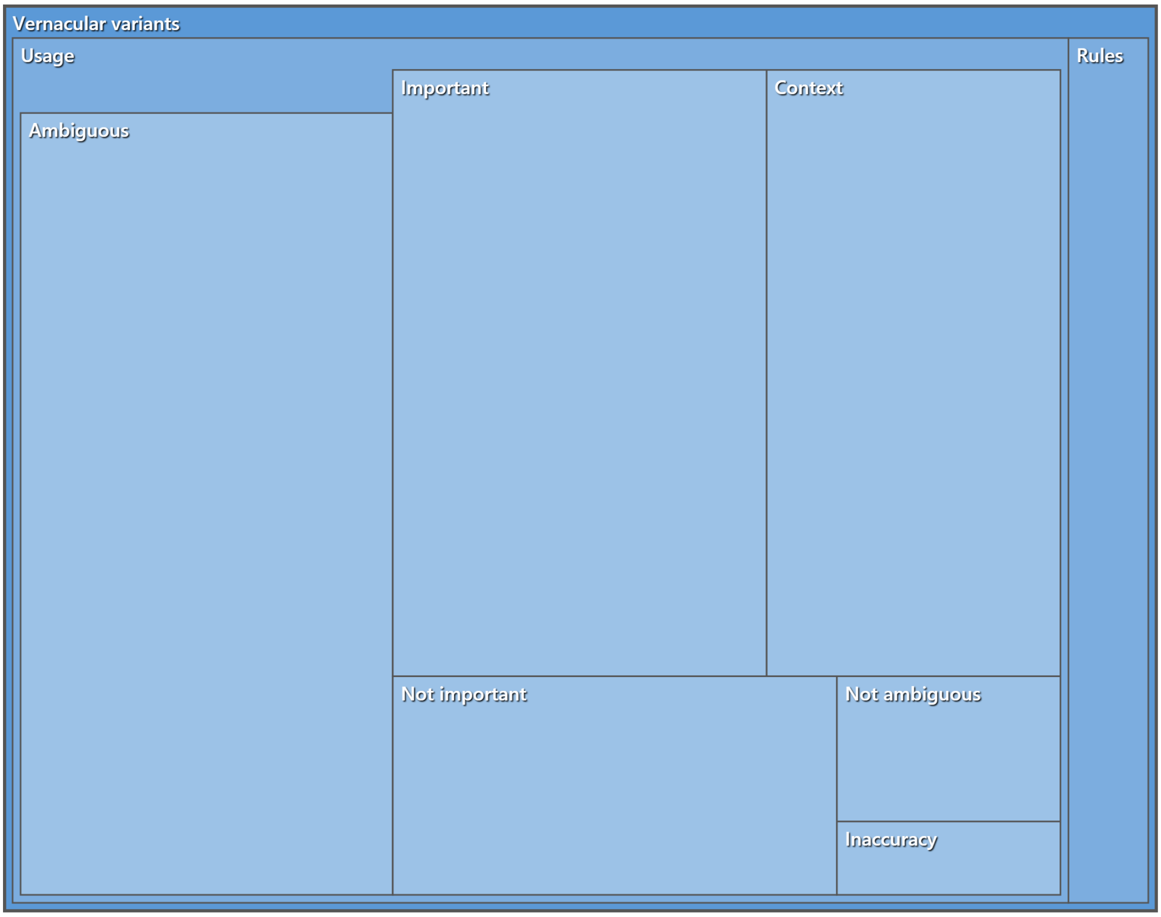
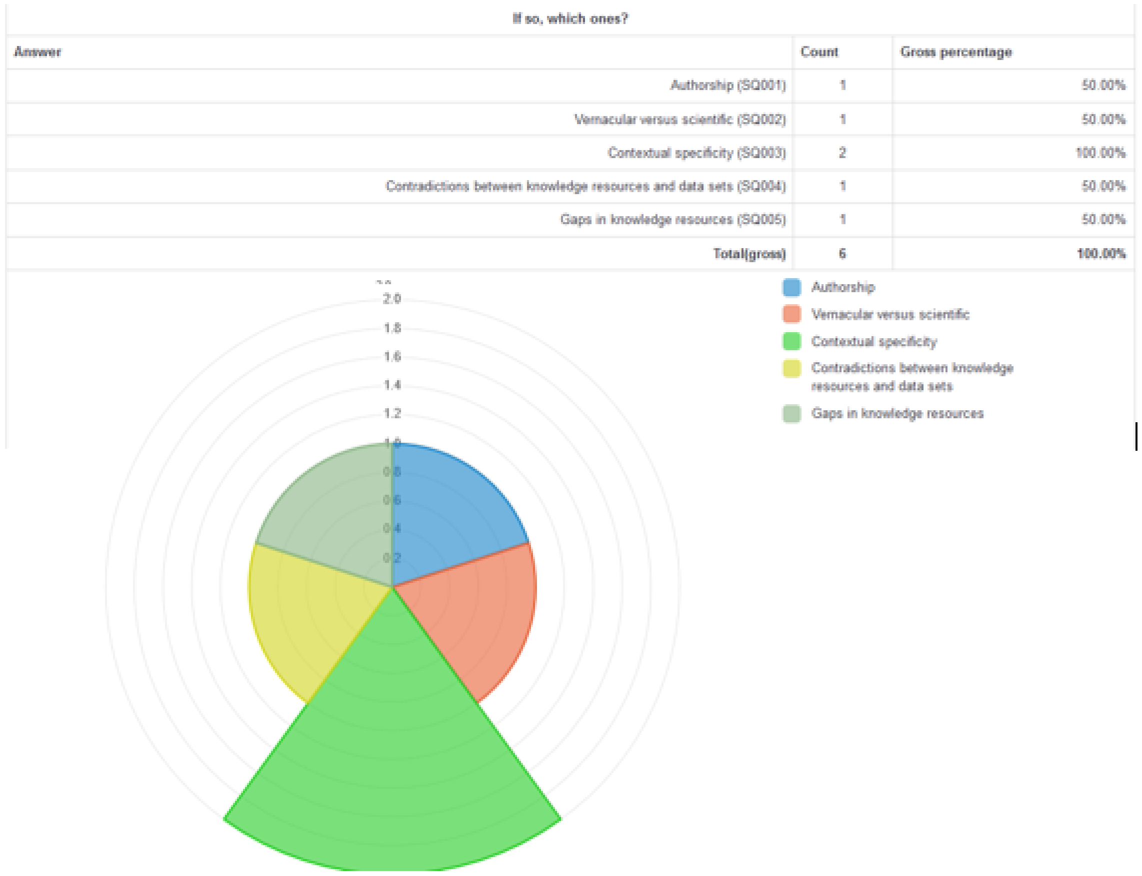


Figure 9.96: Vernacular variants: split between rules and usage. Here it is possible to see that usage was by far more prominent in the conversation than rules.



If so, which ones?

Answer	Count	Gross percentage
Authorship (SQ001)	1	50.00%
Vernacular versus scientific (SQ002)	1	50.00%
Contextual specificity (SQ003)	2	100.00%
Contradictions between knowledge resources and data sets (SQ004)	1	50.00%
Gaps in knowledge resources (SQ005)	1	50.00%
Total(gross)	**6**	**100.00%**

2.0
1.8
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
Authorship
Vernacular versus scientific
Contextual specificity
Contradictions between knowledge resources and data sets
Gaps in knowledge resources

Figure 9.97: Evaluation responses: for which ambiguities do you think my characterisations provide a practical approach?

Appendix A

Corpus data and annotation

This appendix contains everything relating to the corpus test data and related annotation. Any files too large or files that are not suitable for placing within the thesis document itself can be accessed through this link (https://github.com/Sandra-Young-Brighton/Appendices_thesis.git). The reference here provides the file name in the domain folder. Any problems with the git access should be directed to the author of this thesis at s.h.young@brighton.ac.uk.

A.1 Zenodo corpus files

For copyright reasons, these files cannot be produced here. They have been used in accordance with the access the researcher had available according to the University of Brighton subscription. A file with the list of texts can be found in the folder: /Phase 0/

A.2 JEFF corpus files

For copyright reasons, these files cannot be produced here. They have been used in accordance with the access the researcher had available according to the University of Brighton subscription with Wiley. Details of this can be found at [220].

A.3 WEB corpus files

As this corpus was scraped directly from the web from available material, it is available for exploration here. It was used in Phase 2 and also in Phase 3 in the various forms. Files are included in /Phase 2/Corpora/

- Original text file - Vertical file untagged - Vertical file tagged

A.4 Seed words for WEB corpus

Seed words were chosen for the WEB corpus by doing a keyword analysis on the JEFF corpus. This was relevant and necessary because of the wish to have two corpora of comparable subject matters.

List of seed words used to create WEB corpus			
Salmo	cyprinid	freshwater	pike
charr	Jonsson	Oncorhynchus	larval
trout	alpinus	anadramous	salar
otolith	sculpin	whitefish	parr
salmonid	Anguilla	invertebrate	benthic
eel	fish	zooplankton	smolt
predation	fluvialitis	salmon	goby
trutta	roach	lamprey	stickleback
trophic	Perca	spawn	perch
chub	forage	larva	prey
Salvelinus	Coregonus	grayling	

A.5 Name lists for annotations of JEFF and WEB corpora in Phases 1 and 2

Full name lists for the first part of the research are provided in the both in text and JSON format, according to the file names given below. The names lists used for each step and phase of the research are set out here and included in the respective folders according to the phase number:

- Phase 0
  1. Zenodo name list
- Phase 1
  1. JEFF GNRD name list (complete) - tagging
  2. JEFF GNRD name list (genera subset) - Word Sketch call
  3. JEFF GNRD name list (unified) - tagging and WS call

- Phase 2

  1. JEFF GNRD name list (complete) - tagging and Word Sketch call
  2. JEFF GNRD name list (subset) - Word Sketch call for JEFF and WEB corpora in main part of the phase
  3. WEB GNRD name list (complete) - tagging and Word Sketch call

A.6 Names lists for annotation of corpora in the Nomenclature Profile Studies

The annotation for the nomenclature study profiles were based on the variants provided by three different knowledge resources, the VTO, the CoL and the ITIS. The tagging was performed according to the amalgamation of these lists (removing any duplicates), and including the tagging of all taxonomic entity mentions within that family (in the case of Oncorhynchus mykiss, the Salmonidae family) in the test corpora. [Make sure this is in the methods and then put the lists in the folder I think - ask Roger]

A.6.1 Oncorhynchus mykiss

Combination of VTO, ITIS and CoL for Oncorhynchus mykiss (duplicates removed)

Salmonidae family according to VTO

Table A.1: Scientific variants across all three resources (SCI)

Scientific variants
Onchorynchus mykiss (Walbaum, 1792)
Oncorhynchus mykiss
Fario gairdneri (Richardson, 1836)
Onchorhynchus mykiss (Walbaum, 1792)
Onchorrhychus mykiss (Walbaum, 1792)
Oncorhynchus gairdnerii (Richardson, 1836)
Oncorhynchus mykiss aguabonita
Oncorhynchus mykiss gairdneri
Oncorhynchus mykiss gibbsi (Suckley, 1859)
Oncorhynchus mykiss irideus
Oncorhynchus mykiss nelsoni (Evermann, 1908)
Oncorhynchus myskis (Walbaum, 1792)
Parasalmo penshinensis (Pallas, 1814)
Salmo aquilarum
Salmo gairdneri
Salmo gairdneri beardsleei
Salmo gairdneri gairdneri (Richardson, 1836)
Salmo gairdneri gilberti
Salmo gairdneri irideus Gibbons, 1855
Salmo gairdneri kamloops (Jordan, 1892)
Salmo gairdneri Richardson, 1836
Salmo gairdneri shasta
Salmo gairdneri shasta Jordan, 1894
Salmo gairdneri stonei
Salmo gairdnerii
Salmo gairdnerii gairdnerii Richardson, 1836
Salmo gairdnerii irideus Gibbons, 1855
Salmo gairdnerii Richardson, 1836
Salmo gilberti Jordan, 1894
Salmo irideus Gibbons, 1855
Salmo irideux Gibbons, 1855
Salmo kamloops (Jordan, 1892)
Salmo newberrii
Salmo penshinensis Pallas, 1814
Salmo regalis
Salmo rivularis kamloops (Jordan, 1892)
Salmo smaragdus
Salmo whitei
Trutta iridea (Gibbons, 1855)
Oncorhynchus kamloops
Oncorhynchus kamloops Jordan, 1892
Parasalmo mykiss
Parasalmo mykiss (Walbaum, 1792)
Salmo gibbsii
Salmo gibbsii Suckley, 1859
Salmo iridea
Salmo iridea Gibbons, 1855
Salmo irideus argentatus
Salmo irideus argentatus Bajkov, 1927
Salmo kamloops whitehousei
Salmo kamloops whitehousei Dymond, 1931
Salmo masoni
Salmo masoni Suckley, 1860
Salmo mykiss
Salmo mykiss Walbaum, 1792
Salmo nelsoni
Salmo nelsoni Evermann, 1908
Salmo purpuratus
Salmo purpuratus Pallas, 1814
Salmo rivularis
Salmo rivularis Ayres, 1855
Salmo truncatus
Salmo truncatus Suckley, 1859

Table A.2: Vernacular variants across all three resources (COM)

Common variants
Baja California rainbow trout
Brown trout
Coast angel trout
Coast rainbow trout
Coast range trout
Kamchatka salmon
Kamchatka steelhead
Kamchatka trout
Kamloops trout
Rainbow trout
Salmon trout
Silver trout
Steelhead trout
Summer salmon
rainbow trout
redband trout

Table A.3: Salmonidae family: VTO (page 1)

VTO Salmonidae family: tagging			
Prosopium spilonotus	Coregonus fontanae	Salvelinus kronocius	Coregonus nipigon
Prosopium gemmifer	Coregonus bavaricus	Salvelinus drjagini	Coregonus reighardi
Prosopium coulterii	Coregonus lucinensis	Salvelinus taranetzi	Coregonus vandesius
Prosopium williamsoni	Coregonus confusus	Salvelinus schmidti	Coregonus zuerichensis
Prosopium abyssicola	Coregonus ussuriensis	Salvelinus namaycush	Coregonus austriaca
Prosopium cylindraceum	Coregonus fatioi	Salvelinus levanidovi	Coregonus nelsonii
Coregonus nigripinnis	Coregonus duplex	Salvelinus neiva	Brachymystax tumensis
Coregonus tugun	Coregonus pidschian	Salvelinus curilus	Thymallus baicalensis
Coregonus lavaretus	Coregonus heglingus	Salvelinus leucomaenis	Thymallus pallasii
Coregonus migratorius	Coregonus muksun	Salvelinus confluentus	Thymallus flacomaculatus
Coregonus peled	Coregonus pravdinellus	Salvelinus alpinus	Thymallus burejensis
Coregonus chadary	Coregonus oxyrinchus	Salvelinus malma	Thymallus baicalolenensis
Coregonus autumnalis	Coregonus baerii	Salvelinus fontinalis	Thymallus mertensii
Coregonus zenithicus	Coregonus nasus	Salvelinus kuznetzovi	Stenodus nelma
Coregonus artedi	Coregonus baunti	Salvelinus albus	Hucho ishikawae
Coregonus albula	Coregonus maraena	Oncorhynchus gilae	Hucho perryi
Coregonus clupeaformis	Stenodus leucichthys	Oncorhynchus clarkii	Salmo zrmanjaensis
Coregonus hoyi	Hucho hucho	Oncorhynchus tshawytscha	Salmo macedonicus
Coregonus kiyi	Hucho bleekeri	Oncorhynchus chrysogaster	Salmo caspius
Coregonus zugensis	Hucho taimen	Oncorhynchus keta	Salmo pallaryi
Coregonus widegreni	Parahucho perryi	Oncorhynchus kisutch	Salmo rhodanensis
Coregonus wartmanni	Salmo labrax	Oncorhynchus gorbuscha	Salmo sp. (Fink and Fink 1981)
Coregonus suidteri	Salmo carpio	Oncorhynchus mykiss	Salmo taleri
Coregonus palaea	Salmo trutta	Oncorhynchus nerka	Salmo aphelios
Coregonus renke	Salmo letnica	Oncorhynchus masou	Salmo ferox
Coregonus macrophthalmus	Salmo marmoratus	Salvethymus svetovidovi	Salmo coruhensis
Coregonus huntsmani	Salmo fibreni	Brachymystax savinovi	Salmo stomachicus
Coregonus nobilis	Salmo obtusirostris	Brachymystax lenok	Salmo schiefermuelleri
Coregonus laurettae	Salmo salar	Thymallus brevipinnis	Salmo ezenami
Coregonus albellus	Salmo ischchan	Thymallus arcticus	Salmo montenigrinus
Coregonus arenicolus	Salmo platycephalus	Thymallus nigrescens	Salmo balcanicus
Coregonus sardinella	Salmo ohridanus	Thymallus svetovidovi	Salmo rizeensis
Coregonus alpinus	Salvelinus elgyticus	Thymallus tugarinae	Salmo farioides
Coregonus candidus	Salvelinus boganidae	Coregonus bezola	Salmo lumi

Table A.4: Salmonidae family: VTO (page 2)

Salmo akairos	Coregonus pollan	Salvelinus gracillimus	Salvelinus evasus
Salmo dentex	Coregonus danneri	Salvelinus anaktuvukensis	Salvelinus youngeri
Salmo peristericus	Coregonus trybomi	Salvelinus obtusus	Salvelinus neocomensis
Salmo nigripinnis	Coregonus holsata	Salvelinus jacuticus	Salvelinus vasiljevae
Salmo pelagonicus	Coregonus stigmaticus	Salvelinus struanensis	Salvelinus tolmachoffi
Salmo visovacensis	Coregonus clupeoides	Salvelinus trevelyani	Salvelinus profundus
Salmo cettii	Coregonus atterensis	Salvelinus lonsdalii	Salvelinus struanensis
Salmo ciscaucasicus	Coregonus fera	Salvelinus thingvallensis	Salvelinus trevelyani
Salmo abanticus	Coregonus hoferi	Salvelinus lepechini	Salvelinus lonsdalii
Coregonus alpenae	Oncorhynchus iwame	Salvelinus inframundus	Salvelinus thingvallensis
Coregonus sp. (Fink and Fink 1981)	Oncorhynchus aguabonita	Salvelinus czerskii	Salvelinus lepechini
Coregonus subautumnalis	Oncorhynchus penshinensis	Salvelinus colii	Salvelinus inframundus
Coregonus kiletz	Salvelinus scharffi	Salvelinus murta	Salvelinus czerskii
Coregonus pennantii	Salvelinus umbla	Salvelinus agassizii	Salvelinus colii
Coregonus megalops	Salvelinus willoughbii	Salvelinus gracillimus	Salvelinus gritzenkoi
Coregonus anaulorum	Salvelinus fimbriatus	Salvelinus anaktuvukensis	Salvelinus taimyricus
Coregonus hiemalis	Thymallus brevirostris	Salvelinus obtusus	Salvelinus perisii
Coregonus restrictus	Thymallus thymallus	Salvelinus jacuticus	Salvelinus salvelinoinsularis
Coregonus gutturosus	Thymallus grubii	Salvelinus faroensis	Salvelinus aureolus
Coregonus pallasii	Salvelinus killinensis	Salvelinus maxillaris	Salvelinus andriashevi
Coregonus nilssoni	Salvelinus grayi	Coregonus ladogae	Coregonus vessicus
Coregonus johannae	Salvelinus murta	Salvelinus mallochi	Coregonus lutokka
Coregonus balticus	Salvelinus agassizii	Salvelinus krogiusae	Coregonus maxillaris

Table A.5: VTO hierarchy for tagging and Word Sketch API call

Hierarchy
Oncorhynchus
Salmonidae
Salmoniformes
Protacanthopterygii
Euteleostei

A.6.2 Sander lucioperca

Combination of VTO, ITIS and CoL for Sander lucioperca (duplicates removed)

Table A.6: Scientific variants across all three resources

Scientific variants
Sander lucioperca
Sander lucioperca (Linnaeus, 1758)
Centropomus sandat
Centropomus sandat Lacepède, 1802
Lucioperca linnei
Lucioperca linnei Malm, 1877
Lucioperca lucioperca
Lucioperca lucioperca (Linnaeus, 1758)
Lucioperca sandra
Lucioperca sandra Cuvier, 1828
Perca lucioperca
Perca lucioperca Linnaeus, 1758
Sander lucioperca (Linnaeus, 1758)
Stizostedion lucioperca
Stizostedion lucioperca (Linnaeus, 1758)
Stizostedion luciperca (Linnaeus, 1758)
Stizostedium lucioperca (Linnaeus, 1758)

Table A.7: Common variants

Common variants
pike-perch
pikeperch
zander

Table A.8: Percidae family according to VTO, p.1

VTO: Percidae: tagging
Percarina demidoffii
Percarina maeotica
Etheostoma parvipinne
Etheostoma phytophilum
Etheostoma saludae
Etheostoma fusiforme
Etheostoma zonifer
Etheostoma serrifer
Etheostoma collis
Etheostoma gracile
Etheostoma kennicotti
Etheostoma percnurum
Etheostoma pseudovulatum
Etheostoma olivaceum
Etheostoma oophylax
Etheostoma striatulum
Etheostoma virgatum
Etheostoma smithi
Etheostoma squamiceps
Etheostoma corona
Etheostoma crossopterum
Etheostoma barbouri
Etheostoma chienense
Etheostoma nigripinne
Etheostoma flabellare
Etheostoma obeyense
Etheostoma forbesi
Etheostoma neopterum
Etheostoma tuscumbia
Etheostoma davisoni
Etheostoma chlorosomum
Etheostoma fonticola
Etheostoma proeliare
Etheostoma microperca
Etheostoma inscriptum
Etheostoma punctulatum
Etheostoma pallididorsum
Etheostoma cragini
Etheostoma boschungi
Etheostoma trisella
Etheostoma vitreum
Etheostoma sagitta
Etheostoma nianguae
Etheostoma histrio
Etheostoma blennioides
Etheostoma thalassinum
Etheostoma swannanoa
Etheostoma moorei
Etheostoma chuckwachatte
Etheostoma bellum
Etheostoma douglasi
Etheostoma camurum
Etheostoma acuticeps
Etheostoma chlorobranchium
Etheostoma jordani
Etheostoma juliae
Percina uranidea
Percina antesella
Percina tanasi
Percina vigil
Percina shumardi
Percina aurantiaca
Percina macrolepida
Percina caprodes
Percina palmaris
Percina evides
Percina brevicauda
Percina copelandi
Percina aurora
Percina nasuta
Percina phoxocephala
Percina oxyrhynchus
Percina squamata
Percina jenkinsi
Percina cymatotaenia
Percina stictogaster
Percina kathae
Percina apristis
Percina nigrofasciata
Percina aurolineata
Percina crypta
Percina lenticula
Percina sciera
Percina rex

Table A.9: Percidae family according to VTO, p.2

VTO: Percidae: tagging

Etheostoma sitikuense
Etheostoma derivativum
Etheostoma basilare
Etheostoma marmorpinnum
Etheostoma brevispinum
Etheostoma lemniscatum
Etheostoma akatulo
Etheostoma stigmaeum
Etheostoma jessiae
Etheostoma okaloosae
Etheostoma mariae
Etheostoma fricksium
Etheostoma nigrum
Etheostoma susanae
Etheostoma olmstedi
Etheostoma perlongum
Etheostoma longimanum
Etheostoma podostemone
Etheostoma blennius
Etheostoma cinereum
Etheostoma edwini
Etheostoma caeruleum
Etheostoma asprigene
Etheostoma burri
Etheostoma collettei
Etheostoma luteovinctum
Etheostoma bison
Etheostoma lepidum
Etheostoma hopkinsi
Etheostoma australe
Etheostoma artesiae
Etheostoma exile
Etheostoma lawrencei
Etheostoma kantuckeense
Etheostoma lugoi
Etheostoma radiosum
Etheostoma nuchale
Etheostoma spectabile
Etheostoma pottsii
Etheostoma fragi
Etheostoma ditrema
Etheostoma whipplei
Etheostoma grahami
Etheostoma segrex
Etheostoma swaini
Etheostoma tecumsehi
Etheostoma uniporum
Etheostoma rupestre

Etheostoma tippecanoe
Etheostoma etowahae
Etheostoma maculatum
Etheostoma microlepidum
Etheostoma vulneratum
Etheostoma rufilineatum
Etheostoma sanguifluum
Etheostoma rubrum
Etheostoma aquali
Etheostoma wapiti
Etheostoma denoncourti
Etheostoma variatum
Etheostoma euzonum
Etheostoma tetrazonum
Etheostoma kanawhae
Etheostoma osburni
Etheostoma erythrozonum
Etheostoma zonale
Etheostoma lynceum
Etheostoma duryi
Etheostoma coosae
Etheostoma flavum
Etheostoma etnieri
Etheostoma brevirostrum
Etheostoma colorosum
Etheostoma chermocki
Etheostoma ramseyi
Etheostoma rafinesquei
Etheostoma cervus
Etheostoma raneyi
Etheostoma lachneri
Etheostoma pyrrhogaster
Etheostoma baileyi
Etheostoma barrenense
Etheostoma bellator
Etheostoma occidentale
Etheostoma planasaxatile
Etheostoma orientale
Etheostoma tennesseense
Etheostoma zonistium
Etheostoma simoterum
Etheostoma scotti
Etheostoma tallapoosae
Percina burtoni
Percina fulvitaenia
Percina carbonaria
Percina bimaculata
Percina austroperca

Percina maculata
Percina crassa
Percina macrocephala
Percina peltata
Percina williamsi
Percina gymnocephala
Percina pantherina
Percina roanoka
Percina nevisense
Percina notogramma
Percina kusha
Percina smithvanizi
Percina sipsi
Percina suttkusi
Ammocrypta bifascia
Ammocrypta meridiana
Ammocrypta clara
Ammocrypta pellucida
Ammocrypta vivax
Ammocrypta beanii
Sander lucioperca
Sander canadensis
Sander vitreus
Sander volgensis
Sander marinus
Zingel zingel
Zingel asper
Zingel streber
Romanichthys valsanicola
Crystallaria asprella
Crystallaria cincotta
Gymnocephalus cernuus
Gymnocephalus baloni
Gymnocephalus schraetser
Gymnocephalus acerina
Perca schrenkii
Perca flavescens
Perca fluviatilis
Zingel balcanicus
Gymnocephalus acerinus
Etheostoma gutselli
Etheostoma sequatchiense
Etheostoma atripinne
Etheostoma sellare
Perca beaumonti
Perca lepidota
Perca angusta

Table A.10: VTO hierarchy for tagging and Word Sketch API call

Hierarchy
Sander
Percidae
Perciformes
Percomorpha
Acanthopterygii

A.6.3 Salmo trutta

Combination of VTO, ITIS and CoL for Salmo trutta (duplicates removed)

Table A.11: Scientific variants of all three resources (duplicates removed)

Scientific variants
Salmo trutta
Salmo trutta Linnaeus, 1758
Salmo trutta Linnaeus, 1758
Salmo fario major Walecki, 1863
Salmo orientalis McClelland, 1842
Salmo stroemii Gmelin, 1789
Trutta fluviatilis Duhamel, 1771
Trutta salmanata Strøm, 1784
Trutta salmonata Rutty, 1772
Salmo trutta ciscaucasicus (non Dorofeeva, 1967)
Salmo trutta ezenami (non Berg, 1948) (misapplied name)
Fario argenteus
Fario argenteus Valenciennes, 1848
Fario lacustris (Linnaeus, 1758)
Fario trutta (Linnaeus, 1758)
Salar ausonii
Salar ausonii parcepunctata Heckel & Kner, 1858
Salar ausonii semipunctata Heckel & Kner, 1858
Salar ausonii Valenciennes, 1848
Salar bailloni
Salar bailloni Valenciennes, 1848
Salar gaimardi
Salar gaimardi Valenciennes, 1848
Salar macrostigma
Salar spectabilis
Salar spectabilis Valenciennes, 1848
Salmo albus
Salmo albus Bonnaterre, 1788
Salmo albus Walbaum, 1792
Salmo brachypoma
Salmo cumberland Lacepède, 1803
Salmo eriox
Salmo eriox Linnaeus, 1758
Salmo estuarius
Salmo estuarius Knox, 1855
Salmo fario
Salmo fario forestensis
Salmo fario forestensis Bloch & Schneider, 1801
Salmo fario Linnaeus, 1758
Salmo fario loensis
Salmo faris forestensis Bloch & Schneider, 1801
Salmo gadoides
Salmo gadoides Lacepède, 1803
Salmo gallivensis
Salmo gallivensis Günther, 1866
Salmo illanca
Salmo illanca Wartmann, 1783
Salmo islayensis
Salmo islayensis Thomson, 1873
Salmo lacustris
Salmo lacustris Linnaeus, 1758
Salmo lacustris rhenana
Salmo lacustris rhenana Fatio, 1890
Salmo lacustris romanovi
Salmo lacustris romanovi Kawraisky, 1896
Salmo lacustris septentrionalis
Salmo lacustris septentrionalis Fatio, 1890
Salmo lemanus
Salmo lemanus Cuvier, 1829
Salmo levenensis

Table A.12: Scientific variants of all three resources (duplicates removed)

Scientific variants (p.2)	
Salmo brachypoma Günther, 1866	Salmo levenensis Yarrell, 1839
Salmo caecifer	Salmo microps
Salmo caecifer Parnell, 1838	Salmo mistops
Salmo cambricus	Salmo mistops Günther, 1866
Salmo cambricus Donovan, 1806	Salmo montana Walker, 1812
Salmo caspius	Salmo orcadensis
Salmo cornubiensis	Salmo orcadensis Günther, 1866
Salmo cornubiensis Walbaum, 1792	Salmo oxianus
Salmo cumberland	Salmo oxianus Kessler, 1874
Salmo phinoc	Salmo trutta macrostigma
Salmo phinoc Shaw, 1804	Salmo trutta oxianus
Salmo polyosteus	Salmo trutta oxianus Kessler, 1874
Salmo polyosteus Günther, 1866	Salmo trutta trutta
Salmo rappii	Salmo trutta trutta Linnaeus, 1758
Salmo rappii Günther, 1866	Salmo truttula Nilsson, 1832
Salmo saxatilis	Salmo vario
Salmo saxatilis Schrank, 1798	Salmo venernensis
Salmo spurius Pallas, 1814	Salmo venernensis Günther, 1866
Salmo sylvaticus	Trutta fario (Linnaeus, 1758)
Salmo sylvaticus Gmelin, 1789	Trutta fario macroptera
Salmo taurinus Walker, 1812	Trutta lacustris (Linneaus, 1758)
Salmo trutta aralensis	Trutta marina
Salmo trutta aralensis Berg, 1908	Trutta marina Duhamel, 1771
Salmo trutta caspius	Trutta marina Moreau, 1881
Salmo trutta ciscaucasicus	Trutta trutta (Linnaeus, 1758)
Salmo trutta fario	Trutta variabilis
Salmo trutta fario Linnaeus, 1758	Trutta variabilis Lunel, 1874
Salmo trutta lacustris Linnaeus, 1758	Salmo fario loensis Walbaum, 1792

Table A.13: Vernacular variants

Vernacular variants	
Amu-Darya trout	Lake trout
Aral salmon	Loch leven trout
Aral Sea Trout	orange fin
Aral trout	Orkney sea trout
Blacktail	peal
black trout	river trout
brook trout	Salmón
brown trout	salmon trout
Brownie	Sea trout
Finnock	sea-trout
Galway sea trout	sewin
gillaroo	trout
herling	whiting
hirling	whitling

Salmonidae family according to VTO

Table A.14: Salmonidae family: VTO (page 1)

VTO Salmonidae family: tagging			
Prosopium spilonotus	Coregonus fontanae	Salvelinus kronocius	Coregonus nipigon
Prosopium gemmifer	Coregonus bavaricus	Salvelinus drjagini	Coregonus reighardi
Prosopium coulterii	Coregonus lucinensis	Salvelinus taranetzi	Coregonus vandesius
Prosopium williamsoni	Coregonus confusus	Salvelinus schmidti	Coregonus zuerichensis
Prosopium abyssicola	Coregonus ussuriensis	Salvelinus namaycush	Coregonus austriaca
Prosopium cylindraceum	Coregonus fatioi	Salvelinus levanidovi	Coregonus nelsonii
Coregonus nigripinnis	Coregonus duplex	Salvelinus neiva	Brachymystax tumensis
Coregonus tugun	Coregonus pidschian	Salvelinus curilus	Thymallus baicalensis
Coregonus lavaretus	Coregonus heglingus	Salvelinus leucomaenis	Thymallus pallasii
Coregonus migratorius	Coregonus muksun	Salvelinus confluentus	Thymallus flacomaculatus
Coregonus peled	Coregonus pravdinellus	Salvelinus alpinus	Thymallus burejensis
Coregonus chadary	Coregonus oxyrinchus	Salvelinus malma	Thymallus baicalolenensis
Coregonus autumnalis	Coregonus baerii	Salvelinus fontinalis	Thymallus mertensii
Coregonus zenithicus	Coregonus nasus	Salvelinus kuznetzovi	Stenodus nelma
Coregonus artedi	Coregonus baunti	Salvelinus albus	Hucho ishikawae
Coregonus albula	Coregonus maraena	Oncorhynchus gilae	Hucho perryi
Coregonus clupeaformis	Stenodus leucichthys	Oncorhynchus clarkii	Salmo zrmanjaensis
Coregonus hoyi	Hucho hucho	Oncorhynchus tshawytscha	Salmo macedonicus
Coregonus kiyi	Hucho bleekeri	Oncorhynchus chrysogaster	Salmo caspius
Coregonus zugensis	Hucho taimen	Oncorhynchus keta	Salmo pallaryi
Coregonus widegreni	Parahucho perryi	Oncorhynchus kisutch	Salmo rhodanensis
Coregonus wartmanni	Salmo labrax	Oncorhynchus gorbuscha	Salmo sp. (Fink and Fink 1981)
Coregonus suidteri	Salmo carpio	Oncorhynchus mykiss	Salmo taleri
Coregonus palaea	Salmo trutta	Oncorhynchus nerka	Salmo aphelios
Coregonus renke	Salmo letnica	Oncorhynchus masou	Salmo ferox
Coregonus macrophthalmus	Salmo marmoratus	Salvethymus svetovidovi	Salmo coruhensis
Coregonus huntsmani	Salmo fibreni	Brachymystax savinovi	Salmo stomachicus
Coregonus nobilis	Salmo obtusirostris	Brachymystax lenok	Salmo schiefermuelleri
Coregonus laurettae	Salmo salar	Thymallus brevipinnis	Salmo ezenami
Coregonus albellus	Salmo ischchan	Thymallus arcticus	Salmo montenigrinus
Coregonus arenicolus	Salmo platycephalus	Thymallus nigrescens	Salmo balcanicus
Coregonus sardinella	Salmo ohridanus	Thymallus svetovidovi	Salmo rizeensis
Coregonus alpinus	Salvelinus elgyticus	Thymallus tugarinae	Salmo farioides
Coregonus candidus	Salvelinus boganidae	Coregonus bezola	Salmo lumi

Table A.15: Salmonidae family: VTO (page 2)

Salmo akairos	Coregonus pollan	Salvelinus gracillimus	Salvelinus evasus
Salmo dentex	Coregonus danneri	Salvelinus anaktuvukensis	Salvelinus youngeri
Salmo peristericus	Coregonus trybomi	Salvelinus obtusus	Salvelinus neocomensis
Salmo nigripinnis	Coregonus holsata	Salvelinus jacuticus	Salvelinus vasiljevae
Salmo pelagonicus	Coregonus stigmaticus	Salvelinus struanensis	Salvelinus tolmachoffi
Salmo visovacensis	Coregonus clupeoides	Salvelinus trevelyani	Salvelinus profundus
Salmo cettii	Coregonus atterensis	Salvelinus lonsdalii	Salvelinus struanensis
Salmo ciscaucasicus	Coregonus fera	Salvelinus thingvallensis	Salvelinus trevelyani
Salmo abanticus	Coregonus hoferi	Salvelinus lepechini	Salvelinus lonsdalii
Coregonus alpenae	Oncorhynchus iwame	Salvelinus inframundus	Salvelinus thingvallensis
Coregonus sp. (Fink and Fink 1981)	Oncorhynchus aguabonita	Salvelinus czerskii	Salvelinus lepechini
Coregonus subautumnalis	Oncorhynchus penshinensis	Salvelinus colii	Salvelinus inframundus
Coregonus kiletz	Salvelinus scharffi	Salvelinus murta	Salvelinus czerskii
Coregonus pennantii	Salvelinus umbla	Salvelinus agassizii	Salvelinus colii
Coregonus megalops	Salvelinus willoughbii	Salvelinus gracillimus	Salvelinus gritzenkoi
Coregonus anaulorum	Salvelinus fimbriatus	Salvelinus anaktuvukensis	Salvelinus taimyricus
Coregonus hiemalis	Thymallus brevirostris	Salvelinus obtusus	Salvelinus perisii
Coregonus restrictus	Thymallus thymallus	Salvelinus jacuticus	Salvelinus salvelinoinsularis
Coregonus gutturosus	Thymallus grubii	Salvelinus faroensis	Salvelinus aureolus
Coregonus pallasii	Salvelinus killinensis	Salvelinus maxillaris	Salvelinus andriashevi
Coregonus nilssoni	Salvelinus grayi	Coregonus ladogae	Coregonus vessicus
Coregonus johannae	Salvelinus murta	Salvelinus mallochi	Coregonus lutokka
Coregonus balticus	Salvelinus agassizii	Salvelinus krogiusae	Coregonus maxillaris

Table A.16: VTO hierarchy for tagging and Word Sketch API call

Hierarchy
Salmo
Salmonidae
Salmoniformes
Protacanthopterygii
Euteleostei

Appendix B

Sketch Engine files

B.1 Sketch Grammars

The Sketch Grammars are the rules used to identify different grammatical and semantic relations between words in Sketch Engine. To follow is the original, stock Sketch Grammar, followed by the Ecolexicon extension to this Sketch Grammar. After this are the Sketch Grammars used in the thesis: a Sketch Grammar based on the Ecolexicon Grammar for the lempos-retrieved Word Sketches, plus the Sketch Grammars that look to a fourth column for a tag.

Any files too large or files that are not suitable for placing within the thesis document itself can be accessed through this link (https://github.com/Sandra-Young-Brighton/Appendices_thesis.git). The reference here provides the file name in the domain folder. Any problems with the git access should be directed to the author of this thesis at s.h.young@brighton.ac.uk.

Original Sketch Grammar (stock):
Appendices/Sketch Grammars/Sketch Grammar for English.pdf

Ecolexicon Sketch Grammar:
Appendices/Sketch Grammars/Ecolexicon Semantic Sketch Grammar.pdf

Lempos Sketch Grammar:
Appendices/Sketch Grammars/SCICOMGEN_newbreakdown.pdf

4th Column Sketch Grammar:
Appendices/Sketch Grammars/SCI_NN_4th.pdf

4th Column differentiated Sketch Grammar:
Appendices/Sketch Grammars/SCICOMGEN_newbr.pdf

B.2 Lists of names to call Word Sketches

Phase 1: refer to folder Phase 1/Name list/ Phase 2: refer to folder Phase 2/Name lists/ Phase 3: specify and refer back to the Appendix A: collective plus the VTO hierarchy

B.3 Configuration file

```
DEFAULTLOCALE "en_US.UTF-8"
DOCSTRUCTURE "doc"
ENCODING "UTF-8"
FILESTRUCTURE "doc"
INFO ""
LANGUAGE "English"
LPOSLIST ",adjective,-j,adverb,-a,conjunction,-c,noun,-n,numeral,-
m,preposition,-i,pronoun,-d,verb,-v"
NAME "WEB_WEBGNRD"
PATH "/corpora/ca/user_data/sandrayoung/manatee/web_webgnrd"
REFCORPUS "ententen13_tt2_1"
TAGSETDOC "https://www.sketchengine.co.uk/english-treetagger-pipeline-2/"
VERTICAL "| ca_getvertical
'/corpora/ca/user_data/sandrayoung/registry/web_webgnrd' 'docx'"
WPOSLIST
",adjective,J.*,adverb,RB.?,conjunction,CC,determiner,DT,noun,N.*,noun
singular,NN,noun
plural,NNS,numeral,CD,particle,RP,preposition,IN,pronoun,PP.?,verb,V.*,fu
ll stop,SENT"
WSATTR "lempos"
ATTRIBUTE "word" {
    MAPTO "lempos"
}
ATTRIBUTE "tag" {
}
ATTRIBUTE "lempos" {
}
ATTRIBUTE "scientific_name" {
}
ATTRIBUTE "lemma" {
    ARG1 "2"
    DYNAMIC "striplastn"
    DYNLIB "internal"
    DYNTYPE "index"
    FROMATTR "lempos"
    FUNTYPE "i"
}
ATTRIBUTE "lempos_lc" {
    ARG1 "C"
    DYNAMIC "utf8lowercase"
    DYNLIB "internal"
    DYNTYPE "index"
    FROMATTR "lempos"
    FUNTYPE "s"
    LABEL "lempos (lowercase)"
    TRANSQUERY "yes"
}
ATTRIBUTE "lemma_lc" {
    ARG1 "C"
    DYNAMIC "utf8lowercase"
    DYNLIB "internal"
    DYNTYPE "index"
    FROMATTR "lemma"
    FUNTYPE "s"
    LABEL "lemma (lowercase)"
    TRANSQUERY "yes"
}
ATTRIBUTE "lc" {
    ARG1 "C"
    DYNAMIC "utf8lowercase"
    DYNLIB "internal"
    DYNTYPE "index"
    FROMATTR "word"
    FUNTYPE "s"
    LABEL "word (lowercase)"
    TRANSQUERY "yes"
}
STRUCTURE "s" {
}
STRUCTURE "g" {
    DISPLAYBEGIN "_EMPTY_"
    DISPLAYTAG "0"
}
STRUCTURE "doc" {
    ENCODING "UTF-8"
    ATTRIBUTE "url" {
        DYNTYPE "index"
        ENCODING "UTF-8"
        LABEL "URL"
        LOCALE "en_US.UTF-8"
        MULTISEP ","
        MULTIVALUE "n"
        TYPE "MD_MI"
    }
    ATTRIBUTE "parent_folder" {
        DYNTYPE "index"
        ENCODING "UTF-8"
        LABEL "Folder"
        LOCALE "en_US.UTF-8"
        MULTISEP ","
        MULTIVALUE "n"
        TYPE "MD_MI"
    }
    ATTRIBUTE "id" {
        DYNTYPE "index"
        ENCODING "UTF-8"
        LABEL "File ID"
        LOCALE "en_US.UTF-8"
        MULTISEP ","
        MULTIVALUE "n"
        TYPE "MD_MI"
    }
    ATTRIBUTE "filename" {
        DYNTYPE "index"
        ENCODING "UTF-8"
        LABEL "File name"
        LOCALE "en_US.UTF-8"
        MULTISEP ","
        MULTIVALUE "n"
        TYPE "MD_MI"
    }
}
STRUCTURE "p" {
    DEFAULTLOCALE "C"
    ENCODING "UTF-8"
    LANGUAGE ""
    NESTED ""
    ATTRIBUTE "class" {
        DYNTYPE "index"
        ENCODING "UTF-8"
        LOCALE "C"
        MULTISEP ","
        MULTIVALUE "n"
        TYPE "MD_MI"
    }
}
STRUCTURE "li" {
    DEFAULTLOCALE "C"
    ENCODING "UTF-8"
    LANGUAGE ""
    NESTED ""
    ATTRIBUTE "class" {
        DYNTYPE "index"
        ENCODING "UTF-8"
        LOCALE "C"
        MULTISEP ","
        MULTIVALUE "n"
        TYPE "MD_MI"
    }
}
STRUCTURE "a" {
    DEFAULTLOCALE "C"
    ENCODING "UTF-8"
    LANGUAGE ""
    NESTED ""
    ATTRIBUTE "href" {
        DYNTYPE "index"
        ENCODING "UTF-8"
        LOCALE "C"
        MULTISEP ","
        MULTIVALUE "n"
        TYPE "MD_MI"
    }
    ATTRIBUTE "class" {
        DYNTYPE "index"
        ENCODING "UTF-8"
        LOCALE "C"
        MULTISEP ","
        MULTIVALUE "n"
        TYPE "MD_MI"
    }
    ATTRIBUTE "tabindex" {
        DYNTYPE "index"
        ENCODING "UTF-8"
        LOCALE "C"
        MULTISEP ","
        MULTIVALUE "n"
        TYPE "MD_MI"
    }
    ATTRIBUTE "id" {
        DYNTYPE "index"
        ENCODING "UTF-8"
        LOCALE "C"
        MULTISEP ","
```

```
            MULTIVALUE "n"
            TYPE "MD_MI"
        }
        ATTRIBUTE "rel" {
            DYNTYPE "index"
            ENCODING "UTF-8"
            LOCALE "C"
            MULTISEP ","
            MULTIVALUE "n"
            TYPE "MD_MI"
        }
    }
    STRUCTURE "img" {
        DEFAULTLOCALE "C"
        ENCODING "UTF-8"
        LANGUAGE ""
        NESTED ""
        ATTRIBUTE "alt" {
            DYNTYPE "index"
            ENCODING "UTF-8"
            LOCALE "C"
            MULTISEP ","
            MULTIVALUE "n"
            TYPE "MD_MI"
        }
        ATTRIBUTE "src" {
            DYNTYPE "index"
            ENCODING "UTF-8"
            LOCALE "C"
            MULTISEP ","
            MULTIVALUE "n"
            TYPE "MD_MI"
        }
        ATTRIBUTE "title" {
            DYNTYPE "index"
            ENCODING "UTF-8"
            LOCALE "C"
            MULTISEP ","
            MULTIVALUE "n"
            TYPE "MD_MI"
        }
        ATTRIBUTE "rel" {
            DYNTYPE "index"
            ENCODING "UTF-8"
            LOCALE "C"
            MULTISEP ","
            MULTIVALUE "n"
            TYPE "MD_MI"
        }
        ATTRIBUTE "class" {
            DYNTYPE "index"
            ENCODING "UTF-8"
            LOCALE "C"
            MULTISEP ","
            MULTIVALUE "n"
            TYPE "MD_MI"
        }
        ATTRIBUTE "aria-hidden" {
```

```
        DYNTYPE "index"
        ENCODING "UTF-8"
        LOCALE "C"
        MULTISEP ","
        MULTIVALUE "n"
        TYPE "MD_MI"
    }
    ATTRIBUTE "id" {
        DYNTYPE "index"
        ENCODING "UTF-8"
        LOCALE "C"
        MULTISEP ","
        MULTIVALUE "n"
        TYPE "MD_MI"
    }
    ATTRIBUTE "height" {
        DYNTYPE "index"
        ENCODING "UTF-8"
        LOCALE "C"
        MULTISEP ","
        MULTIVALUE "n"
        TYPE "MD_MI"
    }
    ATTRIBUTE "style" {
        DYNTYPE "index"
        ENCODING "UTF-8"
        LOCALE "C"
        MULTISEP ","
        MULTIVALUE "n"
        TYPE "MD_MI"
    }
}
STRUCTURE "div" {
    DEFAULTLOCALE "C"
    ENCODING "UTF-8"
    LANGUAGE ""
    NESTED ""
    ATTRIBUTE "class" {
        DYNTYPE "index"
        ENCODING "UTF-8"
        LOCALE "C"
        MULTISEP ","
        MULTIVALUE "n"
        TYPE "MD_MI"
    }
    ATTRIBUTE "id" {
        DYNTYPE "index"
        ENCODING "UTF-8"
        LOCALE "C"
        MULTISEP ","
        MULTIVALUE "n"
        TYPE "MD_MI"
    }
    ATTRIBUTE "style" {
        DYNTYPE "index"
        ENCODING "UTF-8"
        LOCALE "C"
        MULTISEP ","
        MULTIVALUE "n"
        TYPE "MD_MI"
    }
    ATTRIBUTE "role" {
        DYNTYPE "index"
        ENCODING "UTF-8"
        LOCALE "C"
        MULTISEP ","
        MULTIVALUE "n"
        TYPE "MD_MI"
    }
    ATTRIBUTE "data-dropdown" {
        DYNTYPE "index"
        ENCODING "UTF-8"
        LOCALE "C"
        MULTISEP ","
        MULTIVALUE "n"
        TYPE "MD_MI"
    }
}
STRUCTURE "meta" {
    DEFAULTLOCALE "C"
    ENCODING "UTF-8"
    LANGUAGE ""
    NESTED ""
    ATTRIBUTE "content" {
        DYNTYPE "index"
        ENCODING "UTF-8"
        LOCALE "C"
        MULTISEP ","
        MULTIVALUE "n"
        TYPE "MD_MI"
    }
    ATTRIBUTE "name" {
        DYNTYPE "index"
        ENCODING "UTF-8"
        LOCALE "C"
        MULTISEP ","
        MULTIVALUE "n"
        TYPE "MD_MI"
    }
    ATTRIBUTE "http-equiv" {
        DYNTYPE "index"
        ENCODING "UTF-8"
        LOCALE "C"
        MULTISEP ","
        MULTIVALUE "n"
        TYPE "MD_MI"
    }
    ATTRIBUTE "charset" {
        DYNTYPE "index"
        ENCODING "UTF-8"
        LOCALE "C"
        MULTISEP ","
        MULTIVALUE "n"
        TYPE "MD_MI"
    }
}
```

```
STRUCTURE "script" {
    DEFAULTLOCALE "C"
    ENCODING "UTF-8"
    LANGUAGE ""
    NESTED ""
    ATTRIBUTE "type" {
        DYNTYPE "index"
        ENCODING "UTF-8"
        LOCALE "C"
        MULTISEP ","
        MULTIVALUE "n"
        TYPE "MD_MI"
    }
    ATTRIBUTE "src" {
        DYNTYPE "index"
        ENCODING "UTF-8"
        LOCALE "C"
        MULTISEP ","
        MULTIVALUE "n"
        TYPE "MD_MI"
    }
}
STRUCTURE "span" {
    DEFAULTLOCALE "C"
    ENCODING "UTF-8"
    LANGUAGE ""
    NESTED ""
    ATTRIBUTE "class" {
        DYNTYPE "index"
        ENCODING "UTF-8"
        LOCALE "C"
        MULTISEP ","
        MULTIVALUE "n"
        TYPE "MD_MI"
    }
}
STRUCTURE "ul" {
    DEFAULTLOCALE "C"
    ENCODING "UTF-8"
    LANGUAGE ""
    NESTED ""
    ATTRIBUTE "class" {
        DYNTYPE "index"
        ENCODING "UTF-8"
        LOCALE "C"
        MULTISEP ","
        MULTIVALUE "n"
        TYPE "MD_MI"
    }
}
STRUCTURE "link" {
    DEFAULTLOCALE "C"
    ENCODING "UTF-8"
    LANGUAGE ""
    NESTED ""
    ATTRIBUTE "type" {
        DYNTYPE "index"
        ENCODING "UTF-8"
        LOCALE "C"
        MULTISEP ","
        MULTIVALUE "n"
        TYPE "MD_MI"
    }
    ATTRIBUTE "rel" {
        DYNTYPE "index"
        ENCODING "UTF-8"
        LOCALE "C"
        MULTISEP ","
        MULTIVALUE "n"
        TYPE "MD_MI"
    }
    ATTRIBUTE "href" {
        DYNTYPE "index"
        ENCODING "UTF-8"
        LOCALE "C"
        MULTISEP ","
        MULTIVALUE "n"
        TYPE "MD_MI"
    }
    ATTRIBUTE "media" {
        DYNTYPE "index"
        ENCODING "UTF-8"
        LOCALE "C"
        MULTISEP ","
        MULTIVALUE "n"
        TYPE "MD_MI"
    }
}
STRUCTURE "h3" {
    ENCODING "UTF-8"
}
STRUCTURE "i" {
    ENCODING "UTF-8"
}
STRUCTURE "button" {
    DEFAULTLOCALE "C"
    ENCODING "UTF-8"
    LANGUAGE ""
    NESTED ""
    ATTRIBUTE "class" {
        DYNTYPE "index"
        ENCODING "UTF-8"
        LOCALE "C"
        MULTISEP ","
        MULTIVALUE "n"
        TYPE "MD_MI"
    }
    ATTRIBUTE "type" {
        DYNTYPE "index"
        ENCODING "UTF-8"
        LOCALE "C"
        MULTISEP ","
        MULTIVALUE "n"
        TYPE "MD_MI"
    }
    ATTRIBUTE "name" {
        DYNTYPE "index"
        ENCODING "UTF-8"
        LOCALE "C"
        MULTISEP ","
        MULTIVALUE "n"
        TYPE "MD_MI"
    }
    ATTRIBUTE "id" {
        DYNTYPE "index"
        ENCODING "UTF-8"
        LOCALE "C"
        MULTISEP ","
        MULTIVALUE "n"
        TYPE "MD_MI"
    }
    ATTRIBUTE "data-toggle" {
        DYNTYPE "index"
        ENCODING "UTF-8"
        LOCALE "C"
        MULTISEP ","
        MULTIVALUE "n"
        TYPE "MD_MI"
    }
}
STRUCTURE "br" {
    ENCODING "UTF-8"
}
STRUCTURE "html" {
    DEFAULTLOCALE "C"
    ENCODING "UTF-8"
    LANGUAGE ""
    NESTED ""
    ATTRIBUTE "lang" {
        DYNTYPE "index"
        ENCODING "UTF-8"
        LOCALE "C"
        MULTISEP ","
        MULTIVALUE "n"
        TYPE "MD_MI"
    }
}
STRUCTURE "date" {
    ENCODING "UTF-8"
}
STRUCTURE "nav" {
    DEFAULTLOCALE "C"
    ENCODING "UTF-8"
    LANGUAGE ""
    NESTED ""
    ATTRIBUTE "class" {
        DYNTYPE "index"
        ENCODING "UTF-8"
        LOCALE "C"
        MULTISEP ","
        MULTIVALUE "n"
        TYPE "MD_MI"
    }
    ATTRIBUTE "data-dropdown-item" {
        DYNTYPE "index"
        ENCODING "UTF-8"
        LOCALE "C"
        MULTISEP ","
        MULTIVALUE "n"
        TYPE "MD_MI"
    }
}
STRUCTURE "body" {
    ENCODING "UTF-8"
}
STRUCTURE "title" {
    ENCODING "UTF-8"
}
STRUCTURE "header" {
    DEFAULTLOCALE "C"
    ENCODING "UTF-8"
    LANGUAGE ""
    NESTED ""
    ATTRIBUTE "data-mobilemenu-focussed" {
        DYNTYPE "index"
        ENCODING "UTF-8"
        LOCALE "C"
        MULTISEP ","
        MULTIVALUE "n"
        TYPE "MD_MI"
    }
    ATTRIBUTE "class" {
        DYNTYPE "index"
        ENCODING "UTF-8"
        LOCALE "C"
        MULTISEP ","
        MULTIVALUE "n"
        TYPE "MD_MI"
    }
}
STRUCTURE "head" {
    ENCODING "UTF-8"
}
STRUCTURE "h2" {
    DEFAULTLOCALE "C"
    ENCODING "UTF-8"
    LANGUAGE ""
    NESTED ""
    ATTRIBUTE "class" {
        DYNTYPE "index"
        ENCODING "UTF-8"
        LOCALE "C"
        MULTISEP ","
        MULTIVALUE "n"
        TYPE "MD_MI"
    }
}
STRUCTURE "form" {
    DEFAULTLOCALE "C"
    ENCODING "UTF-8"
    LANGUAGE ""
    NESTED ""
    ATTRIBUTE "method" {
        DYNTYPE "index"
        ENCODING "UTF-8"
        LOCALE "C"
        MULTISEP ","
        MULTIVALUE "n"
        TYPE "MD_MI"
    }
    ATTRIBUTE "id" {
        DYNTYPE "index"
        ENCODING "UTF-8"
        LOCALE "C"
        MULTISEP ","
        MULTIVALUE "n"
        TYPE "MD_MI"
    }
    ATTRIBUTE "action" {
        DYNTYPE "index"
        ENCODING "UTF-8"
        LOCALE "C"
        MULTISEP ","
        MULTIVALUE "n"
        TYPE "MD_MI"
    }
}
STRUCTURE "footer" {
    DEFAULTLOCALE "C"
    ENCODING "UTF-8"
    LANGUAGE ""
    NESTED ""
    ATTRIBUTE "id" {
        DYNTYPE "index"
        ENCODING "UTF-8"
        LOCALE "C"
        MULTISEP ","
        MULTIVALUE "n"
        TYPE "MD_MI"
    }
    ATTRIBUTE "class" {
        DYNTYPE "index"
        ENCODING "UTF-8"
        LOCALE "C"
        MULTISEP ","
        MULTIVALUE "n"
        TYPE "MD_MI"
    }
}
STRUCTURE "ol" {
    ENCODING "UTF-8"
}
STRUCTURE "em" {
    ENCODING "UTF-8"
}
STRUCTURE "sub" {
    DEFAULTLOCALE "C"
    ENCODING "UTF-8"
    LANGUAGE ""
    NESTED ""
    ATTRIBUTE "style" {
        DYNTYPE "index"
        ENCODING "UTF-8"
        LOCALE "C"
        MULTISEP ","
        MULTIVALUE "n"
        TYPE "MD_MI"
    }
}
STRUCTURE "I" {
    ENCODING "UTF-8"
}
STRUCTURE "B" {
    ENCODING "UTF-8"
}
STRUCTURE "top-down" {
    ENCODING "UTF-8"
}
STRUCTURE "docx" {
    DEFAULTLOCALE "C"
    ENCODING "UTF-8"
    LANGUAGE ""
    NESTED ""
    ATTRIBUTE "id" {
        DYNTYPE "index"
        ENCODING "UTF-8"
        LABEL "File ID"
        LOCALE "C"
        MULTISEP ","
        MULTIVALUE "n"
        TYPE "MD_MI"
        UNIQUE "1"
    }
    ATTRIBUTE "filename" {
        DYNTYPE "index"
        ENCODING "UTF-8"
        LABEL "File name"
        LOCALE "C"
        MULTISEP ","
        MULTIVALUE "n"
        TYPE "MD_MI"
    }
}
WSBASE      ""
WSTHES      ""
WSMINHITS   ""
WSDEF "/corpora/ca/user_data/sandrayoung/sg/SCI_newbr_4.m4"
TERMDEF     ""
TERMBASE    "none"
```

Appendix C

Data conversion

You can access the Cytoscape files in the relevant folders on the github repository for exploration should that be required (https://github.com/Sandra-Young-Brighton/Appendices_thesis.git). The Word Sketches for each part of the thesis are also included in the relevant phase folders within the repository. Any problems with the git access should be directed to the author of this thesis at s.h.young@brighton.ac.uk.

Phase 1 In the appendices folder: Phase 1/Cytoscape/JEFF_original.cys and JEFF_unified.cys

Phase 2 All files relating to the data conversion and the data used in Phase 2 can be found in folder Appendices/Phase 2/...

Phase 3 In the appendices folder: Appendices/Phase3/NPS/Cytoscape/...

Table C.1: VTO ranking numbers and equivalents

name	**TAXRANK**	**name**	**TAXRANK**
taxonomic_rank	0	series	31
phylum	1	bio-variety	32
class	2	candidate	33
order	3	cultivar	34
family	4	cultivar-group	35
genus	5	denominationclass	36
species	6	domain	37
subclass	7	graft-chimaera	38
subphylum	8	grex	39
subgenus	9	infraphylum	40
species_group	10	infrafamily	41
species_subgroup	11	infragenerictaxon	42
species_complex	12	infragenus	43
infraorder	13	infrakingdom	44
suborder	14	infraspecies	45
superclass	15	infraspecificTaxon	46
varietas	16	infratribe	47
kingdom	17	patho-variety	48
superfamily	18	specialform	49
infraclass	19	speciesaggregate	50
superorder	20	subvariety	51
parvorder	21	subsubvariety	52
superkingdom	22	subsection	53
subspecies	23	subseries	54
subfamily	24	subspecificaggregate	55
tribe	25	subsubform	56
forma	26	supertribe	57
superphylum	27	supragenerictaxon	58
subtribe	28	subform	59
subkingdom	29	no_rank	60
section	30		

Appendix D

Method evaluation data

All files relating to the technical validation and evaluation can be found in the github repository (https://github.com/Sandra-Young-Brighton/Appendices_thesis.git).

Files with the full breakdowns of comparisons made can be found in the appendices folder /Phase2/ and subsequent folders.

The Phase 2/Comparison with VTO/... files relate to those files in which the precision scores were calculated for the different scenarios, both in an overview analysis and the detailed analysis setting.

Files relating to the dual threshold part of the analysis are included in the /Phase 2/Freqsal/... folder.

Table D.1: Comparison of nomenclature pair relations ID'd per corpus according to different name lists and Word Sketches pulled (frequency filter)

	JEFF corpus		**WEB corpus**		
Filter	**JEFF (JEFF, WS subsection)**	**JEFF (JEFF, WS full)**	**WEB (JEFF, WS subsection)**	**WEB(JEFF, WS full)**	**WEB (WEB, WS full)**
No filter	1218	1715	1351	1581	4014
Relations over 5	227	257	284	323	510
Relations over 10	131	142	173	191	237
Relations over 15	89	96	132	145	167
Relations over 20	67	72	106	113	127
Relations over 25	53	57	92	97	103
Relations over 30	45	47	79	83	85
Relations over 35	41	41	70	73	73
Relations over 40	36	36	61	63	64
Relations over 45	32	32	54	56	57
Relations over 50	30	30	49	50	51
Relations over 55	24	24	43	44	44
Relations over 60	20	20	37	38	38
Relations over 65	19	19	36	37	37
Relations over 70	16	16	33	34	33
Relations over 75	15	15	31	32	31
Relations over 80	15	15	31	32	31
Relations over 85	15	15	26	27	26
Relations over 90	15	15	23	24	24
Relations over 95	14	14	23	24	23
Relations over 100	11	11	21	22	21

Table D.2: Comparison of nomenclature pair relations ID'd per corpus according to different name lists and Word Sketches pulled (salience filter)

	JEFF corpus		**WEB corpus**		
Filter	**JEFF (JEFF, WS subsection)**	**JEFF (JEFF, WS full)**	**WEB (JEFF, WS subsection)**	**WEB(JEFF, WS full)**	**WEB (WEB, WS full)**
No filter	1715	1218	1581	1351	4014
Salience over 0.5	1715	1218	1581	1351	4014
Salience over 1	1715	1218	1581	1351	4014
Salience over 1.5	1715	1218	1581	1351	4014
Salience over 2	1715	1218	1581	1351	4014
Salience over 2.5	1715	1218	1581	1351	4014
Salience over 3	1715	1218	1581	1351	4014
Salience over 3.5	1715	1218	1581	1351	4014
Salience over 4	1715	1218	1581	1351	4014
Salience over 4.5	1714	1217	1577	1349	3996
Salience over 5	1706	1216	1568	1344	3983
Salience over 5.5	1702	1215	1560	1338	3968
Salience over 6	1690	1213	1545	1327	3926
Salience over 6.5	1685	1210	1528	1316	3892
Salience over 7	1665	1190	1501	1289	3861
Salience over 7.5	1644	1170	1474	1262	3813
Salience over 8	1612	1139	1448	1236	3768
Salience over 8.5	1572	1099	1408	1196	3712
Salience over 9	1533	1060	1367	1157	3634
Salience over 9.5	1461	988	1301	1093	3472
Salience over 10	1402	934	1226	1015	3341
Salience over 10.5	1327	860	1131	926	3145
Salience over 11	1251	792	1042	833	2915
Salience over 11.5	1124	681	902	703	2660
Salience over 12	985	556	751	560	2249
Salience over 12.5	856	461	642	457	1922
Salience over 13	725	368	550	370	1600
Salience over 13.5	424	192	354	206	892

Table D.3: Breakdown of synonyms identified in JEFF and WEB (JEFF, WS subsection) corpora

Corpus	Source	Target	Difference
WEB	ACERINA	CERNUA	Gymnocephalus cernuus
WEB	APOLLONIA	MELANOSTOMA	Neogobius melanostomus
JEFF	BARBUS	SCLATERI	related synonym for Luciobarbus sclateri
JEFF	CATOSTOMUS	COMMERSONI	related synonym for Catostomus commersonii
WEB	CATOSTOMUS	COMMERSONI	related synonym for Catostomus commersonii: white sucker is a freshwater cypriniform fish
JEFF	CHONDROSTOMA	POLYLEPIS	related synonym for Pseudochondrostoma polylepis
WEB	CHONDROSTOMA	TOXOSTOMA	South-west European nase (Parachondrostoma toxostoma) is a species of cyprinid fish
JEFF	CHROSOMUS	ERYTHROGASTER	related synonym for Phoxinus erythrogaster
JEFF	COREGONUS	ARTEDII	related synonym for Coregonus artedi
WEB	COREGONUS	ARTEDII	related synonym for Coregonus artedi (lake herring) https://www.itis.gov/servlet/SingleRpt/SingleRpt?search_topic=TSN&
JEFF	COTTUS	BAIRDI	related synonym for Cottus Bairdii
WEB	COTTUS	BAIRDI	related synonym for Cottus Bairdii
JEFF	ENTOSPHENUS	TRIDENTATUS	related synonym for Lampetra tridentata
WEB	ENTOSPHENUS	TRIDENTATUS	related synonym for Lampetra tridentata
JEFF	GAMBUSIA	HOLBROOKI	related synonym for Gambusia affinis
WEB	GAMBUSIA	HOLBROOKI	related synonym for Gambusia affinis
JEFF	GYMNOCEPHALUS	CERNUA	real name, not sure why not in the ruffe or pope fish
WEB	GYMNOCEPHALUS	CERNUA	Eurasian ruffe (Gymnocephalus cernua), also known as ruffe or pope (http://www.catalogueoflife.org/col/details/species/id/349c65df03755f0
WEB	LEUCISCUS	ASPIUS	related synonym for Aspius aspius. asp is a European freshwater fish of the Cyprinid family.
JEFF	LEUCISCUS	PYRENAICUS	related synonym for Squalius pyrenaicus
JEFF	ONCORHYNCHUS	CLARKI	Related synonym for Oncorhynchus clarkii
WEB	ONCORHYNCHUS	CLARKI	Related synonym for Oncorhynchus clarkii
WEB	ONCORHYNCHUS	RHODURUS	Related synonym for Oncorhynchus masou
JEFF	PARASALMO	MYKISS	related synonym for Oncorhynchus mykiss
WEB	PLEURONECTES	AMERICANUS	The winter flounder (Pseudopleuronectes americanus), also known as the black back, is a right-eyed ("dextral") flatfish of the family Pleuronectidae.
WEB	SALMO	CLARKI	Oncorhynchus clarkii - cutthroat trout
JEFF	SALMO	GAIRDNERI	related synonym for Oncorhynchus mykiss
WEB	SALMO	GAIRDNERI	related synonym for Oncorhynchus mykiss
WEB	SCOPHTHALMUS	MAXIMUS	turbot is a species of flatfish in the family Scophthalmidae. Related synonym for Psetta maxima
JEFF	STIZOSTEDION	LUCIOPERCA	related synonym for Sander lucioperca
WEB	STIZOSTEDION	LUCIOPERCA	related synonym for Sander lucioperca
JEFF	STIZOSTEDION	VITREUM	related synonym for Sander vitreus
WEB	STIZOSTEDION	VITREUM	related synonym for Sander vitreus

Appendix E

Nomenclature Profiling Studies

E.1 Links to files of examples of the resources

Examples of ITIS files: Phase 3/Knowledge representation resources/

E.2 Oncorhynchus mykiss: resource variant comparison

E.3 Sander lucioperca: resource variant comparison

E.4 Salmo trutta: resource variant comparison

E.5 Concordances

Concordances for all the three nomenclature profiles can be accessed, along with the Word Sketches, in the relevant Nomenclature Profile Study (NPS) folders in the Appendices archive on github (see https://github.com/Sandra-Young-Brighton/Appendices_thesis.git).

Phase 3/NPS/Oncorhynchus mykiss/...
Phase 3/NPS/Sander lucioperca/...
Phase 3/NPS/Salmo trutta/...

Table E.1: Oncorhynchus mykiss name variants plus resource

Resource	**Classification**	**Name**
VTO	Accepted name	Oncorhynchus mykiss
CoL	Accepted name	Oncorhynchus mykiss (Walbaum, 1792)
ITIS	Accepted name	Oncorhynchus mykiss (Walbaum, 1792)
CoL	Common name	Baiser
CoL	Common name	Baja California rainbow trout
CoL	Common name	Bow
CoL	Common name	Brown trout
CoL	Common name	Coast angel trout
CoL	Common name	Coast rainbow trout
CoL	Common name	Coast range trout
CoL	Common name	Hardhead
CoL	Common name	Kamchatka salmon
CoL	Common name	Kamchatka steelhead
CoL	Common name	Kamchatka trout
CoL	Common name	Kamloops
CoL	Common name	Kamloops trout
CoL	Common name	Kamloops trout
CoL	Common name	Lord-fish
CoL	Common name	Rainbow
CoL	Common name	Rainbow trout
CoL	Common name	Rainbow trout
CoL	Common name	Rainbow trout
CoL	Common name	Rainbow trout
CoL	Common name	Rainbow trout
CoL	Common name	Rainbow trout
CoL	Common name	Rainbow trout
CoL	Common name	Rainbow trout
CoL	Common name	Rainbow trout
CoL	Common name	Rainbow trout
CoL	Common name	Rainbow trout
CoL	Common name	Rainbow trout
CoL	Common name	Rainbow trout
CoL	Common name	Rainbow trout
CoL	Common name	Rainbow trout
CoL	Common name	Rainbow trout
CoL	Common name	Rainbow trout
CoL	Common name	Rainbow trout
CoL	Common name	Rainbow trout
CoL	Common name	Rainbow trout
CoL	Common name	Rainbow trout
CoL	Common name	Rainbow trout
CoL	Common name	Rainbow trout
VTO	Common name	Rainbow trout
ITIS	Common name	rainbow trout
CoL	Common name	Redband
ITIS	Common name	redband trout
CoL	Common name	Salmon trout
CoL	Common name	Silver trout
CoL	Common name	Steelhead
CoL	Common name	Steelhead
CoL	Common name	Steelhead
CoL	Common name	Steelhead
ITIS	Common name	steelhead

Table E.2: Oncorhynchus mykiss name variants plus resource p.2

Resource	Classification	Name
CoL	Common name	Steelhead trout
CoL	Common name	Steelhead trout
CoL	Common name	Steelhead trout
CoL	Common name	Steelhead trout
CoL	Common name	Summer salmon
CoL	Common name	Trout
CoL	Synonym	Fario gairdneri (Richardson, 1836)
CoL	Synonym	Onchorhynchus mykiss (Walbaum, 1792)
CoL	Synonym	Onchorrhychus mykiss (Walbaum, 1792)
CoL	Synonym	Onchorynchus mykiss (Walbaum, 1792)
ITIS	Synonym	Onchorynchus mykiss (Walbaum, 1792)
CoL	Synonym	Oncorhynchus gairdnerii (Richardson, 1836)
VTO	Synonym	Oncorhynchus kamloops
CoL	Synonym	Oncorhynchus kamloops Jordan, 1892
VTO	Synonym	Oncorhynchus mykiss aguabonita
VTO	Synonym	Oncorhynchus mykiss gairdneri
ITIS	Synonym	Oncorhynchus mykiss gibbsi (Suckley, 1859)
VTO	Synonym	Oncorhynchus mykiss irideus
CoL	Synonym	Oncorhynchus mykiss nelsoni (Evermann, 1908)
CoL	Synonym	Oncorhynchus myskis (Walbaum, 1792)
VTO	Synonym	Parasalmo mykiss
CoL	Synonym	Parasalmo mykiss (Walbaum, 1792)
CoL	Synonym	Parasalmo penshinensis (Pallas, 1814)
VTO	Synonym	Salmo aquilarum
VTO	Synonym	Salmo gairdneri
VTO	Synonym	Salmo gairdneri beardsleei
CoL	Synonym	Salmo gairdneri gairdneri (Richardson, 1836)
VTO	Synonym	Salmo gairdneri gilberti
CoL	Synonym	Salmo gairdneri irideus Gibbons, 1855
CoL	Synonym	Salmo gairdneri kamloops (Jordan, 1892)
CoL	Synonym	Salmo gairdneri Richardson, 1836
VTO	Synonym	Salmo gairdneri shasta
CoL	Synonym	Salmo gairdneri shasta Jordan, 1894
VTO	Synonym	Salmo gairdneri stonei
VTO	Synonym	Salmo gairdnerii
CoL	Synonym	Salmo gairdnerii gairdnerii Richardson, 1836
CoL	Synonym	Salmo gairdnerii irideus Gibbons, 1855
CoL	Synonym	Salmo gairdnerii Richardson, 1836
VTO	Synonym	Salmo gibbsii
ITIS	Synonym	Salmo gibbsii Suckley, 1859
CoL	Synonym	Salmo gilberti Jordan, 1894
VTO	Synonym	Salmo iridea
CoL	Synonym	Salmo iridea Gibbons, 1855
VTO	Synonym	Salmo irideus argentatus
CoL	Synonym	Salmo irideus argentatus Bajkov, 1927
CoL	Synonym	Salmo irideus Gibbons, 1855
CoL	Synonym	Salmo irideux Gibbons, 1855
CoL	Synonym	Salmo kamloops (Jordan, 1892)

Table E.3: Oncorhynchus mykiss name variants plus resource p.3

Resource	**Classification**	**Name**
VTO	Synonym	Salmo kamloops whitehousei
CoL	Synonym	Salmo kamloops whitehousei Dymond, 1931
VTO	Synonym	Salmo masoni
CoL	Synonym	Salmo masoni Suckley, 1860
VTO	Synonym	Salmo mykiss
VTO	Synonym	Salmo mykiss
CoL	Synonym	Salmo mykiss Walbaum, 1792
ITIS	Synonym	Salmo mykiss Walbaum, 1792
VTO	Synonym	Salmo nelsoni
CoL	Synonym	Salmo nelsoni Evermann, 1908
VTO	Synonym	Salmo newberrii
CoL	Synonym	Salmo penshinensis Pallas, 1814
VTO	Synonym	Salmo purpuratus
CoL	Synonym	Salmo purpuratus Pallas, 1814
VTO	Synonym	Salmo regalis
VTO	Synonym	Salmo rivularis
CoL	Synonym	Salmo rivularis Ayres, 1855
CoL	Synonym	Salmo rivularis kamloops (Jordan, 1892)
VTO	Synonym	Salmo smaragdus
VTO	Synonym	Salmo truncatus
CoL	Synonym	Salmo truncatus Suckley, 1859
VTO	Synonym	Salmo whitei
CoL	Synonym	Trutta iridea (Gibbons, 1855)

Table E.4: Sander lucioperca name variant comparison plus resource

Resource	**Classification**	**Name**	**Match**
VTO	accepted name	Sander lucioperca	partial match
CoL	accepted name	Sander lucioperca (Linnaeus, 1758)	exact match
ITIS	accepted name	Sander lucioperca (Linneaus, 1758)	exact match
CoL	common	pikeperch	no match
CoL	common	pike-perch	exact match
VTO	common	pike-perch	exact match
CoL	common	zander	exact match
ITIS	common	zander	exact match
VTO	synonym	Centropomus sandat	partial match
CoL	synonym	Centropomus sandat Lacepède, 1802	partial match
VTO	synonym	Lucioperca linnei	partial match
CoL	synonym	Lucioperca linnei Malm, 1877	partial match
VTO	synonym	Lucioperca lucioperca	partial match
CoL	synonym	Lucioperca lucioperca (Linnaeus, 1758)	partial match
VTO	synonym	Lucioperca sandra	partial match
CoL	synonym	Lucioperca sandra Cuvier, 1828	partial match
VTO	synonym	Perca lucioperca	partial match
CoL	synonym	Perca lucioperca Linnaeus, 1758	partial match
VTO	synonym	Sander lucioperca (Linnaeus, 1758)	exact match
VTO	synonym	Stizostedion lucioperca	partial match
ITIS	synonym	Stizostedion lucioperca (Linnaeus, 1758)	exact match
CoL	synonym	Stizostedion lucioperca (Linnaeus, 1758)	exact match
CoL	synonym	Stizostedion luciperca (Linnaeus, 1758)	no match
CoL	synonym	Stizostedium lucioperca (Linnaeus, 1758)	no match

Table E.5: Salmo trutta name variant comparison plus resource

Resource	Classification	Name	Match
VTO	accepted	Salmo trutta	partial match
CoL	accepted	Salmo trutta Linnaeus, 1758	exact match
ITIS	accepted	Salmo trutta Linnaeus, 1758	exact match
CoL	ambiguous synonym	Salmo fario major Walecki, 1863	no match
CoL	ambiguous synonym	Salmo orientalis McClelland, 1842	no match
CoL	ambiguous synonym	Salmo stroemii Gmelin, 1789	no match
CoL	ambiguous synonym	Trutta fluviatilis Duhamel, 1771	no match
CoL	ambiguous synonym	Trutta salmanata Strøm, 1784	no match
CoL	ambiguous synonym	Trutta salmonata Rutty, 1772	no match
CoL	common	Amu-Darya trout	no match
CoL	common	Aral salmon	no match
CoL	common	Aral Sea Trout	no match
CoL	common	Aral trout	no match
CoL	common	Blacktail	no match
CoL	common	Brook trout	no match
CoL	common	Brown trout	exact match
ITIS	common	brown trout	exact match
CoL	common	Brownie	no match
CoL	common	Finnock	no match
CoL	common	Galway sea trout	no match
CoL	common	Gillaroo	no match
CoL	common	Herling	no match
CoL	common	Hirling	no match
CoL	common	Lake trout	no match
CoL	common	Loch leven trout	no match
CoL	common	Orange fin	no match
CoL	common	Orkney sea trout	no match
CoL	common	Peal	no match
CoL	common	River trout	no match
CoL	common	Salmón	no match
CoL	common	Salmon trout	no match
CoL	common	Sea trout	exact match
VTO	common	Sea trout	exact match
CoL	common	Sea-trout	no match
CoL	common	Sewin	no match
CoL	common	Trout	no match
CoL	common	Whiting	no match
CoL	common	Whitling	no match
CoL	misapplied name	Salmo trutta ciscaucasicus (non Dorofeeva, 1967)	no match
CoL	misapplied name	Salmo trutta ezenami (non Berg, 1948) (misapplied name)	no match
VTO	synonym	Fario argenteus	partial match
CoL	synonym	Fario argenteus Valenciennes, 1848	partial match
CoL	synonym	Fario lacustris (Linnaeus, 1758)	no match
CoL	synonym	Fario trutta (Linnaeus, 1758)	no match
VTO	synonym	Salar ausonii	partial match
CoL	synonym	Salar ausonii parcepunctata Heckel & Kner, 1858	no match
CoL	synonym	Salar ausonii semipunctata Heckel & Kner, 1858	no match

Table E.6: Salmo trutta name variant comparison plus resource p.2

Resource	Classification	Name	Match
CoL	synonym	Salar ausonii Valenciennes, 1848	partial match
VTO	synonym	Salar bailloni	partial match
CoL	synonym	Salar bailloni Valenciennes, 1848	partial match
VTO	synonym	Salar gaimardi	partial match
CoL	synonym	Salar gaimardi Valenciennes, 1848	partial match
VTO	synonym	Salar macrostigma	no match
VTO	synonym	Salar spectabilis	partial match
CoL	synonym	Salar spectabilis Valenciennes, 1848	partial match
VTO	synonym	Salmo albus	partial match
CoL	synonym	Salmo albus Bonnaterre, 1788	partial match
CoL	synonym	Salmo albus Walbaum, 1792	partial match
VTO	synonym	Salmo brachypoma	partial match
CoL	synonym	Salmo brachypoma Günther, 1866	partial match
VTO	synonym	Salmo caecifer	partial match
CoL	synonym	Salmo caecifer Parnell, 1838	partial match
VTO	synonym	Salmo cambricus	partial match
CoL	synonym	Salmo cambricus Donovan, 1806	partial match
VTO	synonym	Salmo caspius	no match
VTO	synonym	Salmo cornubiensis	partial match
CoL	synonym	Salmo cornubiensis Walbaum, 1792	partial match
VTO	synonym	Salmo cumberland	partial match
CoL	synonym	Salmo cumberland Lacepède, 1803	partial match
VTO	synonym	Salmo eriox	partial match
CoL	synonym	Salmo eriox Linnaeus, 1758	partial match
VTO	synonym	Salmo estuarius	partial match
CoL	synonym	Salmo estuarius Knox, 1855	partial match
VTO	synonym	Salmo fario	no match
VTO	synonym	Salmo fario forestensis	partial match
CoL	synonym	Salmo fario forestensis Bloch & Schneider, 1801	partial match
CoL	synonym	Salmo fario Linnaeus, 1758	no match
VTO	synonym	Salmo fario loensis	no match
CoL	synonym	Salmo faris forestensis Bloch & Schneider, 1801	no match
VTO	synonym	Salmo gadoides	partial match
CoL	synonym	Salmo gadoides Lacepède, 1803	partial match
VTO	synonym	Salmo gallivensis	partial match
CoL	synonym	Salmo gallivensis Günther, 1866	partial match
VTO	synonym	Salmo illanca	partial match
CoL	synonym	Salmo illanca Wartmann, 1783	partial match
VTO	synonym	Salmo islayensis	partial match
CoL	synonym	Salmo islayensis Thomson, 1873	partial match
VTO	synonym	Salmo lacustris	partial match
CoL	synonym	Salmo lacustris Linnaeus, 1758	partial match
VTO	synonym	Salmo lacustris rhenana	partial match
CoL	synonym	Salmo lacustris rhenana Fatio, 1890	partial match
VTO	synonym	Salmo lacustris romanovi	partial match
CoL	synonym	Salmo lacustris romanovi Kawraisky, 1896	partial match
VTO	synonym	Salmo lacustris septentrionalis	partial match
CoL	synonym	Salmo lacustris septentrionalis Fatio, 1890	partial match
VTO	synonym	Salmo lemanus	partial match
CoL	synonym	Salmo lemanus Cuvier, 1829	partial match
VTO	synonym	Salmo levenensis	partial match

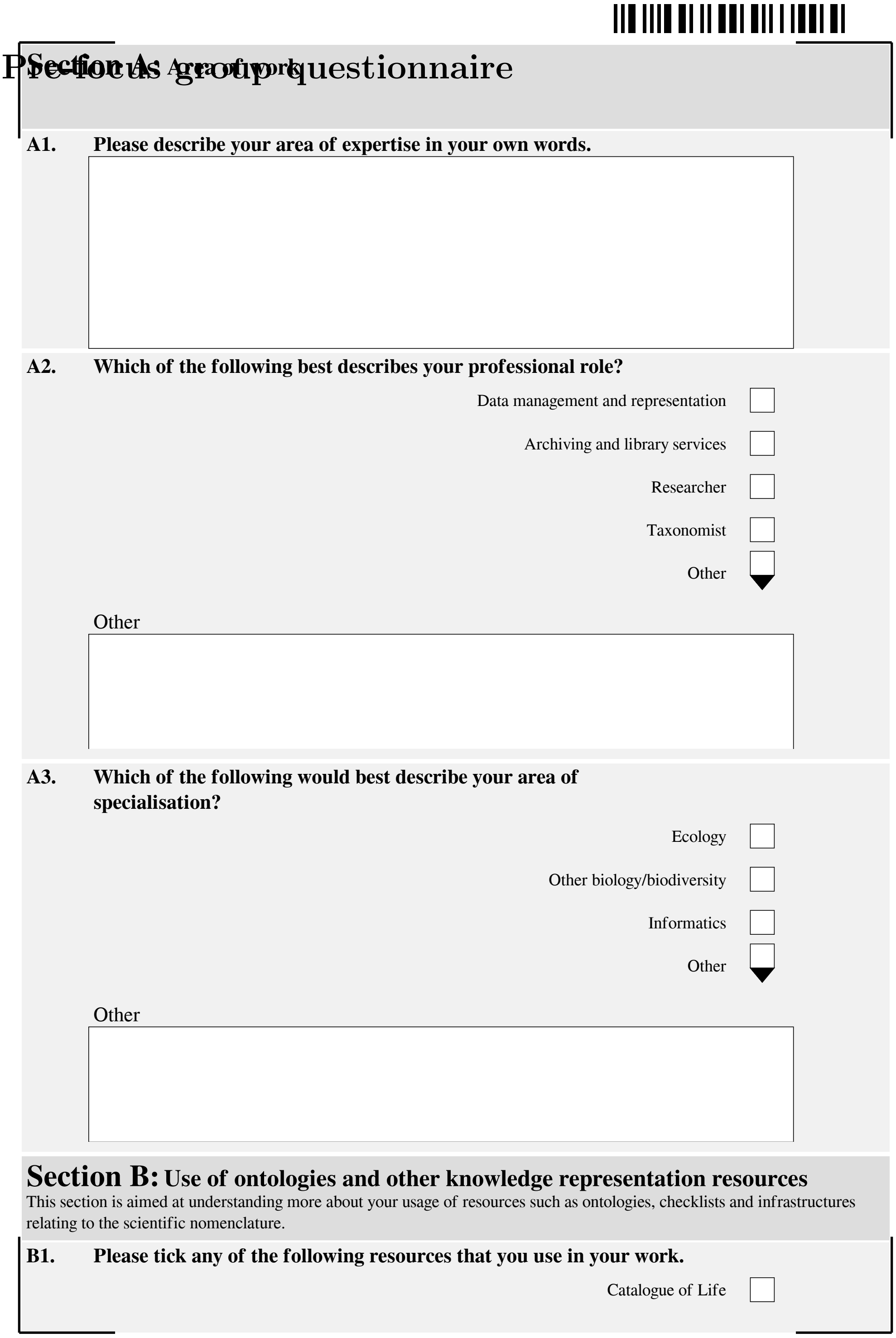
Table E.7: Salmo trutta name variant comparison plus resource p.3

Resource	**Classification**	**Name**	**Match**
CoL	synonym	Salmo levenensis Yarrell, 1839	partial match
VTO	synonym	Salmo microps	no match
VTO	synonym	Salmo mistops	partial match
CoL	synonym	Salmo mistops Günther, 1866	partial match
CoL	synonym	Salmo montana Walker, 1812	no match
VTO	synonym	Salmo orcadensis	partial match
CoL	synonym	Salmo orcadensis Günther, 1866	partial match
VTO	synonym	Salmo oxianus	partial match
CoL	synonym	Salmo oxianus Kessler, 1874	partial match
VTO	synonym	Salmo phinoc	partial match
CoL	synonym	Salmo phinoc Shaw, 1804	partial match
VTO	synonym	Salmo polyosteus	partial match
CoL	synonym	Salmo polyosteus Günther, 1866	partial match
VTO	synonym	Salmo rappii	partial match
CoL	synonym	Salmo rappii Günther, 1866	partial match
VTO	synonym	Salmo saxatilis	partial match
CoL	synonym	Salmo saxatilis Schrank, 1798	partial match
CoL	synonym	Salmo spurius Pallas, 1814	no match
VTO	synonym	Salmo sylvaticus	partial match
CoL	synonym	Salmo sylvaticus Gmelin, 1789	partial match
CoL	synonym	Salmo taurinus Walker, 1812	no match
VTO	synonym	Salmo trutta aralensis	partial match
CoL	synonym	Salmo trutta aralensis Berg, 1908	partial match
VTO	synonym	Salmo trutta caspius	no match
VTO	synonym	Salmo trutta ciscaucasicus	no match
VTO	synonym	Salmo trutta fario	partial match
CoL	synonym	Salmo trutta fario Linnaeus, 1758	partial match
CoL	synonym	Salmo trutta lacustris Linnaeus, 1758	no match
VTO	synonym	Salmo trutta macrostigma	no match
VTO	synonym	Salmo trutta oxianus	partial match
CoL	synonym	Salmo trutta oxianus Kessler, 1874	partial match
VTO	synonym	Salmo trutta trutta	partial match
CoL	synonym	Salmo trutta trutta Linnaeus, 1758	partial match
CoL	synonym	Salmo truttula Nilsson, 1832	no match
VTO	synonym	Salmo vario	no match
VTO	synonym	Salmo venernensis	partial match
CoL	synonym	Salmo venernensis Günther, 1866	partial match
CoL	synonym	Trutta fario (Linnaeus, 1758)	no match
VTO	synonym	Trutta fario macroptera	no match
CoL	synonym	Trutta lacustris (Linneaus, 1758)	no match
VTO	synonym	Trutta marina	partial match
CoL	synonym	Trutta marina Duhamel, 1771	partial match
CoL	synonym	Trutta marina Moreau, 1881	partial match
CoL	synonym	Trutta trutta (Linnaeus, 1758)	no match
VTO	synonym	Trutta variabilis	partial match
CoL	synonym	Trutta variabilis Lunel, 1874	partial match
CoL	misapplied name	Salmo fario loensis Walbaum, 1792	no match

Appendix F

Focus group materials

F.1 Pre-focus group questionnaire



Section A: Area of work

A1. **Please describe your area of expertise in your own words.**

A2. **Which of the following best describes your professional role?**

- Data management and representation □
- Archiving and library services □
- Researcher □
- Taxonomist □
- Other □

Other

A3. **Which of the following would best describe your area of specialisation?**

- Ecology □
- Other biology/biodiversity □
- Informatics □
- Other □

Other

Section B: Use of ontologies and other knowledge representation resources

This section is aimed at understanding more about your usage of resources such as ontologies, checklists and infrastructures relating to the scientific nomenclature.

B1. **Please tick any of the following resources that you use in your work.**

- Catalogue of Life □

Vertebrate Taxonomy Ontology □

Encyclopedia of Life □

FishBase □

NCBI classification □

None of the above □

Other □

Other

□

B2. **Please list any other knowledge representation resources (scientifc nomenclature databases, checklists, ontologies) you use frequently.**

□

B3. **Choose from the reasons below why you use these resources. If you use these resources for other reasons too, you can add your own in the "other" section.**

Check name variant status □

Check taxon classification of name variant □

Map data to resource □

Compare taxon classification of name variant between resources □

Use resource to annotate data □

Other □

Other

□

B4. **Select from the following (or add) any problems you have with the resources.**

Incomplete □

Conflicting information □

Inaccurate information □

Ambiguous information □

Lacking detail □

N/A □

Other □

Other

B5. **Select positive features of the resources from the following (or specify any not listed in the "other" section).**

Comprehensive □

Well-curated □

Accurate □

Clear □

Other □

Other

B6. **If you have any other comments about your experiences using scientific nomenclature knowledge representation resources (ontologies, taxonomies, checklists, databases) that were not covered in the previous questions, please leave them here.**

Section C: Scientific nomenclature usage

This set of questions will ask you briefly about scientific nomenclature usage in general, the ambiguity or clarity in its usages, and your experience of this in your work.

C1. Which of the following characteristics would you associate with consistent scientific nomenclature usage?

One variant □

Multiple variants □

Authorship □

Open nomenclature abbreviations □

One spelling □

Multiple spelling □

Static in time □

Geographical-specificity □

Time-dependent □

Other □ ▼

Other

□

C2. Which of the following characteristics do you associate with ambiguous nomenclature usage?

One variant □

Multiple variants □

Authorship □

Open nomenclature abbreviations □

One spelling □

Multiple spelling □

Static in time □

Geographical-specificity □

Time-dependent □

Other □

Other

C3. **If applicable, please provide examples of taxa for which scientific nomenclature tends to be used:**

Consistently □

Comment

Ambiguously □

Comment

Other □

Other

C4. **Which types of ambiguity do you face as regards scientific nomenclature in your work?**

Vernacular usage □

Contextual ambiguity □

Inconsistent usage of authorship □

N/A □

Other □

Other

C5. **Are some types of ambiguity more difficult to deal with than others?**

Yes □

No □

C6. Please order the types of ambiguity you have selected/provided in difficulty to handle (1 = most difficult to 3 = least difficult)

Vernacular usage □□□□□□□□□□

Contextual ambiguity □□□□□□□□□□

Inconsistent usage of authorship □□□□□□□□□□

N/A □□□□□□□□□□

Section D: Misspellings and synonyms

This section briefly asks you opinion about the usage of misspellings and synonyms in the scientific nomenclature.

D1. In the scientific nomenclature, where an accepted taxon name has a number of variants, do you consider them to be synonyms?

Yes □

No □

Depends □

D2. If you do not consider variants to be synonyms, what do you consider them to be?

Invalid variants □

Variants □

Synonyms □

Alternatives □

Depends □

Other □▼

Other

D3. How frequently do you come across variants in your work?

Very infrequently □

Infrequently □

Sometimes □

Frequently □

Very frequently □

All the time □

D4. Where do you see variants used more frequently (if anywhere)?

Academic journals □

Citizen science articles □

Webpages □

Textbooks □

All of the above □

None of the above □

Other □

Other

D5. Does the use of these variants cause any ambiguity in your work?

Yes □

No □

D6. If they do cause ambiguity, can you describe what sort of ambiguity?

D7. How common are misspellings in the scientific literature?

Very rare □

Rare □

Uncommon □

Common □

Very common □

All the time □

D8. **Do misspelled variants ever appear more frequently than the correctly spelled variant?**

Yes □

No □

D9. **If you know of any examples, please list here.**

□

Section E: Usage and meaning of vernacular variants

This section will explore your opinions on vernacular name usage and its interplay with scientific nomenclature. It will also explore a little about possible ambiguity that can arise.

E1. **How often do you come across vernacular variants of the nomenclature in your work?**

Very infrequently □

Infrequently □

Sometimes □

Frequently □

Very frequently □

All the time □

E2. **In your opinion, do vernacular variant names usually have a more specific or broader meaning?**

Broader □

More specific □

E3. **Which of the following bits of information (if any) can vernacular variants be used to convey, that are not conveyed by their scientific equivalents?**

Geographical-specificity □

Comment

□

Domain-specificity □

Comment

□

Time-specificity ▼

Comment

Context-specificity ▼

Comment

Lifestage-specificity ▼

Comment

Unknown variants ▼

Comment

Other ▼

Other

E4. Does the usage of common variants cause ambiguity in your work?

Yes □

No □

E5. What sort of ambiguity? Please order the types of ambiguity in order of importance (1 = most important, 5 = least important).

	1	2	3	4	5
Cover multiple taxa	□	□	□	□	□
Vague	□	□	□	□	□
Unknown variants	□	□	□	□	□
Contextual ambiguity	□	□	□	□	□
Inconsistently-used variants	□	□	□	□	□

Thank you for completing the pre-focus group questionnaire. I look forward to meeting you on the day.

If you have any questions prior to this, please contact me at:
s.h.young@brighton.ac.uk

F.2 Focus group outline

Focus group: slides with related questions for group

(15 minutes for my introduction, ontologies, problem and approach)

Nomenclature use and stability study: introduction slide

- Introduce my work, me, the approach I am taking
- Mention ontologies: if people do not know move onto the following two slides to explain (in fact I think they are useful anyway because it presents ontologies in a way that I perceive them)

Hello! Thank you for agreeing to participate in this focus group/outreach day. Firstly, I will introduce myself and my project briefly, then I will ask you all to introduce yourselves.

I am Sandra, PhD student at CEM, University of Brighton. Translator and interpreter (linguist background) and for the last 3 ½ years have been here working on my PhD. The focus of the PhD is on knowledge representation and integration. It considers difficulties experienced in many domains of accurately integrating data due to ambiguities in natural language. Things such as ontologies are increasingly popular in this task but there have been problems identified which I aimed to tackle in my research.

Ontology: slide 1

- Ontology as more complex relations
- Demonstration of why can infer information etc. so why useful

[Does everyone know what an ontology is? For anyone who doesn't know, an ontology is a formal and explicit conceptualisation of a domain or defined thing. It is so useful in data integration and searching because of the way it defines classes (or concepts) and the relations between them, which allows for much broader searches, for example, because it can use these models to infer information (different words used for the same term, relations to this term with other things, etc.)].

Ontology: slide 2

- Ontology as a taxonomy
- While some ontologies are more complete (different sorts of relations) others are more like taxonomies, as shown here
- How my research has worked in the first instance: could be developed to demonstrate more complex relations

The problem

- Ontologies really useful
- However, question as to accurate integration of data
- Ontologies by nature have to exclude at some point (explicit conceptualisation), don't necessarily take into account how presented in the data
- Biodiversity and scientific nomenclature and why identified

The problem I seek to address in my research is related to applying these ontologies to data and the issues in which data will be ignored because it does not seem to fit with the logic of the ontology, or worse, incorrectly integrated because, despite it not fitting with the conceptual model, this is not picked up. [pick up on this later with the brown trout and lake and brook trout bits]. The domain focus of my research is the biodiversity literature, and specifically looking at modelling scientific nomenclature and vernacular usage within. The biodiversity literature is particularly relevant because of one, the importance of scientific nomenclature for organising and ordering biodiversity data, but two because of the complexities relating to its usage and meaning (how multiple, fluid it is, etc.).

Purpose is not to create an ontology as such but to use linguistic analysis techniques to create a comparable model/representation. Future work who knows.

My approach

- Explain how I use linguistic clues in the text to identify relations
- Then the Word Sketches
- Then I convert to graphs for visualisation purposes

Focus group plan

- Introductions
- Nomenclature Profile Studies: introduction
- Hierarchy identification
- Knowledge representation: your experiences and my analysis
- Corpus data: analyses and discussion

For this focus group, therefore, I was interested in people who have a knowledge and understanding of scientific nomenclature and its usage, from different perspectives. You have already filled out a pre-focus group questionnaire, which I have used to mould the shape of today's discussion. The questionnaire has focused on your general thoughts towards nomenclature usage, vernacular usage and ambiguities that arise in these usages. Today I am going to show you the results of my work, so we can discuss the outcomes relating to the pre-focus group questionnaire, but in the context of real data. First I will ask you to introduce yourselves though ☺

Introduction (5 minutes)

Introduce yourselves

- Ask the participants to introduce themselves: specialism, biological background if any, role
- Ask participants what sort of contexts they use scientific nomenclature/read scientific nomenclature
- What difficulties in meaning do they come across? Where are they most likely to have these difficulties? How do they tackle these issues?

[redacted]

[redacted]

[redacted]

Nomenclature profile studies **(5 minutes)**

- Introduce the work I am going to present: explain how they work [knowledge representation comparison to look at variation in reliable resources, then look at the corpora, then compare against the resources and the corpora]

Hierarchy identification **(5 minutes)**

- Explain that as scientific nomenclature and vernacular variants: multi-part (options of how to deal with)
- Each word as unit or unified
- Explain that in the NPS I focused on the latter because of the wish to consider multi-word terms as unit "term as unit" but could be developed further in other way if deemed useful
- From a brief look at the graphs [have spares ready to show people to discuss]
- Go back through: what are the main features of each (species level as joining nodes or genus as hub nodes versus common names/general terms as hub nodes and species as around)

Questions (focus group):

1. Is there anything in the different representations that call out to you as useful one way or the other? If so, what?

Discussion about knowledge representation (including my bit) (10-15 minutes)

Knowledge representation: your experiences

- Lead on from their answers in the pre-focus group questionnaire (maybe not need a lot of these) – pull out if any differences or similarities in the responses and explore

[redacted]

Questions (focus group):

1. Ask about how they choose, what is important criteria, do these criteria vary? What are your criteria? Is it context dependent? Related to your role?
2. Do you find resources to be consistent?
3. Do you understand why different resources present information in different ways/include different information and is that useful?
4. Do you have to use different resources for different purposes?
5. Or multiple resources on a task?
6. What do you think is lacking?
7. What is good?

1. [redacted] both put comprehensive, accurate etc. do you choose these resources specifically because they are the ones that are accurate? Or generally accurate?
2. [redacted] commented that used different resources to compare taxon classification between different resources: is this necessary for all orders of beings in your experience or just in some areas?
3. How do the different classifications affect how you write about/your choices on how to interpret data?
4. [redacted] also mentions the comparative trait-extinction analyses: can you give any further detail?
5. When you are doing these things is this for narrative or database data (are there differences)?
6. [redacted] talked about using the resource to annotate: can you give me a bit more detail on that? (then see if [redacted] has anything to share)
7. [redacted] also says conflicting and ambiguous – then in the comment about the time-lag/mismatch between IUCN and most current recognised avian taxonomy: two questions – one relating to the comparison in general – you use to compare to see if everyone agrees? But the most current recognised avian taxonomy implies that there is one agreed taxonomy at a time – can you explain? Sorry if it is a silly question but hard and also very important for me to have clear in my mind.

Knowledge representation: NPS

- As I mentioned, in the studies I compared three resources: explain why and how I compared
- Describe the 3

[redacted] Go into different detail depending on whether they know/use any of them [redacted]

- If do use: ask opinion on them
- Have the slides there but can also go onto the internet to show

<u>*Questions (focus group):*</u>

1. If you use any of the three resources: why, what is your opinion of them/it?
2. Any issues with these sorts of databases: that compile many different ones?

Discussion in general about ambiguity and the general try for 20-30 minutes

Corpus data: general results [overarching characteristics in the data]

- Authorship
- Common name usage
- Coverage/Variation (explain why I have used both of these words)

Here ambiguity: authorship inconsistent usage; contextual ambiguity (find where to explain)

Questions [ask before providing any information about the data] (focus group questions)

This is the kind of focus group bit – then I go on to share my data and it is more a discussion with outreach and feedback about my data

1. What sort of data do you primarily work with (database, narrative, etc.)?
2. What sort of narrative texts would you usually work with?
3. What are your opinions on authorship usage in the literature?
4. Do you think there are differences between narrative text and databases in the adherence to authorship?

5. In your experience, how are scientific nomenclature variants used in biodiversity literature? Could specify which literature: academic articles, taxonomic circumscriptions, newspapers, and so on.
   In the questionnaire say that find variants etc. in all forms. Are there any differences in what/how you would expect differences to arise?
6. Can you describe variation of usage of scientific nomenclature variants and any ambiguities that may arise from this – and discuss if your opinions are the same/different. Also any differences between different sorts of media (journal articles, circumscriptions, webpages, citizen science).
7. Ambiguity in the questionnaire returned mixed results. Can we discuss reasons for thinking ambiguous or not in the variation of nomenclature?
8. [redacted] multiple variants and geographical specificity being both consistent and ambiguous usage – can [redacted] explain further?
9. Ambiguity as regards geographical-specificity and static in time – can you explain what you think about these points?
10. Does this multiplicity cause ambiguity? Or does it not have to? Is it only a problem for taxonomists for example?

Explain the contextual ambiguity (are some scientific variants used in context to mean one thing or another or is it sometimes unclear in the context what the variant means) and then authorship inconsistency is whether used or not, or part of the authorship etc.

11. Not on the basis of this can you answer the question better?
12. How do the decisions of splitters and lumpers affect you in your work? (put in the little anecdote about dictionary splitters and lumpers)
13. Species and subspecies and ambiguity: go into more detail (and link back to the conversation about the "current accepted avian taxonomy") – this links to the above question
14. As regards variants can I ask you three to discuss how you would define the following terms (in the context of scientific nomenclature):
    Synonyms
    Misspellings
    Are there other terms you can think of that it would be worth discussing? Does this cause ambiguity? (depending on how people answer we may see that people have different ideas about what it means)

15. In your experience, how are vernacular variants used in biodiversity literature? Could specify which literature: academic articles, taxonomic circumscriptions, newspapers, and so on.
    Where do you come across them?
    Say you come across them infrequently – is that because infrequently use narrative data?
    Rachel says very frequently (here is the difference linked to the type of data she works with for example)
    Geographical and context-specificity: open a discussion on these two (ask what they understand by these descriptions first).

16. Can you describe variation of usage of vernacular variants and any ambiguities that may arise from this – and discuss if your opinions are the same/different. Also any differences between different sorts of media (journal articles, circumscriptions, webpages, citizen science).
    First check what they all understood by each option.
    In this can discuss the different levels of importance given [redacted]
    [redacted] important: multiple taxa, vague, unknown and inconsistently used. Less important: contextual ambiguity.
    [redacted]: most important: inconsistently used. Mid: vague, unknown, contextual ambiguity. Least important: multiple taxa.
    [redacted] (Most to least): multiple taxa, vague, inconsistently used, vague, contextual
    Probe about these – explain what I meant by contextual ambiguity if necessary. Ask them what they understood. I meant (for clarity) that because of multiple usage in context it is not clear what mean. In the same context it can mean various things (sometimes if consistently used, variants not ambiguous because clear by the context, but if this is not the case, then this generates contextual ambiguity).
    With inconsistent usage: how do you find out what it means? Are there times in which you do not know?

After this can go into discussion about the specific issues.

Authorship **5-10 minutes**

- Just describe how there were none of the identified in the corpora with the full authorship. Some had just Linnaeus.

<u>Questions</u>

1. Is this normal in your areas of work?
2. Is this linked to the narrative text or is it more generalised?
3. Is it to do with the domain? If we had entomologists here would they have a different perspective? Or for example, [redacted] you mentioned the Elasmobranchi is it different there?
4. Discuss the implications of this: from the biology side does this have implications
5. From the technical side (not tagging, for example)? Historically? [these questions can be directed in general at everyone and should elicit different responses]

Variant coverage: summary page **(10-15 minutes)**

- Show the differences in variation
- The effect of authorship
- The contrast between Onc my and sal trutta in coverage
- High coverage of common names all round

[Have more details if they want in the subsequent slides]

Variant variation (spread or trend): Oncorhynchus mykiss

- Discuss

Variant variation (spread or trend): Sander lucioperca

- Discuss
-

Variant variation (spread or trend): Salmo trutta

- Discuss

With the variation:

- Think about what is important to pull out of here
- Coverage versus spread (concentration of terms)
- Vernacular coverage
- Difference between Salmo trutta and Oncorhynchus mykiss
- Spread of usage: perhaps important to see if one surprised but also if these trends are normal or if there are areas in which this is much more messy

<u>Questions (after seeing)</u>

1. What are potential difficulties that may arise because of this variation (if any)? [Ambiguity of meaning, connecting like with like, processing capabilities, being aware that one term is congruent or not with another term]
   Maybe more probing about whether including or excluding in tagging?
   Or just excluding or including in general? Benefits and issues with that.
2. To what extent does this variation affect your work? If I have a mix of biologist, historian and informatics: want to ask this question in a way that will ensure that I pick up on the nuances of how these differences affects people differently [or find that some shared].
3. About the terms themselves: comments on the variation/differences in the variation, any insight

Specific points of ambiguity

- Use the previous discussion to lead onto that. My research focuses on the terminological side of scientific nomenclature usage. However, some of these patterns could be used to indicate if there are differences (or at least perceived differences) between the variants. Really not sure if these are true synonyms. I have not done a review of the circumscriptions of these scientific nomenclature variants. Not sure if [redacted] could shed some light as we go along. Just looking at the usage as per the evidence in the test corpora. Obviously this means that only what is there in the corpora can be used as evidence: lack of evidence cannot be used to prove anything, simply that there is no data to say either way. Important to bear in mind.

- Split ambiguities into four: vernacular, spelling, official taxonomic validity versus usage, contradictions between data and resource

Questions: ask before showing actual data **(5-10 minutes)**

This again is more focus group, then when look at the data it is more outreach

1. Discuss the categories – what do you understand by these categories?
2. Are there other ambiguity types you would expect to be identified?
3. How would you classify ambiguity?
4. Give examples of ambiguity you experience?

- Explain ambiguities as I have categorised them. Make comments about how they responded to my questions. Add anything that has come out of the discussion.

1. Any comments?

Go into explaining my data: this is more outreach and then just get comments on my interpretation, the presentation. Also mention that most related to common name usage even though different issues

Vernacular ambiguity: broader meaning (use this one to explain my graphs a bit)

- Trout
- This is clear but questions the classification of trout there. Not wrong but the nuance of meaning is not presented in the CoL and is shown here

Question

1. Can it be useful to have a more general term included like that without any explanation? If so, why?
2. Is that a problem? If so, why?

Before going into broader or narrower meaning, explain why. Also look at the focus group answers.

- Common names: particularly in focus group see that extra information only marked as geographical (can discuss?)

- [redacted] steelhead and rainbow trout what are the differences? Are there other examples? Can people give examples?

Vernacular meaning: broader or narrower meaning

- Sea trout
- Explain. Explain small amount of data also how the method lets you check how it is correct (go into Sketch Engine or show a concordance)
- But shows the different context: sea trout (with the clarification as Salmo trutta) or sea trout as the anadromous form of Salmo trutta. These sorts of things can be incorrectly imposed.

Question

1. How do you see this distinction? Do you agree with my interpretation?
2. Are there times where these sorts of distinctions are important in your work? If so, when?
3. Can you describe any other sorts of distinctions like this?

Ambiguities: increased specificity of usage

- Kamchatka
- Geographical

Questions

1. Only limited data but ask [redacted] if makes any sense
2. Ask others if this is something come across? Are there differences in meaning?
3. Are these the sorts of variants you meant when said geographical? Are these the only kinds? Can you give me examples of the ones you were thinking of?

Ambiguities: increased specificity of usage

- Steelhead and rainbow trout
- Consistently parents (show that more general than each of the scientific nomenclature variants they can represent)
- But multiple terms for each, feed into Onc my
- Show how I have seen in the data indications meaning more specific (variation and then graphs)
- Read about: looks like anadromous again involved.
- None of the resources take this into account (guessing because taxonomically not important)

Questions

1. For any of you in your work it this sort of distinction important?
2. If so when, if not why?

Ambiguities: nomenclature spelling

- Salmo gairdneri or gairdnerii (considering past discussion – do you consider this a misspelling?)
- Discuss which may be proper spelling

Questions

1. How does misspelling affect your work?
2. Can you think of any examples in which misspellings appear more frequently than valid taxonomic names?
3. How can the non-inclusion of this in databases cause issues to accessing data?
4. Or issues in tagging or annotating data?

Ambiguities: name usage (variant or accepted name)

- Sander lucioperca versus Stizostedion lucioperca
- Show the results, explain which seems like the most used
- Interesting that JEFF preferred old term (is it a time-related thing)?

Questions

1. What does/can it mean when a variant like this is used?
2. Is this common?
3. If [redacted] is there, does he know about this? Can he give me a back story?
4. Do these sorts of issues cause problems? If so, how?

Contradiction: common names and links to scientific nomenclature

Brown trout and Oncorhynchus mykiss

Lake trout and brook trout and their respective nomenclature: ask if anyone knows

Demonstrate how this can be used to determine how the terms are being used: indicate issues, not prove

Questions

1. Are these common names that strictly linked to the ones shown in the corpora?
2. Are there times in which particularly the latter two could be used as Salmo trutta? Or other Salmonidae?
3. How much does the ambiguous nature of common names cause problems in your work?

End

- Any questions they might have, then go into the following

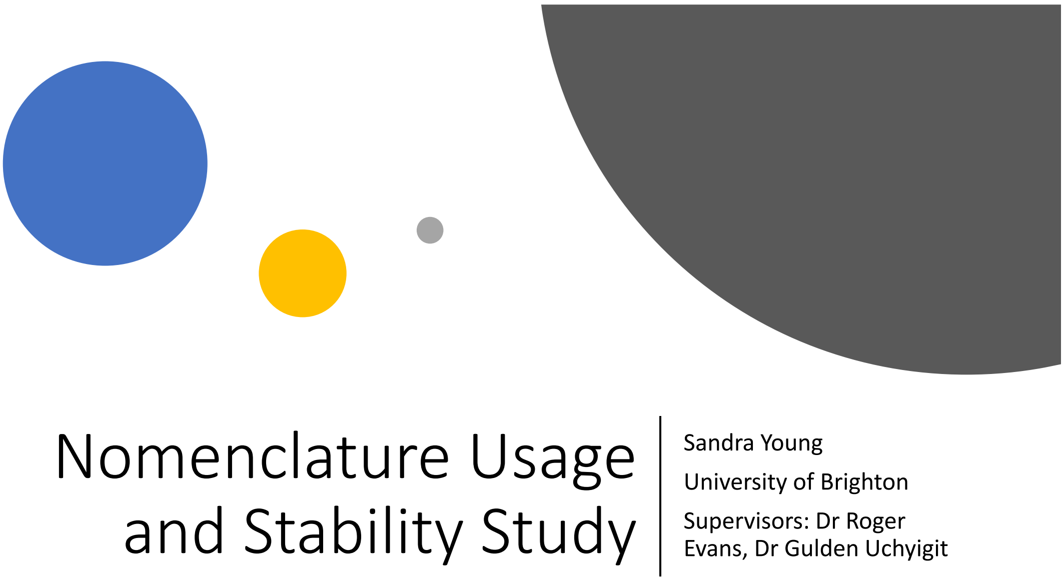
Approach and presentation of findings (15 minutes)(OUT)

1. Considering the issues discussed in the focus group today, do you think the network graph presentations accurately reflect the:
   - ambiguities present in the data??
   - differences between the different data sets
2. Considering the issues discussed in the focus group today, do you think the network graph presentations can be used to:

- disambiguate meaning between usage of different terms in the data?
- highlight hidden information about term usage not included in the ontologies?

For both of the above questions: [yes/no] If so, why? Why not?

F.3 Focus group slides



Nomenclature Usage and Stability Study

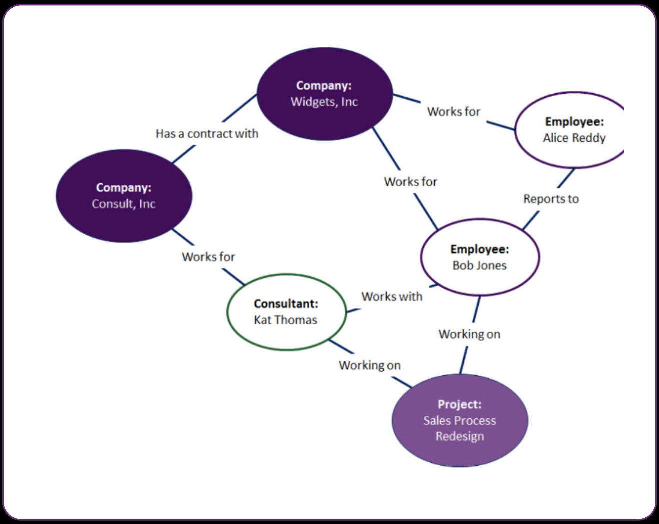
Sandra Young

University of Brighton

Supervisors: Dr Roger Evans, Dr Gulden Uchyigit

Ontologies

Image from: https://towardsdatascience.com/ontology-and-data-science-45e916288cc5 accessed on 13/02/2020



- *Yusof, Norlia et al. “Ontology modeling of Malaysian food composition.” 2016 Third International Conference on Information Retrieval and Knowledge Management (CAMP) (2016): 149-154.*

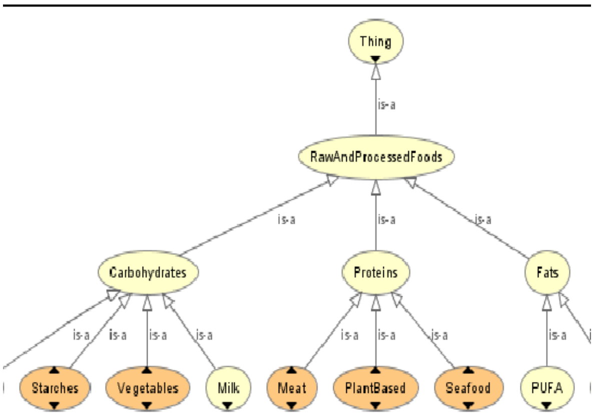


Figure 5. The class hierarchy of raw and processed food.

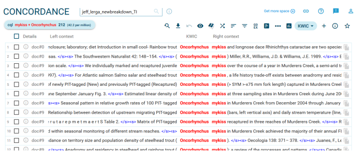
Problem

Ontologies are really useful

However, question as to accurate integration of data

Ontologies by nature have to exclude at some point (explicit conceptualisation), don’t necessarily take into account how presented in the data

Biodiversity and scientific nomenclature and why identified



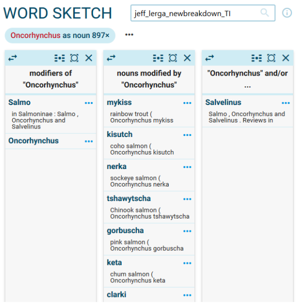
CONCORDANCE jeff_lerga_newbreakdown_TI

Get more space

cql mykiss + Oncorhynchus 212 (42.2 per million)

KWIC

		Details	Left context	KWIC		Right context
1	□	doc#0	nclosure; laboratory; diet Introduction in small cool- Rainbow trout	Oncorhynchus	mykiss	and longnose dace Rhinichthys cataractae are two species
2	□	doc#0	sas. </s><s> The Southwestern Naturalist 42: 148–154. </s><s> (	Oncorhynchus	mykiss	) Miller, R.R., Williams, J.D. & Williams, J.E. 1989. </s><s> E
3	□	doc#9	ion scale. </s><s> We individually marked and recaptured juvenile	Oncorhynchus	mykiss	over the course of a year in Murderers Creek, a semi-arid tr
4	□	doc#9	997). </s><s> For Atlantic salmon Salmo salar and steelhead trout	Oncorhynchus	mykiss	, a life history trade-off exists between anadromy and resid
5	□	doc#9	of newly PIT-tagged (New) and previously PIT-tagged (Recaptured)	Oncorhynchus	mykiss	(> SYM >-x75 mm fork length) captured in Murderers Creel
6	□	doc#9	une September January Fig. 3. </s><s> Estimated linear density of	Oncorhynchus	mykiss	at three sampling sites in Murderers Creek during June 20
7	□	doc#9	s><s> Seasonal pattern in relative growth rates of 100 PIT- tagged	Oncorhynchus	mykiss	in Murderers Creek from December 2004 through January
8	□	doc#9	Relationship between detection of upstream migrating PIT-tagged	Oncorhynchus	mykiss	(bars, left vertical axis) and daily stream temperature (line,
9	□	doc#9	e r u t a r e p m e t m a e r t S Table 2. </s><s> Matrix of PIT-tagged	Oncorhynchus	mykiss	recaptured in three reaches of Murderers Creek. </s><s> R
10	□	doc#9	d within seasonal monitoring of different stream reaches. </s><s>	Oncorhynchus	mykiss	in Murderers Creek achieved the majority of their annual Fl
11	□	doc#9	idance on territory size and population density of steelhead trout (	Oncorhynchus	mykiss	). </s><s> Oecologia 138: 371– 378. </s><s> Juanes, F., Le
12	□	doc#9	</s><s> Anadromy and residency in steelhead and rainbow trout (	Oncorhynchus	mykiss	): a review of the processes and patterns. </s><s> Canadia



WORD SKETCH jeff_lerga_newbreakdown_TI

Oncorhynchus as noun 897× •••

modifiers of "Oncorhynchus"

Salmo •••
in Salmoninae : Salmo , Oncorhynchus and Salvelinus

Oncorhynchus •••

nouns modified by "Oncorhynchus"

mykiss •••
rainbow trout (Oncorhynchus mykiss

kisutch •••
coho salmon (Oncorhynchus kisutch

nerka •••
sockeye salmon (Oncorhynchus nerka

tshawytscha •••
Chinook salmon (Oncorhynchus tshawytscha

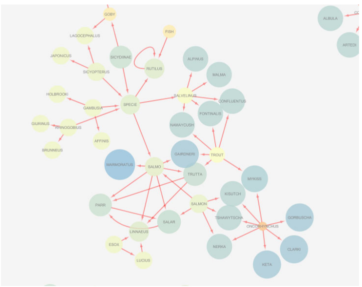
gorbuscha •••
pink salmon (Oncorhynchus gorbuscha

keta •••
chum salmon (Oncorhynchus keta

clarki •••

"Oncorhynchus" and/or ...

Salvelinus •••
Salmo , Oncorhynchus and Salvelinus . Reviews in



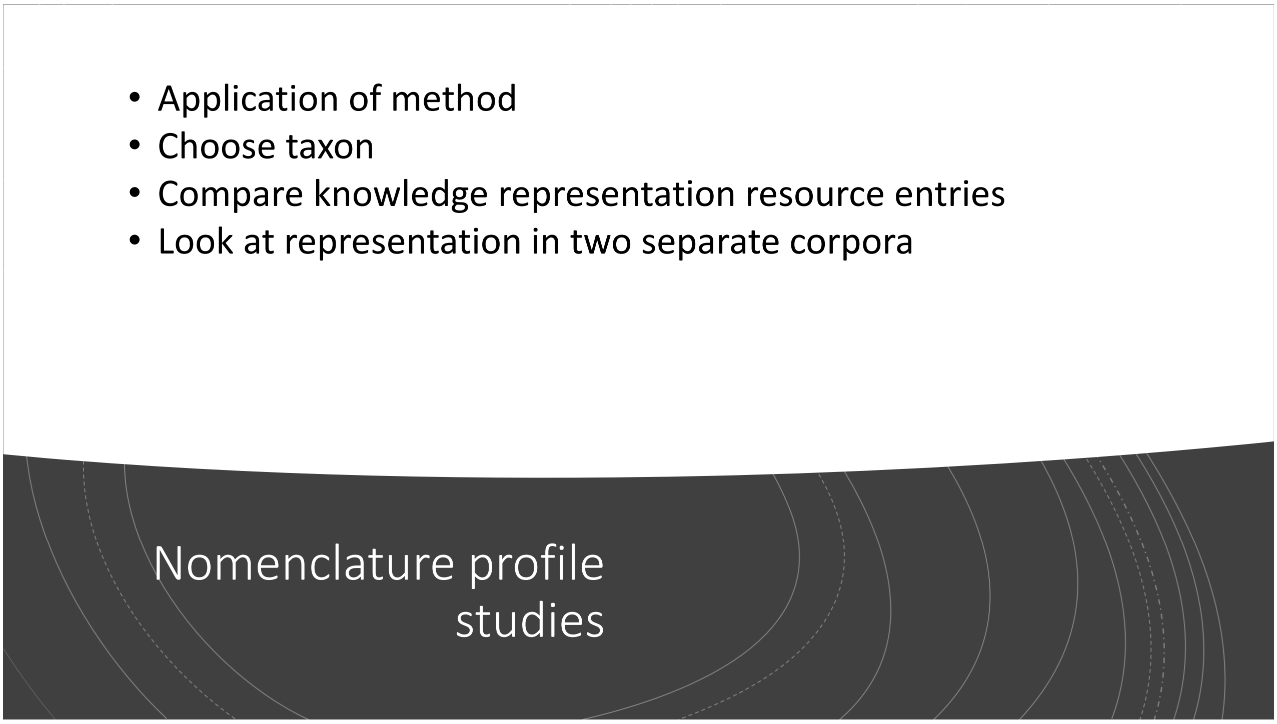
My approach

Plan for today

- Introductions
- Nomenclature Profile Studies: an introduction
- Hierarchy identification
- Knowledge representation: your experiences and my analysis
- Nomenclature and vernacular usage: discussion
- Corpus data: analyses and discussion

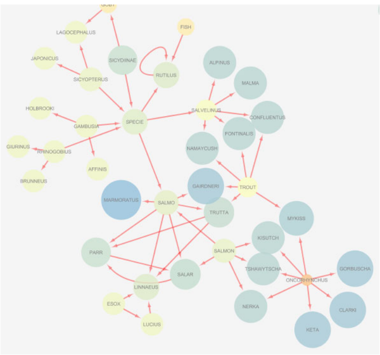
Plan for today

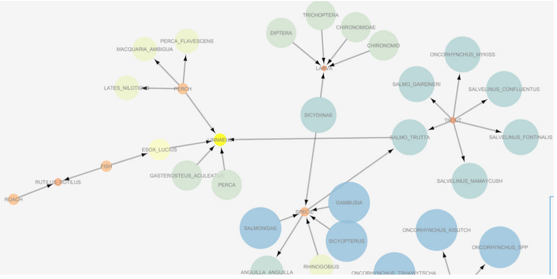
- Introductions
- Nomenclature Profile Studies: an introduction
- Hierarchy identification
- Knowledge representation: your experiences and my analysis
- Nomenclature and vernacular usage: discussion
- Corpus data: analyses and discussion



- Application of method
- Choose taxon
- Compare knowledge representation resource entries
- Look at representation in two separate corpora

Nomenclature profile studies





Hierarchy identification

- Left image of original corpus (no extra processing)
- Right image of unified corpus (multi-word names unified)

Plan for today

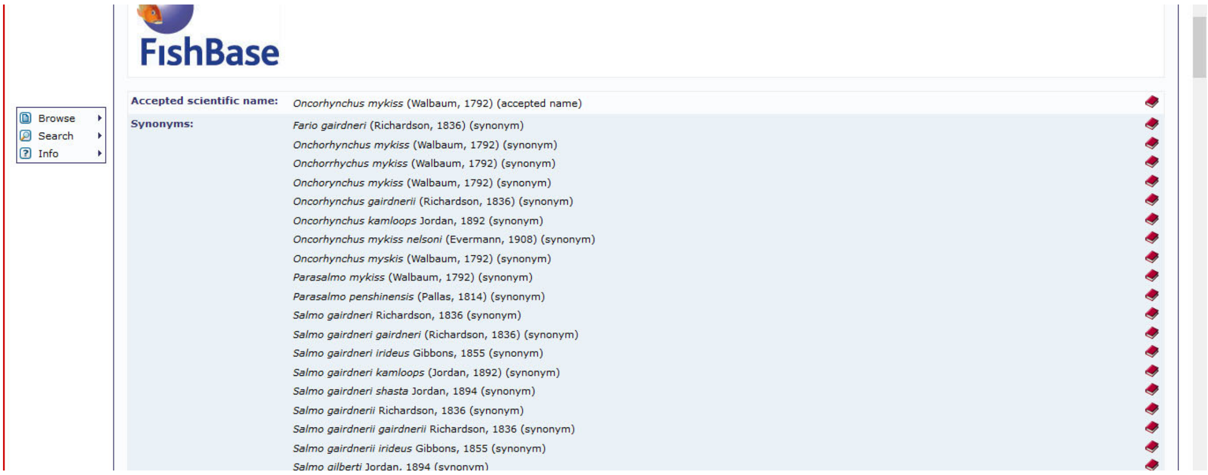
- Introductions
- Nomenclature Profile Studies: an introduction
- Hierarchy identification
- Knowledge representation: your experiences and my analysis
- Nomenclature and vernacular usage: discussion
- Corpus data: analyses and discussion

Knowledge representation resources

VERTEBRATE TAXONOMY ONTOLOGY

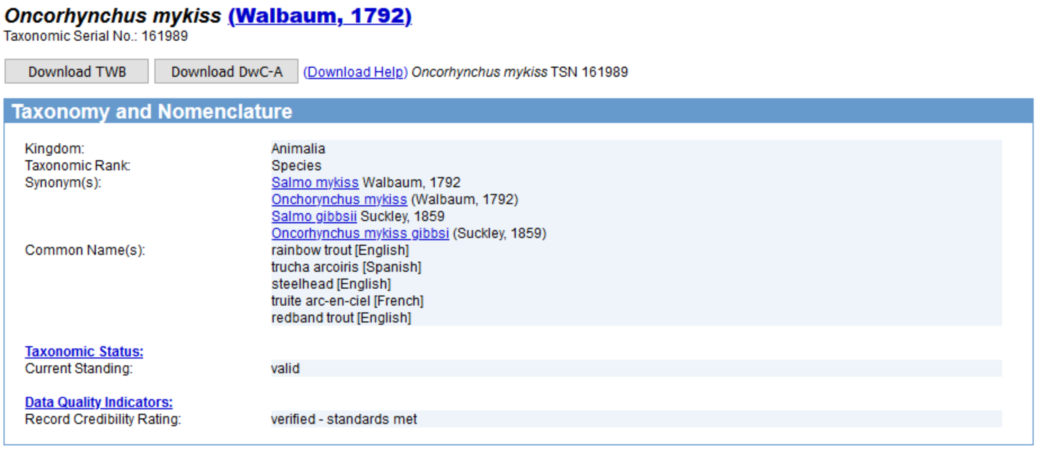
INTEGRATED TAXONOMIC INFORMATION SYSTEM

CATALOGUE OF LIFE



Accepted scientific name:	*Oncorhynchus mykiss* (Walbaum, 1792) (accepted name)
Synonyms:	*Fario gairdneri* (Richardson, 1836) (synonym)
	Onchorhynchus mykiss (Walbaum, 1792) (synonym)
	Onchorrhychus mykiss (Walbaum, 1792) (synonym)
	Onchorynchus mykiss (Walbaum, 1792) (synonym)
	Oncorhynchus gairdnerii (Richardson, 1836) (synonym)
	Oncorhynchus kamloops Jordan, 1892 (synonym)
	Oncorhynchus mykiss nelsoni (Evermann, 1908) (synonym)
	Oncorhynchus myskis (Walbaum, 1792) (synonym)
	Parasalmo mykiss (Walbaum, 1792) (synonym)
	Parasalmo penshinensis (Pallas, 1814) (synonym)
	Salmo gairdneri Richardson, 1836 (synonym)
	Salmo gairdneri gairdneri (Richardson, 1836) (synonym)
	Salmo gairdneri irideus Gibbons, 1855 (synonym)
	Salmo gairdneri kamloops (Jordan, 1892) (synonym)
	Salmo gairdneri shasta Jordan, 1894 (synonym)
	Salmo gairdnerii Richardson, 1836 (synonym)
	Salmo gairdnerii gairdnerii Richardson, 1836 (synonym)
	Salmo gairdnerii irideus Gibbons, 1855 (synonym)
	Salmo gilberti Jordan, 1894 (synonym)

Oncorhynchus mykiss (CoL)



***Oncorhynchus mykiss* (Walbaum, 1792)**

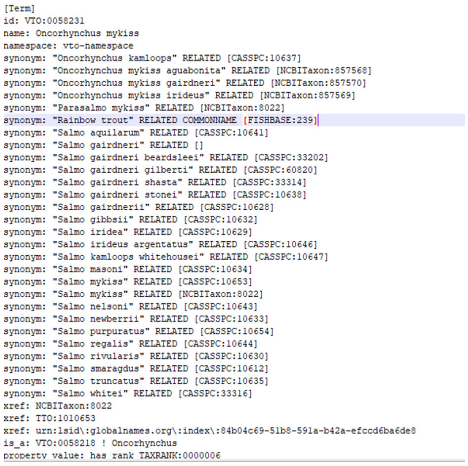
Taxonomic Serial No.: 161989

(Download Help) *Oncorhynchus mykiss* TSN 161989

Taxonomy and Nomenclature

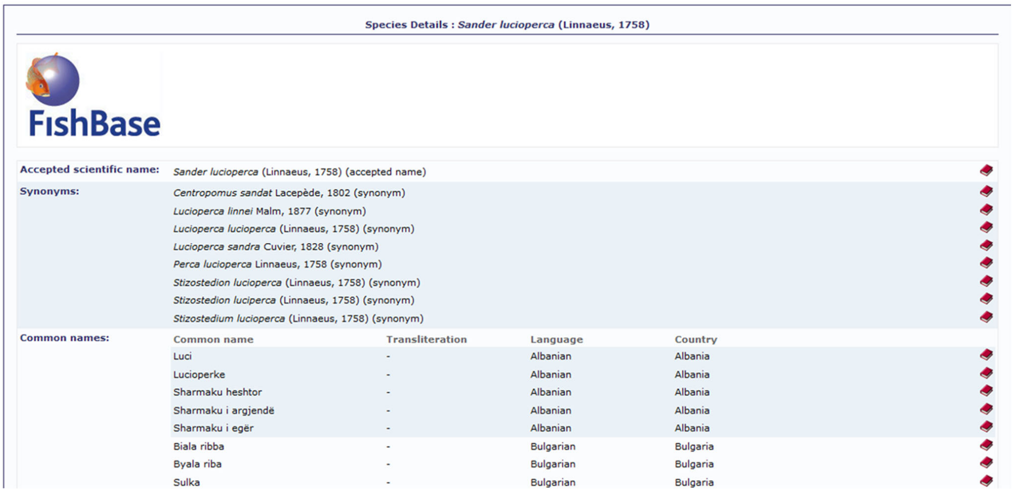
Kingdom:	Animalia
Taxonomic Rank:	Species
Synonym(s):	Salmo mykiss Walbaum, 1792 Onchorynchus mykiss (Walbaum, 1792) Salmo gibbsii Suckley, 1859 Oncorhynchus mykiss gibbsi (Suckley, 1859)
Common Name(s):	rainbow trout [English] trucha arcoiris [Spanish] steelhead [English] truite arc-en-ciel [French] redband trout [English]
Taxonomic Status:	
Current Standing:	valid
Data Quality Indicators:	
Record Credibility Rating:	verified - standards met

Oncorhynchus mykiss (ITIS)



```
[Term]
id: VTO:0058231
name: Oncorhynchus mykiss
namespace: vto-namespace
synonym: "Oncorhynchus kamloops" RELATED [CASSPC:10637]
synonym: "Oncorhynchus mykiss aguabonita" RELATED [NCBITaxon:857568]
synonym: "Oncorhynchus mykiss gairdneri" RELATED [NCBITaxon:857570]
synonym: "Oncorhynchus mykiss irideus" RELATED [NCBITaxon:857569]
synonym: "Parasalmo mykiss" RELATED [NCBITaxon:8022]
synonym: "Rainbow trout" RELATED COMMONNAME [FISHBASE:239]
synonym: "Salmo aquilarum" RELATED [CASSPC:10641]
synonym: "Salmo gairdneri" RELATED []
synonym: "Salmo gairdneri beardsleei" RELATED [CASSPC:33202]
synonym: "Salmo gairdneri gilberti" RELATED [CASSPC:60820]
synonym: "Salmo gairdneri shasta" RELATED [CASSPC:33314]
synonym: "Salmo gairdneri stonei" RELATED [CASSPC:10638]
synonym: "Salmo gairdnerii" RELATED [CASSPC:10628]
synonym: "Salmo gibbsii" RELATED [CASSPC:10632]
synonym: "Salmo iridea" RELATED [CASSPC:10629]
synonym: "Salmo irideus argentatus" RELATED [CASSPC:10646]
synonym: "Salmo kamloops whitehousei" RELATED [CASSPC:10647]
synonym: "Salmo masoni" RELATED [CASSPC:10634]
synonym: "Salmo mykiss" RELATED [CASSPC:10653]
synonym: "Salmo mykiss" RELATED [NCBITaxon:8022]
synonym: "Salmo nelsoni" RELATED [CASSPC:10643]
synonym: "Salmo newberrii" RELATED [CASSPC:10633]
synonym: "Salmo purpuratus" RELATED [CASSPC:10654]
synonym: "Salmo regalis" RELATED [CASSPC:10644]
synonym: "Salmo rivularis" RELATED [CASSPC:10630]
synonym: "Salmo smaragdus" RELATED [CASSPC:10612]
synonym: "Salmo truncatus" RELATED [CASSPC:10635]
synonym: "Salmo whitei" RELATED [CASSPC:33316]
xref: NCBITaxon:8022
xref: TTO:1010653
xref: urn:lsid\:globalnames.org\:index\:84b04c69-51b8-591a-b42a-efccd6ba6de8
is_a: VTO:0058218 ! Oncorhynchus
property_value: has_rank TAXRANK:0000006
```

Oncorhynchus mykiss (VTO)



Species Details : *Sander lucioperca* (Linnaeus, 1758)

FishBase

Accepted scientific name: *Sander lucioperca* (Linnaeus, 1758) (accepted name)

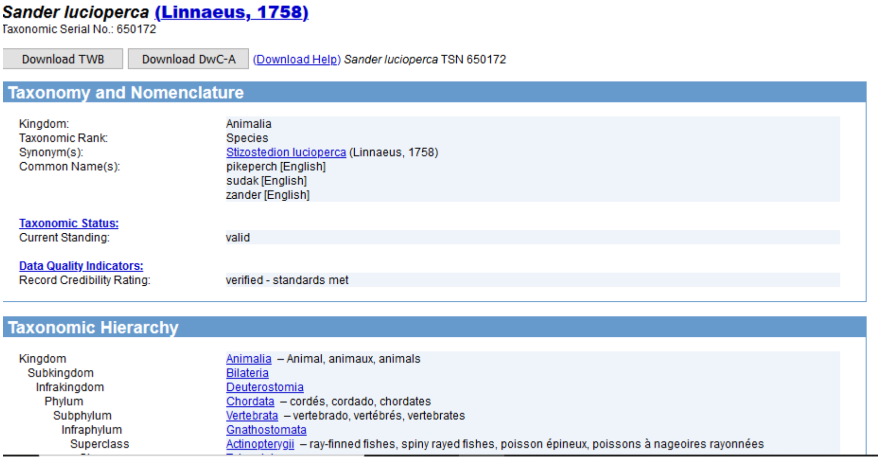
Synonyms:

Centropomus sandat Lacepède, 1802 (synonym)
Lucioperca linnei Malm, 1877 (synonym)
Lucioperca lucioperca (Linnaeus, 1758) (synonym)
Lucioperca sandra Cuvier, 1828 (synonym)
Perca lucioperca Linnaeus, 1758 (synonym)
Stizostedion lucioperca (Linnaeus, 1758) (synonym)
Stizostedion luciperca (Linnaeus, 1758) (synonym)
Stizostedium lucioperca (Linnaeus, 1758) (synonym)

Common names:

Common name	Transliteration	Language	Country
Luci	-	Albanian	Albania
Lucioperke	-	Albanian	Albania
Sharmaku heshtor	-	Albanian	Albania
Sharmaku i argjendë	-	Albanian	Albania
Sharmaku i egër	-	Albanian	Albania
Biala ribba	-	Bulgarian	Bulgaria
Byala riba	-	Bulgarian	Bulgaria
Sulka	-	Bulgarian	Bulgaria

Sander lucioperca (CoL)



***Sander lucioperca* (Linnaeus, 1758)**
Taxonomic Serial No.: 650172

(Download Help) *Sander lucioperca* TSN 650172

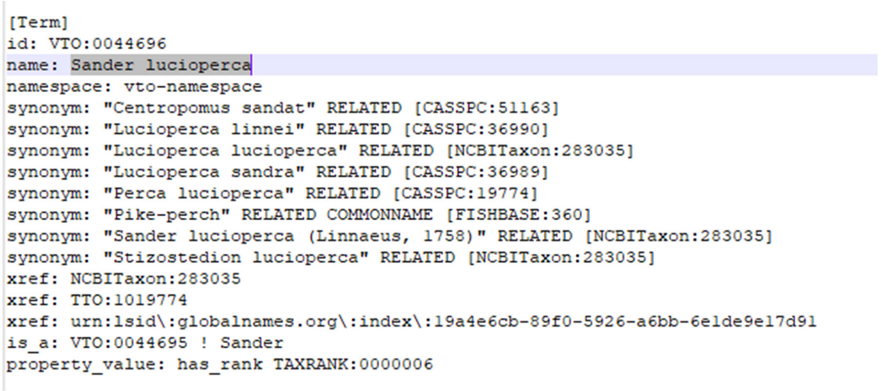
Taxonomy and Nomenclature

Kingdom:	Animalia
Taxonomic Rank:	Species
Synonym(s):	Stizostedion lucioperca (Linnaeus, 1758)
Common Name(s):	pikeperch [English] sudak [English] zander [English]
Taxonomic Status:	
Current Standing:	valid
Data Quality Indicators:	
Record Credibility Rating:	verified - standards met

Taxonomic Hierarchy

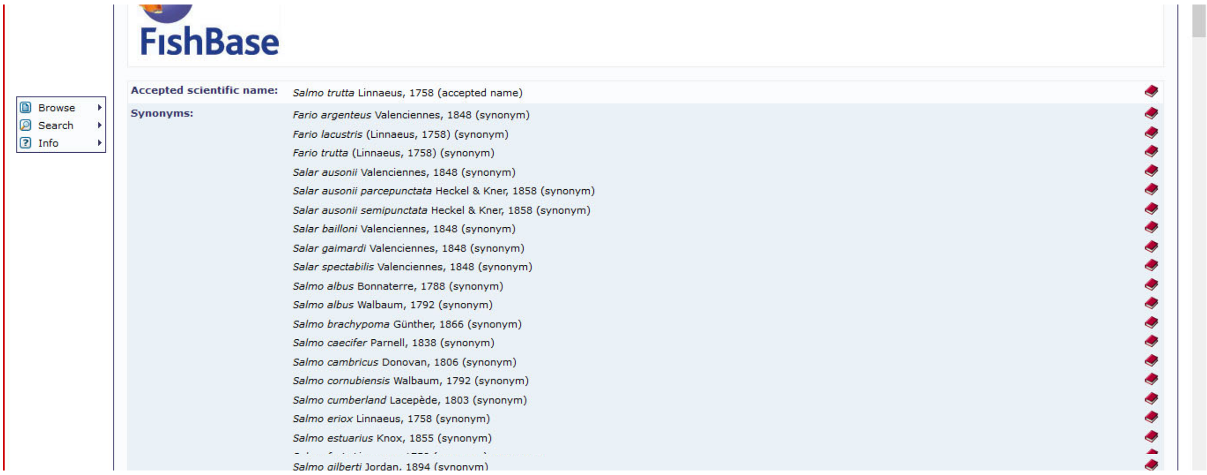
Kingdom	Animalia – Animal, animaux, animals
Subkingdom	Bilateria
Infrakingdom	Deuterostomia
Phylum	Chordata – cordés, cordado, chordates
Subphylum	Vertebrata – vertebrado, vertébrés, vertebrates
Infraphylum	Gnathostomata
Superclass	Actinopterygii – ray-finned fishes, spiny rayed fishes, poisson épineux, poissons à nageoires rayonnées

Sander lucioperca (ITIS)



```
[Term]
id: VTO:0044696
name: Sander lucioperca
namespace: vto-namespace
synonym: "Centropomus sandat" RELATED [CASSPC:51163]
synonym: "Lucioperca linnei" RELATED [CASSPC:36990]
synonym: "Lucioperca lucioperca" RELATED [NCBITaxon:283035]
synonym: "Lucioperca sandra" RELATED [CASSPC:36989]
synonym: "Perca lucioperca" RELATED [CASSPC:19774]
synonym: "Pike-perch" RELATED COMMONNAME [FISHBASE:360]
synonym: "Sander lucioperca (Linnaeus, 1758)" RELATED [NCBITaxon:283035]
synonym: "Stizostedion lucioperca" RELATED [NCBITaxon:283035]
xref: NCBITaxon:283035
xref: TTO:1019774
xref: urn:lsid\:globalnames.org\:index\:19a4e6cb-89f0-5926-a6bb-6e1de9e17d91
is_a: VTO:0044695 ! Sander
property_value: has_rank TAXRANK:0000006
```

Sander lucioperca (VTO)

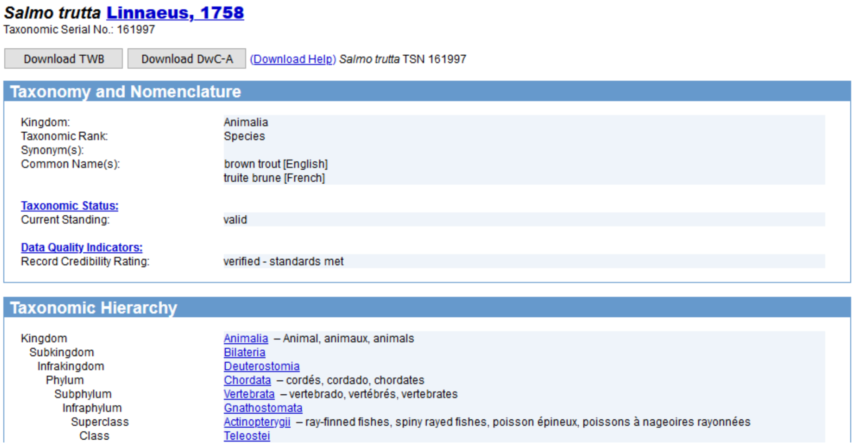


Accepted scientific name: *Salmo trutta* Linnaeus, 1758 (accepted name)

Synonyms:

Fario argenteus Valenciennes, 1848 (synonym)
Fario lacustris (Linnaeus, 1758) (synonym)
Fario trutta (Linnaeus, 1758) (synonym)
Salar ausonii Valenciennes, 1848 (synonym)
Salar ausonii parcepunctata Heckel & Kner, 1858 (synonym)
Salar ausonii semipunctata Heckel & Kner, 1858 (synonym)
Salar bailloni Valenciennes, 1848 (synonym)
Salar gaimardi Valenciennes, 1848 (synonym)
Salar spectabilis Valenciennes, 1848 (synonym)
Salmo albus Bonnaterre, 1788 (synonym)
Salmo albus Walbaum, 1792 (synonym)
Salmo brachypoma Günther, 1866 (synonym)
Salmo caecifer Parnell, 1838 (synonym)
Salmo cambricus Donovan, 1806 (synonym)
Salmo cornubiensis Walbaum, 1792 (synonym)
Salmo cumberland Lacepède, 1803 (synonym)
Salmo eriox Linnaeus, 1758 (synonym)
Salmo estuarius Knox, 1855 (synonym)
[illegible]
Salmo gilberti Jordan, 1894 (synonym)

Salmo trutta (CoL)



Salmo trutta **Linnaeus, 1758**
Taxonomic Serial No.: 161997

(Download Help) *Salmo trutta* TSN 161997

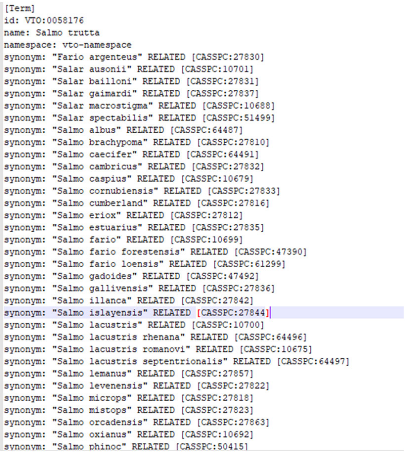
Taxonomy and Nomenclature

Kingdom:	Animalia
Taxonomic Rank:	Species
Synonym(s):	
Common Name(s):	brown trout [English] truite brune [French]
Taxonomic Status:	
Current Standing:	valid
Data Quality Indicators:	
Record Credibility Rating:	verified - standards met

Taxonomic Hierarchy

Kingdom	Animalia – Animal, animaux, animals
Subkingdom	Bilateria
Infrakingdom	Deuterostomia
Phylum	Chordata – cordés, cordado, chordates
Subphylum	Vertebrata – vertebrado, vertébrés, vertebrates
Infraphylum	Gnathostomata
Superclass	Actinopterygii – ray-finned fishes, spiny rayed fishes, poisson épineux, poissons à nageoires rayonnées
Class	Teleostei

Salmo trutta (ITIS)



```
[Term]
id: VTO:0058176
name: Salmo trutta
namespace: vto-namespace
synonym: "Fario argenteus" RELATED [CASSPC:27830]
synonym: "Salar ausonii" RELATED [CASSPC:10701]
synonym: "Salar bailloni" RELATED [CASSPC:27831]
synonym: "Salar gaimardi" RELATED [CASSPC:27837]
synonym: "Salar macrostigma" RELATED [CASSPC:10688]
synonym: "Salar spectabilis" RELATED [CASSPC:51499]
synonym: "Salmo albus" RELATED [CASSPC:64487]
synonym: "Salmo brachypoma" RELATED [CASSPC:27810]
synonym: "Salmo caecifer" RELATED [CASSPC:64491]
synonym: "Salmo cambricus" RELATED [CASSPC:27832]
synonym: "Salmo caspius" RELATED [CASSPC:10679]
synonym: "Salmo cornubiensis" RELATED [CASSPC:27833]
synonym: "Salmo cumberland" RELATED [CASSPC:27816]
synonym: "Salmo eriox" RELATED [CASSPC:27812]
synonym: "Salmo estuarius" RELATED [CASSPC:27835]
synonym: "Salmo fario" RELATED [CASSPC:10699]
synonym: "Salmo fario forestensis" RELATED [CASSPC:47390]
synonym: "Salmo fario loensis" RELATED [CASSPC:61299]
synonym: "Salmo gadoides" RELATED [CASSPC:47492]
synonym: "Salmo gallivensis" RELATED [CASSPC:27836]
synonym: "Salmo illanca" RELATED [CASSPC:27842]
synonym: "Salmo islayensis" RELATED [CASSPC:27844]
synonym: "Salmo lacustris" RELATED [CASSPC:10700]
synonym: "Salmo lacustris rhenana" RELATED [CASSPC:64496]
synonym: "Salmo lacustris romanovi" RELATED [CASSPC:10675]
synonym: "Salmo lacustris septentrionalis" RELATED [CASSPC:64497]
synonym: "Salmo lemanus" RELATED [CASSPC:27857]
synonym: "Salmo levenensis" RELATED [CASSPC:27822]
synonym: "Salmo microps" RELATED [CASSPC:27818]
synonym: "Salmo mistops" RELATED [CASSPC:27823]
synonym: "Salmo orcadensis" RELATED [CASSPC:27863]
synonym: "Salmo oxianus" RELATED [CASSPC:10692]
synonym: "Salmo phinoc" RELATED [CASSPC:50415]
```

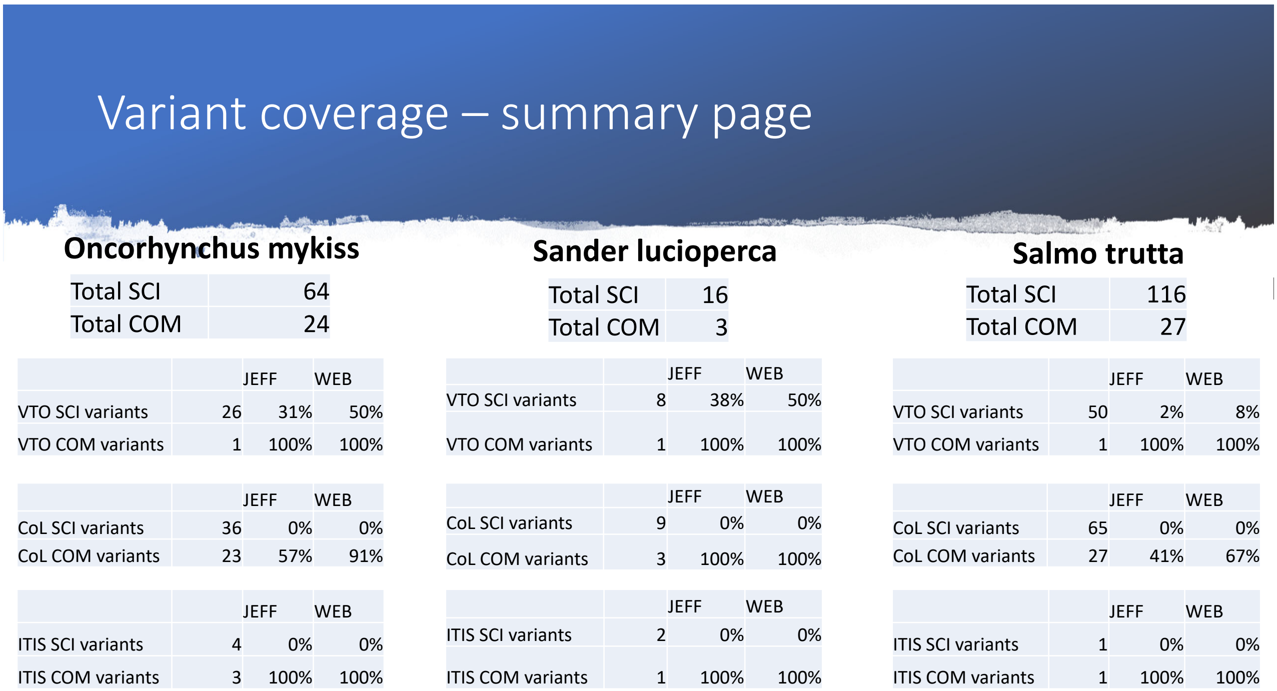
Salmo trutta (VTO)

Plan for today

- Introductions
- Nomenclature Profile Studies: an introduction
- Hierarchy identification
- Knowledge representation: your experiences and my analysis
- Nomenclature and vernacular usage: discussion
- Corpus data: analyses and discussion

Plan for today

- Introductions
- Nomenclature Profile Studies: an introduction
- Hierarchy identification
- Knowledge representation: your experiences and my analysis
- Nomenclature and vernacular usage: discussion
- Corpus data: analyses and discussion



Variant coverage – summary page

Oncorhynchus mykiss

Total SCI	64
Total COM	24

		JEFF	WEB
VTO SCI variants	26	31%	50%
VTO COM variants	1	100%	100%

		JEFF	WEB
CoL SCI variants	36	0%	0%
CoL COM variants	23	57%	91%

		JEFF	WEB
ITIS SCI variants	4	0%	0%
ITIS COM variants	3	100%	100%

Sander lucioperca

Total SCI	16
Total COM	3

		JEFF	WEB
VTO SCI variants	8	38%	50%
VTO COM variants	1	100%	100%

		JEFF	WEB
CoL SCI variants	9	0%	0%
CoL COM variants	3	100%	100%

		JEFF	WEB
ITIS SCI variants	2	0%	0%
ITIS COM variants	1	100%	100%

Salmo trutta

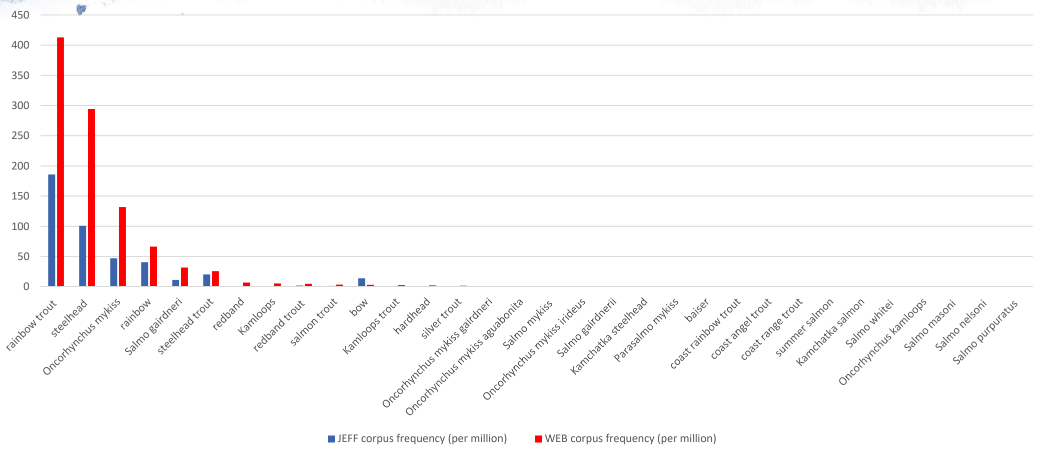
Total SCI	116
Total COM	27

		JEFF	WEB
VTO SCI variants	50	2%	8%
VTO COM variants	1	100%	100%

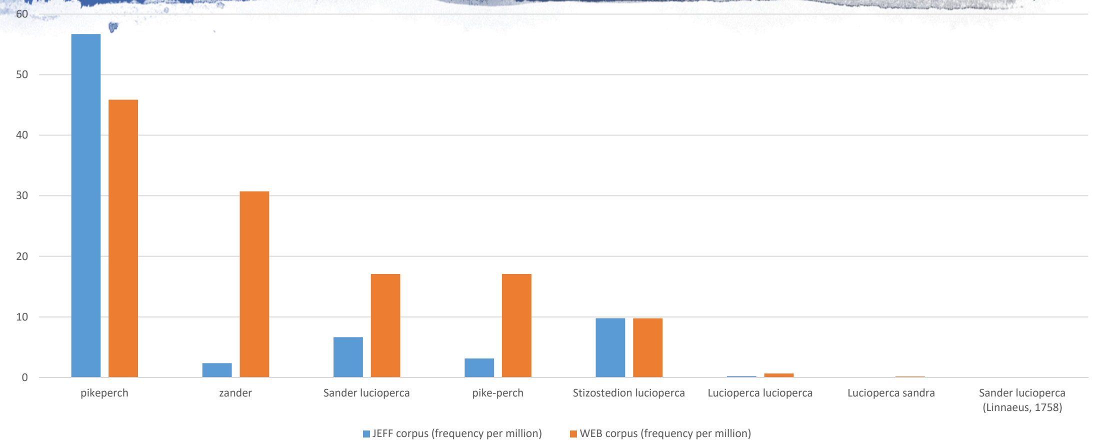
		JEFF	WEB
CoL SCI variants	65	0%	0%
CoL COM variants	27	41%	67%

		JEFF	WEB
ITIS SCI variants	1	0%	0%
ITIS COM variants	1	100%	100%

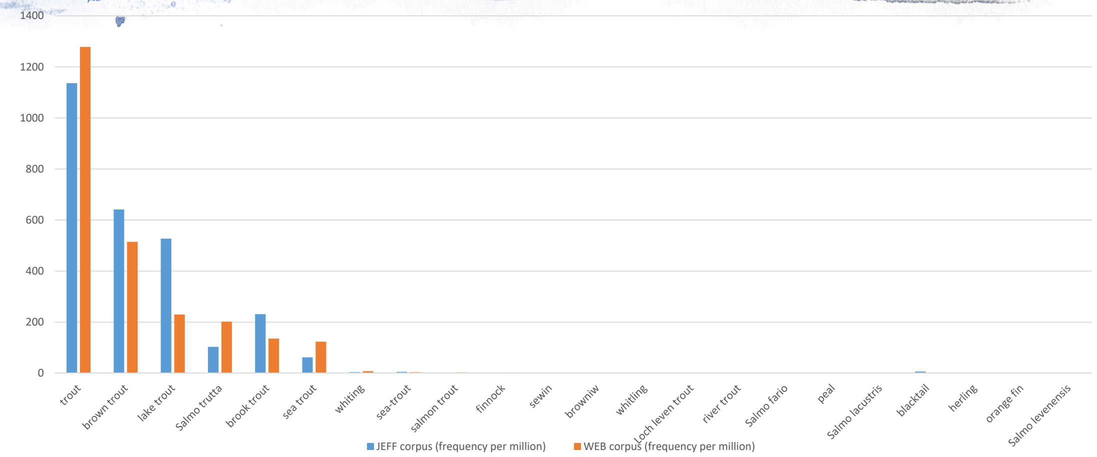
Variant variation – Oncorhynchus mykiss



Variant variation – Sander lucioperca



Variant variation – Salmo trutta



Ambiguities identified

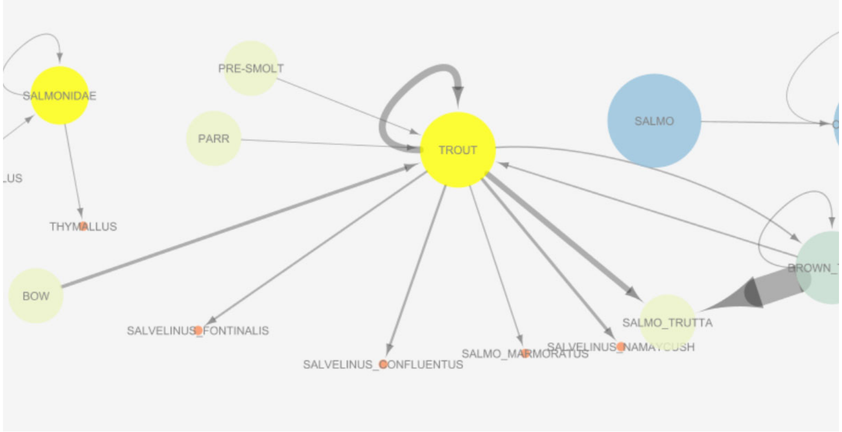
Vernacular

Spelling

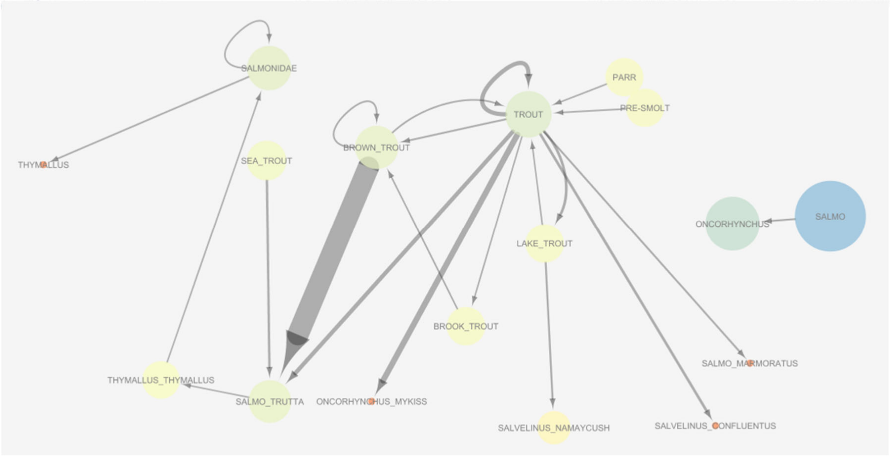
Accepted term versus most used

Contradictions between test data and knowledge resources

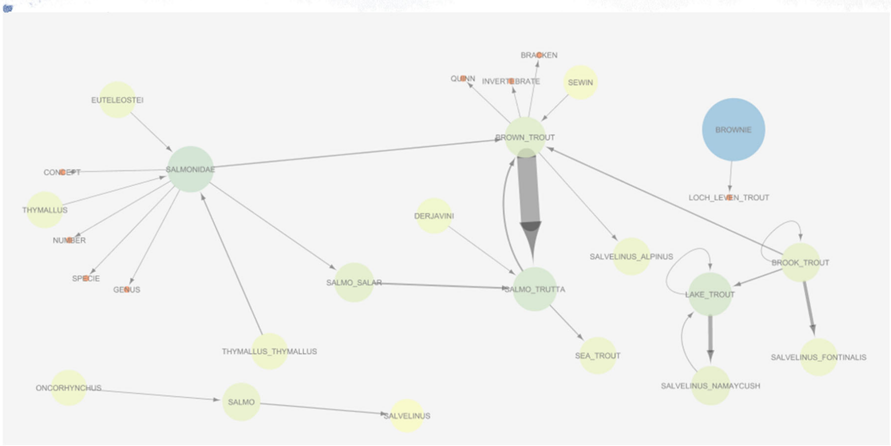
Vernacular ambiguities: broader meaning



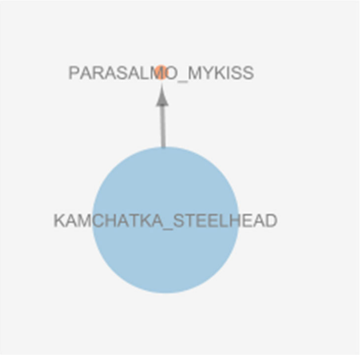
Vernacular ambiguities: broader or narrower meaning

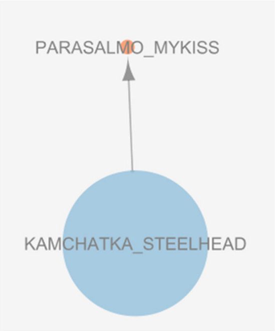


Vernacular ambiguities: broader or narrower meaning

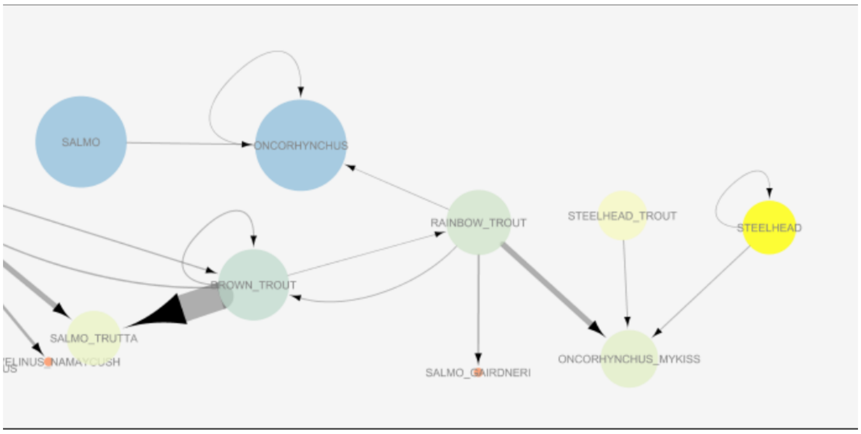


Ambiguities: increased specificity of usage

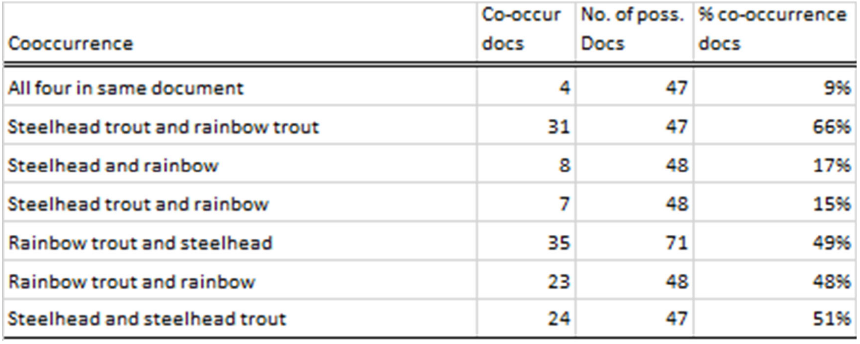




Ambiguities: increased specificity of usage (JEFF)

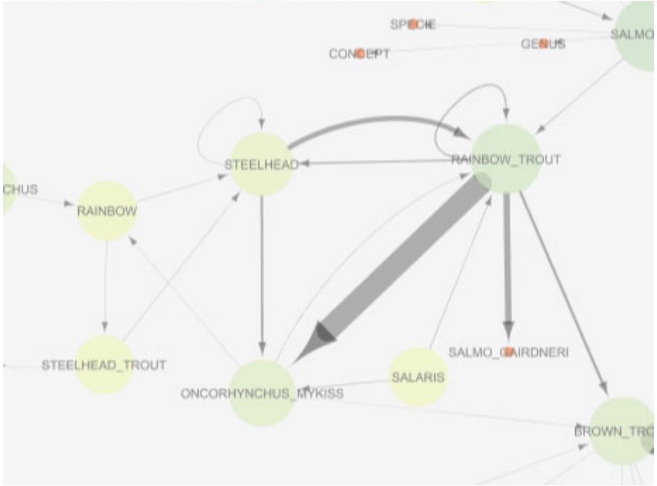


Ambiguities: increased specificity of usage (JEFF)



Cooccurrence	Co-occur docs	No. of poss. Docs	% co-occurrence docs
All four in same document	4	47	9%
Steelhead trout and rainbow trout	31	47	66%
Steelhead and rainbow	8	48	17%
Steelhead trout and rainbow	7	48	15%
Rainbow trout and steelhead	35	71	49%
Rainbow trout and rainbow	23	48	48%
Steelhead and steelhead trout	24	47	51%

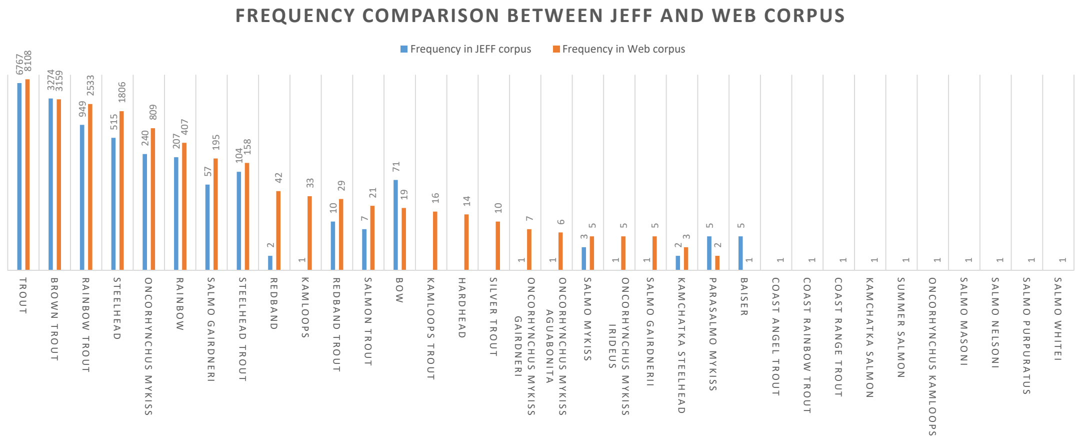
Ambiguities: increased specificity of usage (WEB)



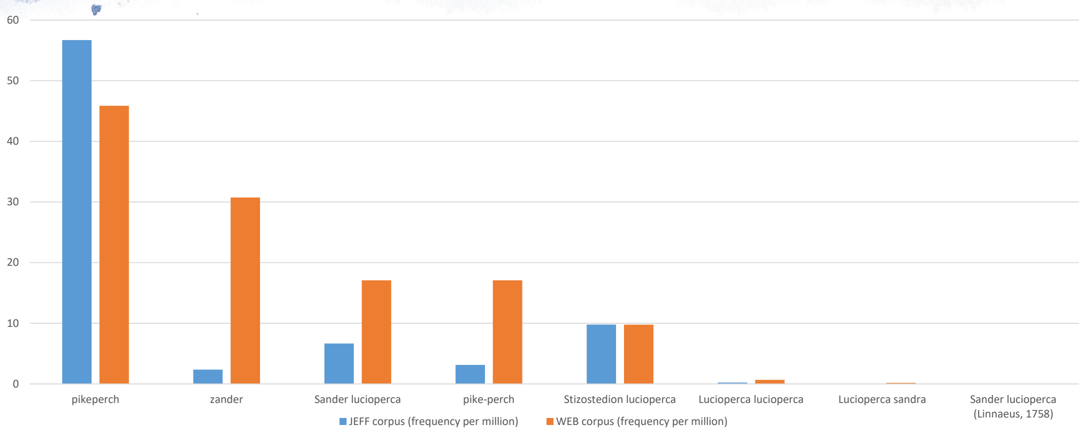
Ambiguities: increased specificity of usage (WEB)

Co-occurrence	Co-occur docs	No. of poss. Docs	% co-occurrence docs
All four in same document	15	72	21%
Steelhead trout and rainbow trout	57	72	79%
Steelhead and rainbow	41	120	34%
Steelhead trout and rainbow	27	120	23%
Rainbow trout and steelhead	101	120	84%
Rainbow trout and rainbow	96	139	69%
Steelhead and steelhead trout	48	72	67%

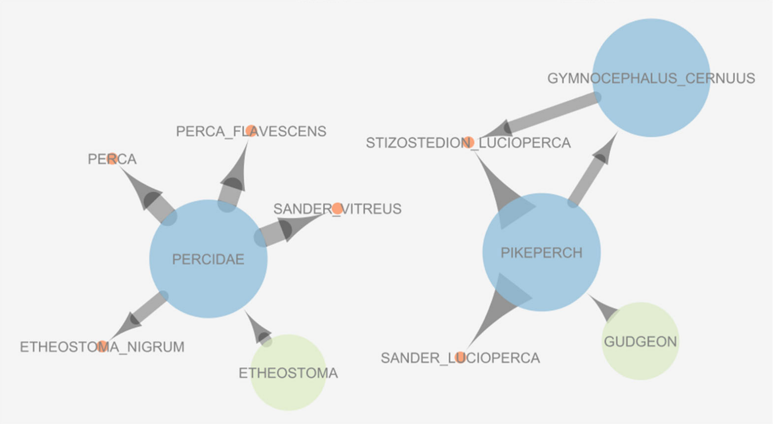
Ambiguities: nomenclature spelling



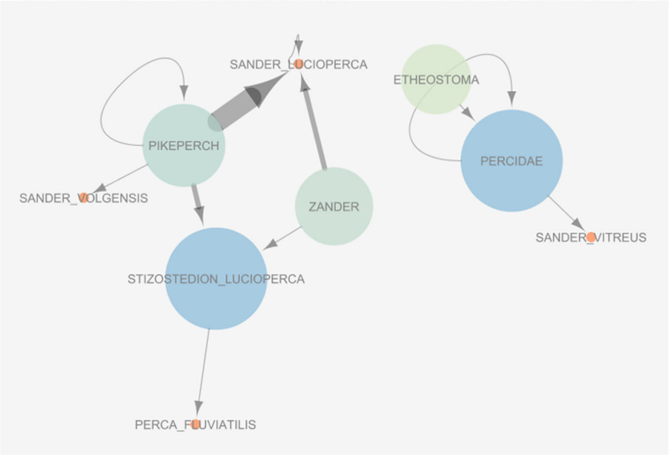
Ambiguities: accepted name versus variant usage



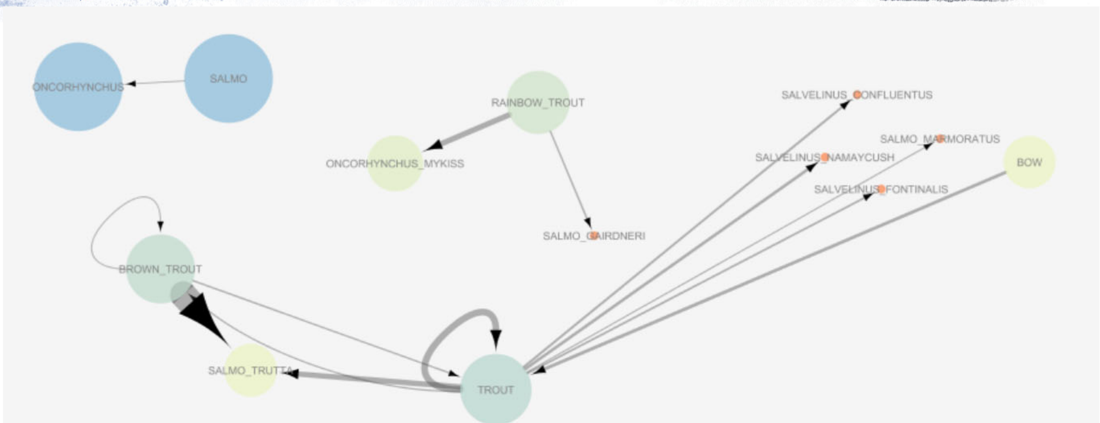
Ambiguities: accepted name versus variant usage



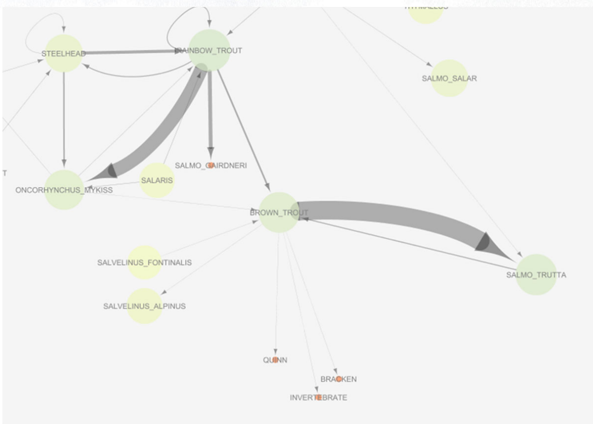
Ambiguities: accepted name versus variant usage



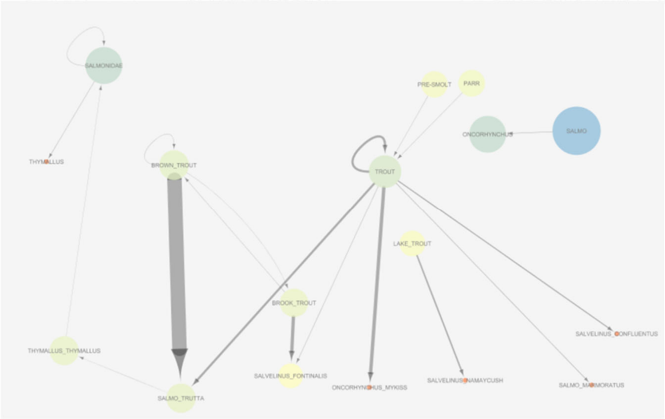
Contradiction with knowledge resource (JEFF)



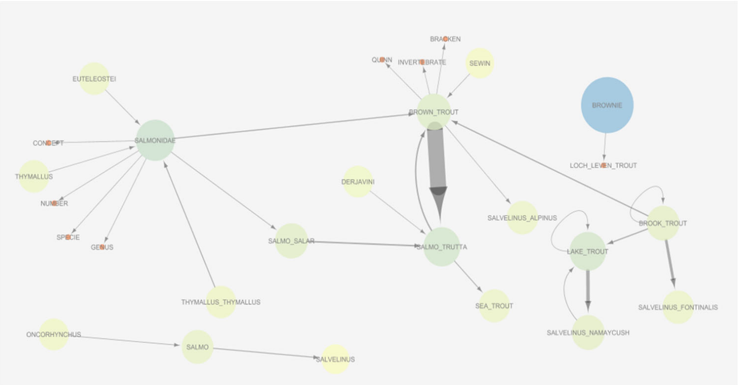
Contradiction with knowledge resource (WEB)

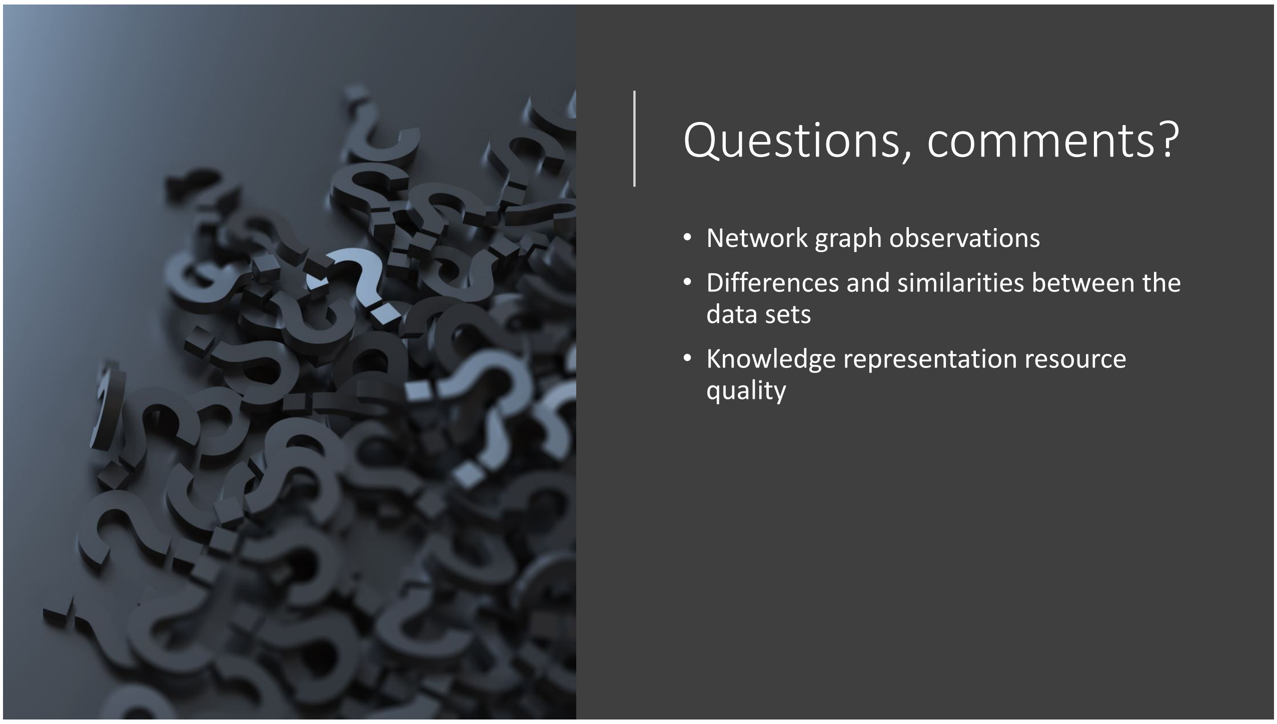


Contradiction with knowledge resource (JEFF)



Contradiction with knowledge resource (WEB)

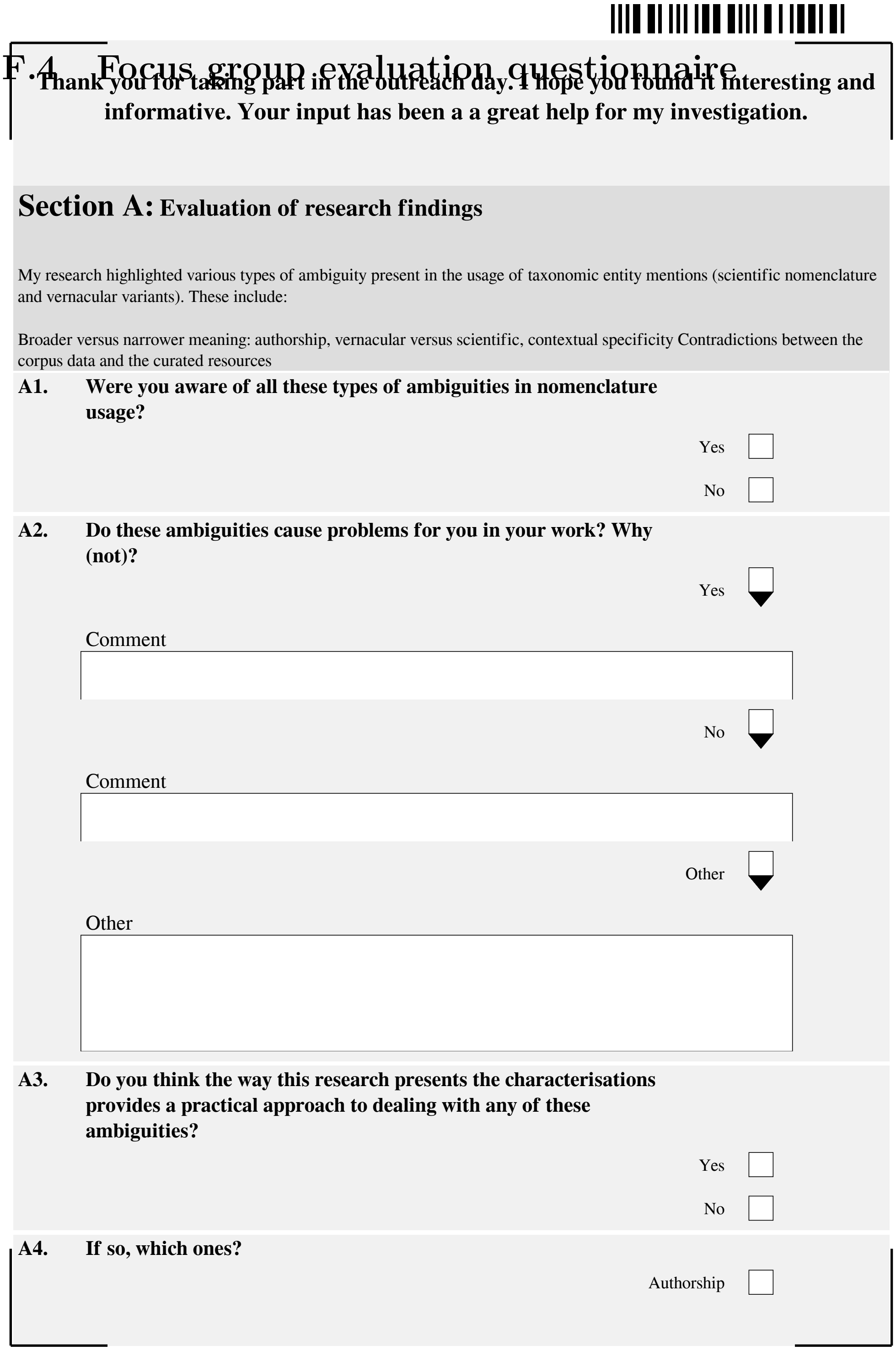




Questions, comments?

- Network graph observations
- Differences and similarities between the data sets
- Knowledge representation resource quality

F.4 Focus group evaluation questionnaire



Thank you for taking part in the outreach day. I hope you found it interesting and informative. Your input has been a a great help for my investigation.

Section A: Evaluation of research findings

My research highlighted various types of ambiguity present in the usage of taxonomic entity mentions (scientific nomenclature and vernacular variants). These include:

Broader versus narrower meaning: authorship, vernacular versus scientific, contextual specificity Contradictions between the corpus data and the curated resources

A1. Were you aware of all these types of ambiguities in nomenclature usage?

Yes □

No □

A2. Do these ambiguities cause problems for you in your work? Why (not)?

Yes □

Comment

No □

Comment

Other □

Other

A3. Do you think the way this research presents the characterisations provides a practical approach to dealing with any of these ambiguities?

Yes □

No □

A4. If so, which ones?

Authorship □

Vernacular versus scientific □

Contextual specificity □

Contradictions between knowledge resources and data sets □

Gaps in knowledge resources □

A5. How useful could this research be for your work (1 = least useful to 5 = most useful)?

1 □

2 □

3 □

4 □

5 □

A6. Can you think of a possible application of this method in your work (such as for checking consistency of usage across data sets, checking for suitable knowledge resources, etc.)

Yes □

No □

Maybe □

A7. If so, what application could you foresee?

Section B: Evaluation of outreach day

B1. What is the most interesting thing you learned today (if anything)?

B2. **How would you rate the outreach day overall? (1 = lowest to 5 = highest)**

1 □

2 □

3 □

4 □

5 □

Thank you for taking the time to fill out this evaluation questionnaire and for taking part in the outreach day.

F.5 Pre-focus group questionnaire results

Please refer to Focus group folder, file name “responses_prefocusgroup.csv”.

F.6 Focus group transcript

SY: Thank you for joining and for agreeing to participate in this outreach day/focus group, particularly with everything going on with the coronavirus. I will start with a bit about me, the purpose of my study. I will explore a bit about the results of the focus group questionnaire. As I am recording can we try to speak one at a time. I have invited you specifically because I wanted people working with scientific nomenclature in similar but different area to get feedback about your opinions about usage.

I am Sandra, I am a PhD student at the University of Brighton. I have been there for about 3.5 years. I'm in CEM. My background is in translation and interpreting, so very much a linguistic background. My study looks at the knowledge representation and integration and looking at the difficulties in integration of knowledge because of ambiguities inherent in natural language. In knowledge integration and interpretation ontologies are a very important data structure, are you all comfortable with ontologies?

P1: Yep, what working definition are you using?

SY: Ontology as formal and explicit representation/conceptualisation of a domain. I have identified that ontologies are very useful but because of the ambiguity of natural language. But problem about this explicit conceptualisation is you can exclude relevant information or inaccurately impose a classification. Identified biodiversity and nomenclature: taxonomy and the way scientific nomenclature is used to describe it. Because of taxonomic format but also because of the hypothetical and changing manners of classification of species. The approach I have taken is an approach that arises from lexicography. Dictionary making. It looks at adapting features that, for example. In lexicography we could do this to analyse language to create dictionary entries – look at large numbers of documents, look at the collocations, so pairs of words that appear in different contexts. Here you can see the different sentences and these are called concordances and the contexts they come out in. Specifically there is a thing called Word Sketch that identified relations between words due to the grammatical relations between them. I have adapted the took to look specifically at links between scientific and vernacular variants and the relations between them. I will explain the graph more in a little bit.

Basically the plan today, first I'd like you to introduce yourselves and your roles and the different ways you tend to interact with scientific nomenclature.

I will then introduce nomenclature profile studies, and then look at hierarchy identification in which I will explain the graphs.

Then we will explore the issues relating to knowledge representation, taxonomies and ontologies and nomenclature usage, exploring further what you responded to in the focus group questionnaire.

Finally ask about my data: how I have analysed it and see if you have any comments.

That OK?

P1: Yes

P3: Yes

P2: Yes

SY: Can you all introduce yourselves?

P1: [redacted] I'm a senior lecturer in ecology and conservation at the University of Brighton. I've been here five years now. In terms of how I use taxonomy and nomenclature its during my teaching so during my day to day teaching but also my research, a lot of which relates specifically to birds, particularly diversity. So, going back to my PhD I was using a global birds' trait dataset, when I was having to deal with differences in different avian taxonomies on a day to day basis. That was certainly a challenge. Which taxonomy to go with, and how to deal with people who split and lump species. But it is ongoing. I still do sometimes take a global approach sometimes, in my research and that does involve having to deal with a lot of bird species, but it's predominantly from bird taxonomy, that's what I tend to use.

P3: I'm the tech lead of the informatics group at the Natural History Museum. I do a lot of open data, so getting the specimen collections online – via the open data portal, which involves joining up the taxonomy from our internal collections management database with GBIF and other systems so we can publish it as open data. Currently I am working on building sort of natural language processers for our historical literature, so automatically joining traits and mining for traits and joining them up with botanical classifications.

P2: [redacted]
So I use taxonomy again, like P1 in my teaching, teaching about species classifications and things. My recent research has been on a single species, on the white rhinosaurus. But even there you have people who want to split species or lump them again. But some of the work before that has been ongoing on amphibian and reptile diversity, where l have been involved in species' assessments and species' descriptions, where you get junior synonyms and undescribed species by other people described that are sat in databases and things like that, where you can get some confusion about whether you are talking about the same thing, unless you specifically use the identifiers provided in some databases. That is my experience of things.

SY: Do you all work a lot with database information? Do you see differences in ambiguity with databases versus narrative text?

P3: Yes, there is. But we get the same problems within the databases that we used within the museum because obviously that's transcribed based on specimen labels. So you get the same narrative problems, with misspellings and redescriptions of names.

SY: As with splitting and lumping: how can you go about identifying, do you try to get back to which taxonomy they are following? And how do you do that?

P1: Yes. It's a case by case situation, I have to say, basically. It is a bit of detective work. Sometimes you would hit lucky and you would be able to find out what taxonomy has been used. Other times it would be a bit of a dead end and you have to make certain assumptions, because you just don't know. So it really does vary, from my personal experience anyway.

SY: And then when you make those assumptions, do you clarify the assumptions in the work to ensure transparency?

P1: Yes, I try to yes.

SY: Do you find that the way that people present or clarify what taxonomies they use differs on domain? Species? Conservation? Ecology? Do people approach the problem in different ways? Or are people more clear and better at defining their thought purpose or more case by case?

P1: I'll have to think about that.

P3: I think that case by case. There are so may existing taxonomies people can download and install, to validate their work. So it's whether or not they are I guess savvy enough to do a bit of research to sort of link theirs up. But no, I don't think it is domain specific, or not from our experience.

P2: Fully agree.

P1: [Nods head in agreement]

SY: Explanation about NPS, and representation of terms as multiple units (word units) or term per unit (unified). Any thoughts?

P1: I don't have any specific thoughts at the moment. I am just taking some time to have a proper look at the… diagram you are presenting.

P3: Same here.

[Stuff about zooming.]

P1: Sorry could you just clarify what was the selection process for the central words, where you have other aspects coming from it? So, for example you have larva, you have perch, you have trout.

SY: These graphs were taken from the collection of WS I had extracted from the data. The processing steps went from: I got the corpus, body of documents, then used GNRD to identify scientific names that existed within the corpus, which I then annotated to put SCI if grammatically related to other words. I also identified more general terms and common names (through corpus analysis). Tagged with general-type or common tag. Then pulled all the WS – collections of pairs of words and the relation between them. Parent/Child or Child/Parent. Ensured there was no duplication of the terms . As you have reciprocal relations – if the same pair of words identified twice but in the reciprocal relation. So you have a list of word pairs with the identification of the type of relation. And then put into Cytoscape, which produces the graphs for the different pair. And identify each word, which is a node. With the relations between them, which represent the relation. The arrow points ot the child of the relation.

P1: Yep, yep.

P2: Not sure if it applies here but I will ask the question anyway. I see that says Linnaeus there. In terms of scientific authorship, so who first described the species. In my experience, it depends whether you have got parentheses around the name or not whether something has synonyms, I am not sure if you have come across that in your coding, where there is a further grammatical layer to that as well. In Linneaus, if it has brackets around it for example Perca flavescens Linneaus. If it doesn't have brackets it means that the species has not been relisted or reclassified. And if it does have brackets, it means that it was originally described as something else. I am not sure if this is relevant to this part of it. But just a note on…

SY: I was not aware of this and it did not form part of this analysis.

P2: I stumbled over that because I had written a name and then they said, no that's wrong because it's been redescribed and you need to have the brackets around the scientific name and year. So, you know, Linneaus and like 1800 or whenever he described things. But that's just a point on coding, not perhaps what your question was but perhaps the knowledge that parentheses would come up as a factor and it might be something you would need to incorporate into your analysis. So just a thought on that. Happy to go back to the actual question now.

SY: Really useful. Thank you. If it has no bracket, it has not been reclassified. With the bracket, it has.

P2: Yes, but just google and cite it. Author citation. Zoological name and nomenclature. So just a point on names, I guess.

SY: Any thought about the different representations or shall we move on.

P2: I can only see the left one at the moment.

SY: If I do that?

P1, P3: Yes, that's better.

P3: I was just going to day, I think it is a nice model of the data, actually. Are you doing graph analysis over that as well?

SY: Some. I identified through the analysis and playing with Cytoscape that there were certain characteristics that seem to form patterns. You see the larger dots…it's so long I have done this bit. It was related to neighbourhood connectivity… they were incidental characteristics. I identified characteristics that in which there were patterns of the analyses used in graph analysis: it was neighbourhood connectivity and closeness centrality and they had inverse relations depending on the bit of the graph they were involved in.

P3: Sorry, you dropped off at the beginning of that, but I kind of I understand.

SY: Thanks for sending in the FG questionnaires. Explore a little bit more.

SY: What knowledge resources do you use and why? And any problems you face with them or not?

All of you said you use the taxonomic resources to check name variant status. Also many to check classification of the name variant. I am interested because all use various different resources. Can you discuss when you use one resource or another or always a mix? If there are specific reasons as to why you use one resource or another to discuss?

P1: I can jump in there. Once again it's kind of a bird focus. A lot of my research links in with extinction risk, so I use the IUCN Red List data quite a bit. I use a taxonomy that is endorsed and used by BirdLife International – which is the custodian of the bird section of the IUCN Red List. So I tend to always go to their most up-to-date checklist, based on the taxonomies that they use. And that's what I use. But that is what I use on a kind of global scale. But I do use a British taxonomy as well if I am focused on more national based data. Because there are some slight differences For me it depends on the regional scale, actually. So whether I am doing sort of more global research or looking at more national checklists. So I guess that is a starter for that particular question you asked.

SY: Asking about the differences: can I ask if there is more or less information or differences in categorisation (between the taxonomies)?

P1: There are some slight differences in their actual classifications, across the different platforms that I use. There are slight differences in naming consensus and sometimes even differences in terms of the higher-level taxa, so kind of family level and even orders as well. So yeah, there are some differences.

[Missed question opportunities: how are these classification differences shown – is it just in descriptions, or are different nomenclature terms used – you say there are differences in the higher level taxa, can you give me an example?]

SY: Do the others find this as well?

P3: The differences in the taxonomies? So yeah. I mean one of the projects we are working on at the moment, is based on the Wilson Reeder mammal taxonomy. I was published 11 years ago. And so trying to join up the data that was published under that taxonomy, with what is now considered the standard, which is the ASM, mammal diversity database. And the number of redescriptions and synonyms... So trying to match up those two different taxonomies, even for a mammal taxonomy, yeah is incredibly complicated.

[Missed question opportunities: how do you go about that?]

P2: A lot of databases obviously work as crawlers or whatever, they extract stuff from papers and that sort of things. But with the papers I recently I have been involved in they tend to all defer to GenBank basically so it's, uhh.. whether you are working to genetics or not, but it would be tied specifically to the genome, so it will be an identifier 16rnsDNA,…RNA or something. And that will be encoded in GenBank specifically with a name and now will link back to that. And I mean that certainly works when you know what you have molecularly. But at least with lots of amphibian and reptile papers they are supposed to be able to say what it is specifically. And that way if it changes it is still tied to an actual, because that's obviously the most fundamental unit of description. I guess database-wise there are some differences within them. But at least for amphibian research they now rely on GenBank…. But again, I am not a geneticist. So I have only be involved in sort of the data collection side and some of the more conservation based aspects, so I am a bit of a novice in how it all works.

P3: That's interested because I am involved in paper doing data angling for a new polychete worm description, so a new species that was discovered. And they, we were struggling to get names published because we don't have the genetic breakdown along with it. So they are now refusing to describe new species without the DNA.

P2: Yes, that's not surprising. But, interesting

P1: Yes.

SY: Interesting. Difference taxonomies and main focus or regional or global might use one or another. For example writing a paper focused on something with a regional focus – how traced back. Is that across the board?

P1: Um I can go first again. Yes and if I am reviewing a paper and they don't tell me what taxonomy they are following I will make a comment and say please tell us what taxonomy you are following – for that transparency like you said.

[Missed question: how can these be integrated with the varying structures of taxonomies?]

P3: Yeah and we try and use GBIF, try and with a link back to the GBIF specimen items wherever possible. Because obviously that incorporates a lot of regional differences, that sort of thing, Catalogue of Life, taxonomic background.

SY: What are the problems with taxonomic resources? What situations?

P3: I guess just, um, trying to track changes in redescriptions over the years, particularly for some historical ones that aren't in GBIF. But on the whole, it's quite easy to track down, once you… [SY: understand?]. Yeah, yeah, exactly. And I find Wikipedia and Wikidata are useful as well. Because there's often references to things.

P1: Yeah, I mean so, the datazone that BirdLife International use, and the taxonomy that they use. They're quite good, you go to any species with a common name or what you think the scientific name is. And it does give you a history, or if the name's changed and the synonyms that exist as well, if any. So that has been very useful to me.

P2: Sorry I have got my email window up I have to minimise it. [Repeat the question] Initially, but, as long as you know there is a problem there, then that's fine. If you are not aware of the fact that there are several names is what causes the difficulty. And often it's not until you have read several papers or some old papers and you're like what's this, is that the same thing. And then you check. So often it's… if you're starting on the taxonomical side and it's just the fact you are mentioning a species and you have no knowledge of its history or descriptions. I guess it could lead to missing certain research papers but usually you would use multiple keywords and terms to locate the information you wanted. I guess there are a few times I got a little confused. But it just takes time basically. As long as you are aware of the problem, then it's OK to resolve most of the time.

P3: I agree.

SY: What you said clears where you talk about the lag or mismatch between the IUCN and taxonomies – with the Birdlife International. That is what you were talking about there.

P1: Yes.

In the NPS I chose 3 different resources: VTO… etc. A couple of you use the CoL in your work? I don't think anyone use the others?

P1: A little bit.

P3: ITIS yes, not the VTO.

P1: When I tend to look at other platforms is when I am looking at non-avian taxas, non-bird taxas when I tend to go out of my comfort zone, that's when I tend to look at those other bigger platforms.

SY: How is CoL?

P3: It's quite good, we run a platform called ScratchPads, which allows scientists to create their own taxonomies and describe their species and specimens, and one of the import mechanisms we use on that is to allow them to import the CoL. And that is probably the most popular tool, that we have. Yeah, on the whole… we have done some analysis on how much they have modified it after the import and it's not hugely modified afterwards. So, it seems the scientists using the platform seem quite happy with it as well.

SY: Just examples of the different pages.

Can you discuss where you find the greatest ambiguity in scientific nomenclature usage? And if these problems are more present in one type of data or another?

P1: I think it is a little bit mixed. I think it depends on the paper, on the journal. When I'm reading scientific articles they don't always clearly highlight where the data, or what taxonomy they are using. But a number of journal articles do. So it can be quite mixed when looking at scientific literature. But in terms of databases versus narrative: I still personally find it think it is mixed. I don't particularly think it is more one so or another. It depends on the database, it depends on the narrative that you are looking at. That is just my own personal experience. I can't think of any standout exceptions to that. I don't know if it is different for P2 or P3, but...I haven't noticed anything stand out in terms of issues.

P3: I guess for me, publications the one issue we have is when we're trying to parse a description often their species description is grouped at familial level. And so you need to have a semantic understanding of the entire narrative. Because they'll have the specifics of a species but then the broader, high level description which is applicable across familial level... But yeah just some publications do it that way, which makes extracting the data harder. But I guess for a human reader it makes a lot of sense.

P2: Probably the same, no real apparent differences. Depends on sources of things and what you're reading. Take say, rhinos as a specific example, how I mean obviously the media don't get it right, for most things, in popular science or news article and stuff. That's where I normally spot the biggest discrepancies, or differences between things. But the scientific papers normally you can trace what taxonomies they're using and things. Not that it doesn't really apply and it doesn't matter. But take the white rhino, which I know a lot about. Take the two species that there are and say… I mean in the media. I mean, it's hard to say, what is a subspecies? Even scientists would debate that. So we have two types of white rhino, a Southern and a Northern and one is going extinct and one isn't. A lot of it is down to terminology. That's where I see the ambiguity basically, in non-scientific writing. Which is less of an issue in this discussion, I guess but… I think all scientists get annoyed about but it is obviously difficult to police.

SY: Ambiguities: species or subspecies. Does it depend what you are doing whether that matters or not?

P2: It only matters if you are trying to treat those subspecies or species separately. So if your conservation initiative or media piece is about specifically about one thing or another and you are classifying it through name rather than through a geographic indicator, it would matter. But yeah, I guess it can get a little confusing otherwise I suppose.

P1: It can get confusing if you are doing trait, character-related analyses as well. Is this unit a species or a subspecies, are you merging trait-data together when perhaps maybe it should be split? And it's also, I guess this links back to the previous question, you had. If you are given, I don't know a trait database, you are given some trait data, it is really important that you know what species and what taxonomy is being assigned to it. Because you may need to pool that data if you are using a different taxonomy or might have to try and split the data somehow otherwise. It can get quite difficult if you are interested in trait and characteristics data but there isn't that transparency about what is being used in taxonomy as well.

Sandra: Splitters and lumpers – how does it affect you? It is similar to subspecies and species? Being able to trace it or not?

P1: What immediately jumps to my mind. And once again coming from both the conservation background is that it is still the case really that species are the unit of conservation. There are cases, once again going back to birds and you have had a species that is near threatened or perhaps a species of concern. And perhaps this gets split into a number of species and their threat status gets potentially upgraded – because they have a smaller population size, smaller range size. So there are a number of kind of discussion papers that have looked at the impacts of taxonomic changes and splitting and lumping and how that impacts conservation status. So there are some quite significant conservation implications as to how you define a species. And that's not just with birds.

P3: Yeah and so the project we are working on at the moment. It is obviously doing climate change modelling based on geographical location of the traits associated with the species. So if we can't actually tie down exactly what species it is and identify the traits to go along with that species, the models break down and the data too. Yeah, so it is vital.

P2: I agree with what P1 was saying. The direct conservation implications of say for the frog species I have worked with. Where they are species then if you split it, it might immediately be critically endangered whereas before it was just vulnerable. Or somewhere like Madagascar where they have some frog research out there a lot. Or primate species such as the sported lima, it used to be one species across Madagascar, now 20 species. And they are all obviously critically endangered. And that mostly been split on genetics, not even traits and things. Depends on the field. Bigger impacts where splitting just by genetics. Everything seems to have a different criteria. In amphibians I think you take a 4% difference in the RNA you are looking at. Sometimes it can almost be quite arbitrary how you are splitting it, but it will have real impacts on how you conserve it. Conservation issues… because that is what I know. I'm sure there's other impact to.

P1: [Nice paper – will send]

SY: Ask about terms synonyms and misspellings. How you define them? On case to case basis – idea of a misspelling and idea of a synonym?

P1: I was just gonna say. With birds, one of the quite frustrating things, I don't know if it is the same for other taxonomic groups, but with birds quite often with synonyms there are very subtle differences. It could just be the last couple of letters in the specific name. I know if I have been searching a database for a specific species and I just happen to have used one of the spellings over another I can, you can easily miss a species. So sometimes it can sometimes be very subtle differences, that are synonyms, recognised synonyms I am not really answering your question but it's something that I have noticed is particularly prevalent. If something ends in an "a" or a "us". It can be quite subtle but it can made a big difference when you search for species, unless you use a wild-card search function which is what I tend to do nowadays. But yeah they can be quite subtle differences.

P2: Yeah I agree, again not really answering your question either. One of the things you get in species classification the oldest name takes priority. I forget the term. But basically if the original classifier 200 years ago originally called it something that didn't make any sense, in terms of the description, etymology of what the word means. You still get stuck with that, that's the official name for it. It's the oldest that takes it. So you might have something that is incorrect – I can't think of any examples but there are some good ones – where something is actually called something that it isn't, in the Latin name. So that is something that does come up. So if you read the name and you understood some of it, you would mis-conclude without further checking. Not really answering your question but…

P3: I guess for me mapping between ASM database and the older version of it. In the new version they are actually commenting on the fact that the Wilson and Reeder taxonomy misspelt some of the species' names. So you can see those propagated through all the different sources that have then used the Wilson and Reeder taxonomy. But now it's been corrected. So, having a definitive taxonomy without spelling mistakes is good.

VERNACULAR VARIANTS

Different responses about vernacular variants: different data

P3: So I think on the whole, the only time we have a problem with vernacular names is with the current project. With the trait mining in publications, because they kind of set the scene before describing the species sometimes. Especially in older publications. Then use the vernacular to describe the landscape with a few useful traits thrown in. Apart from that we don't actually use vernacular names very much.

P1: I use vernacular names quite a bit in my teaching, I always try to provide the scientific name as well. I am fascinated with vernacular names for birds anyway because there are so many, they are very regional, very kind of localised names as well. But I try to steer away my students from being reliant upon them but I wouldn't… it's a huge part of ornithology, the vernacular names and it has a lot of history. But yeah, always try to stick to the scientific names.

P2: I use the scientific names but then I will lapse into vernacular or common names just for readability. But I will define them before that. If you have a particularly horrible Latin name you try to use it less. If you are writing about E-coli, in a different field, it's easy to talk about E-coli. But yeah, only for readability. But it's always, I just stick to scientific names, the same.

SY: It's just about usage and how much it comes up. Interested. What it is like in your daily lives. Looking at different info: any info that vernacular names can provide about geographical context and contextual specificity. P1 has mentioned regional differences.

P1: With, obviously not just focused on birds but it's not always the case that scientific name has much to do with morphology with the species and sometimes the vernacular name can actually be better from an identification perspective. It can do a better job of describing the appearance of a species. Obviously it depends on which vernacular name you're using. Some aren't particularly informative at all in helping to identify a species. So there's that to think about. It is very variable, obviously. Yes, sometimes the vernacular name can be more helpful in terms of identification, than the scientific name. Sometimes the scientific name could be named after a celebrity or Coca-cola or something [P3: Yeah, yeah]. It can be not at all informative the scientific name sometimes.

P3: And the only thing, I don't know if in citizen science using vernacular name encourages more occurrence records to be deposited.

P2: Nothing to add.

P1: Fair point that P3 was making, we do quite a bit of citizen science work. I encourage people to use iNaturalist and things like that. Of course there is no expectation that citizen scientists are trained taxonomists, so yeah, in those situations use of the common name is a given. But with iNaturalist you can upload an observation, provide the common name but then you'd have the online iNaturalist community helping to get it down to a proper scientific observation.

SY: Explain the contextual ambiguity/authorship inconsistency. Ambiguity problems in vernacular – how problematic? Or generally not so many problems – because of using common in combination with the scientific nomenclature etc.?

P1: For me it links back to the citizen science. So for example when people say I have seen a gull, or a bird that is black, I've seen a black bird. That is very hard. When you have multiple gull species, and which species did you actually see. So yeah that can be hard, the generic descriptions of a species. So yeah, particularly for citizen science it can be very hard to know if they are using the correct name, and if they use a generic name what can you do with that realistically.

P2: One frustrating thing that I find. Take something like Google search for images, if you put in a Latin name or a common name, the stuff you get back... I know it is not an authoritative resource, because it's a crawler and it pulls stuff in from everywhere. But particularly for a frog or an insect it very rarely reflects what it is that you are searching for. And the reason for that is because people mislabel things with different names, different common names or there might be an image which has multiple names on the page. A bit of an aside there, but say if you want to find the species or a frog, say Rufus anfropensis, or something like that, type it in and it will keep giving you images and it will all come from different sources, with different things. It doesn't really matter for me, as a scientist because I know most of the time what species I am looking for. But for people just trying to ID things it is not a good resource. As an example where you try and find out what your species look like you will get some very strange pictures up normally. But a very minor annoyance really in the scheme of things really, but it is something that does happen.

SY: Authorship – lack of – surprised or normal?

P1: I was just going to quickly say that I have, in some of my publications when I've mentioned species I haven't provided always the full authorship, especially in my earlier papers, I just haven't. A lot of journal articles, particularly with a conservation focus, authorship isn't always given. I think different journals, different kind of requirements, that's just on my own experience. But yeah.

P2: Some use it, some don't. On a taxonomic paper you would use it because it's important, but in most other situations it is something you should use. But, only some journals ask for it.

SY: No authorship does it cause confusion or not?

P1: I don't know, just trying to think. I guess the potentially the lack of transparency and lack of acknowledgement. But a lot of journals I publish in don't explicitly ask that information to be provided.

P3: I guess the one time where it can be useful is if you do have a particular discrepancy in the name that you are trying to track down, then history of the name, having the author is useful but usually it's fine not to have it, I think. That's on the data mining side.

P2: I guess most people just drop it. Technically a full species' name does need that qualifier. It is not just the Latin, it's the author as well. That's just how it is. The zoological nomenclature rules. The bracket is there to help people. But if nobody knows what it means…most scientists don't use it because it's not relevant to them…. It's a paper trail really. But you don't use it really apart from it you are describing species really. Or in taxonomic papers. At least from my experience. It's not really necessary most of the time. It depends if it's changing a lot. It depends on the field. Some stuff are static for hundreds of years. Other stuff is frequently changing. Because there are still parasitic groups, or species complexes particularly for amphibians, where people know that what something is dubbed as is wrong, so then it helps to be as specific as possible because it hasn't been described properly yet. It's context specific.

P3: I don't know if there are discrepancy between journals in different areas. We mainly do sort of entomology. I know botanists are better at describing new species. They have that conference every year don't they where they get together and sign off all the new names. So I don't know if they are better at citing names maybe?

P1: Interesting, yeah.

P3: It might be interesting to find out.

SY: For me it is to see if it causes any ambiguities, unless if there are specific ambiguities as to differences of opinions.

P3: Yes, I think it is more about the paper trail. A more accurate paper trail is the most important.

MY DATA

SY: NPS – looking at the different variants that appeared in the different resources.

P1: I'm amazed at the different number of variants, scientific variants for what is it, the brown trout, the Salmo trutta. It's a lot of different variants.

SY: About the spread of usage.

P2: I wonder with things like fish I wonder how much it is ties to things like the field that they are published in. Like say aquaculturists there is probably less… not to criticise other fields but some things that aren't about the taxonomy, there are lots of scientific papers that are published by specialists in one area but they are not necessarily taxonomists. I wonder if say, for a particular fish species where they might be being farmed in big systems, commercially they might not care what they're calling it, because they might all know what they mean, it might be something they use in business but I wonder how much of that is field dependent. I would say maybe fish in particular where you have all kinds of strange papers and stuff on specific areas. And how whether they aren't necessarily using the same names as other people. You do see that in some literature where some of the more applied stuff perhaps written by types of academic, or not academics even, you might see more name variation.

P3: Yes, that's really interesting. Fish are probably one of the few areas where lots of members of the public care exactly what type of species it is. I guess orchids as well, are areas that could…

P1: Yeah. Yeah…

P2: I think botany in general.

P1: Did you look further at thematic analysis as to the context that these names were used in?

SY: No, across the different corpora. [Wrong - I did look at where in the article]

P1: Another PhD probably, wouldn't it?

SY: Ambiguity – if intra-domain everyone understands – where specifics aren't so important but in taxonomy it is.

P2: Somethings when the species names changes people don't use the new names because of local usage and like you said recognition like say something like acacias as a big family of trees now. Half of acacias are longer acacias. They've got split and have been renamed to another genus so like Acacia tortilis or whatever it's called, it's now something else tortilis, but all of the people in the field or the field guys and the people who know what an Acacia is, and we all know what an Acacia is probably. So they stick with the old usage just for ease for the, again I get annoyed at papers, even some tree books I've seen for example, have said although the new classification that there should be this we are going to stick with the historic names because everyone knows what it is. So it depends on, on that too. Yeah. General usage.

SY: Arguments for both ways – are there sometimes disagreements where you have these changes? Any other disagreement about why to assume the new name?

P2: There are very specific guidelines from the zoological nomenclature. They say no, you have to use this name, like I was saying with the oldest name guidelines or whatever. So the guidelines are specific, but that doesn't mean that in non-academic or non-taxonomic papers that people will stick to it, I suppose. At least from what I know. I don't know, I could be wrong.

P1: Ornithologists are quite an opinionated group of people and there are a whole forums devoted to debating changes or proposed changes in avian taxonomy. And, yeah, that's gonna be the same for other taxonomic groups as well. But yeah, you can argue all you like but if there's been an authoritative change, that is the name that should be used.

P3: That's actually one of the few, a few places where you can see on ScratchPads people importing and cataloguing with life taxonomy. And then there will be changes. You see, they they've been putting it back to how to the current taxonomic structure that they're familiar with.

P2: You can look at that with Wikipedia for a few species and see the backwards and forwards of this.

P1: Yeah, yeah.

SY: Say I'm in, in the data that I looked at then identified these four different areas where there are potential ambiguities in usage. About vernacular, spelling, authorship and So I'd like your opinion on the way that I've split. This one has been that spelling. And then I've got this one which is accepted versus the most used term. And then contradictions are seen test day to a knowledge resources that that the identified I don't know what you think about these as grips for ambiguity, this one if you don't understand what I mean by any of them. If you disagree with these being relevant ambiguity is on if there's anything else that you'd add.

P1: And could you just clarify the final one, the one in blue, contradictions one.

SY: The blue one was that I identified in a couple of cases in which my test data and this is relating to linking of inaccurate, inaccurate terms with a specific species. So it looks like there's a contradiction between the inclusion of a particular variant within the taxon that it's been included. But it's just it's when it says contradictions it's just that this is what my data suggests. It's not anything conclusive but that was what that was, what the meaning was in that.

P1: Do you have many cases of that?

SY: There were three and it was all relating to the common name that had been included. That through my days would suggest that it was my data and a very non expert look, search on Google, but like the data suggesting that we're having that.

P3: Yeah, I mean, I think the accepted term versus most used, that's kind of what we were just talking about wasn't it. People have like a favourite name and they're not going to change unless they're forced to. But yeah, that contradictions between some test data, knowledge resources. And just that kind of include, I know lots of this sort of name resolution services. I mean, some of them were created about 10 years ago, they're still live, but the names in them are now out of date a lot of the resources which doesn't actually resolve the currently accepted name which propagates the inaccuracies I guess.

SY: Definitely. When we look here, when I put ambiguity here than I thought to split up the different ambiguities that you find with vernacular clue, and you can see that there's a broader meaning. So in this case, trout was included in the Catalogue of Life as a synonym for both Oncorhynchus mykiss and Salmo trutta. It's not that it's wrong, it's just that it's even more broad than saying like brown trout, then just including trout and I just wanted to show this as an example of what I mean by broader meaning and how this comes up in my data. So you can see that trout is linked to lots of lots of different species in the data. And you can see that it the data shows it being a parent of those in general. Yeah, so just wondering if that is the sort of things that actually represent them. And if you think that this causes ambiguity, or if you know a reason as to why such a general term could be of use?

P2: I think it can be quite species specific.

P1: Yeah, it was like when I was talking about, you know, the gulls and it depends on the context, generally speaking, it's not that informative, those kind of names. But yeah, it depends on the wider narrative that you're looking at? Yeah, yeah.

P3: Yeah, I completely agree with P1. It depends on the sort of the context and narrative. It's some, I feel like you're trying to extract data from a paper that uses in terms like that. They often use familial rather than sort of common, vernacular names and have a broader concept and drill down further on, but I guess it's just sort of trying to understand the semantic meaning of how the term's used and annotate it.

SY: Broader or narrower meaning. Sea trout as a parent of Salmo trutta. And then in the web corpus Salmo trutta as a parent of sea trout. Can go back into the data to look at instances to see specific context. Sea trout, with Salmo trutta as explanation versus sea trout as anadromous form of Salmo trutta. So it's different contexts in which in which these terms can be used. And so yeah, we're like going back to the discussion that we had about whether extra information that can be found in and vernacular terms or do you think this causes ambiguity? Do you think it can add extra clarity? Do you think that there are different sides to vernacular terms in the way that they can either cause ambiguity or add extra clarity?

P1: And whether they yeah, they can definitely cause ambiguity. In the sense that some people use a given common name to represent different species. But at the same time, vernacular names often kind of more accessible and it's what people tend to readily use. But it comes back to the sense but I would never for example, I wouldn't be allowed to anyway, I would never publish a paper where I just use common name, I would always have the scientific name used alongside. But yeah going back to my previous point where it actually sometimes the given vernacular name can be more informative to somebody in terms of a species identification in terms of what it what it looks like its appearance or its locality. But I see them I see the vernacular and the scientific name is kind of going side by side. Not…

SY: Yeah. I think I think maybe I need to clarify and something in the, in the data. I mean, none of none of the none of these papers would just have had the common names. I mean, these are the relations that are coming out. It's because they're found next to each other in sentences across the data. So it's very much that they're used together as I think P2 was saying for the readability for the usability If you use the, the scientific term and then throughout the narrative, then what will be used if you're always talking about the same, the same taxon, then know you use the common name that you choose to use or event or a number of common names, and many, many of these. And so, one of the things that I was interested in here is because of this issue with ontologies. And being able to accurately integrate the data, is identifying the importance of vernacular names in this and if there are the ambiguity, can these sorts of graph be used to identify how they're being used in these particular in a particular context, so that if there are multiple interpretations of them, then they can separate out. In this context it's being used in this way. And so it should be mapped in a particular way or it shouldn't be mapped in that way. Yes, it's definitely not that they're being used in isolation.

P1: Yeah. I mean, I mean, I guess another, I mean, I'm sure you're aware of this, but not every, it's only minority of species that have been described that have a commonly used vernacular name, anyway. I mean birds with a bit of a special case, and I guess mammals as well, in the sense that they pretty much all have a commonly used vernacular name, but a lot of species, especially those that have been newly described, invertebrate species, or plant species as well, but just they just have a scientific name, they might obviously, have a very niche vernacular name that local communities use but not a globally used common name.

P2: It also depends if the species you're talking about appears outside of the scientific literature or not. If it's like a deep-sea worm and it's only ever talked about in this scientific paper, even if it's got a common name, no one will use it.

P2: And there's no there's no ruling for common names something could have, from what from what I know, at least, I could be wrong. The species could have 10 common names, they are all equally valid. There's no such thing as an official common name, I don't think.

P1: But it's interesting. I don't know how, for example, the IUCN Red List or Birdlife International, how they decided on the common name that they use. For birds. But I don't know what the.. because, for some of them I question because I would use a slightly different common name for some of the birds.

P2: What are the authorities for it. Because in papers when you describe a species, you don't even need to give a common name. You can suggest one but people don't have to use it. [P1: Yeah.] I'd say, for all species descriptions I've been involved in. You explain why you pick the Latin name. And I think only in a minority of cases do you present a common name. People might use one using the Latin name. But often they don't, I suppose. I'm not sure really how it works to be honest.

P1: Yeah, yeah.

P2: They are strange. Yeah, they are vernacular. I mean…

P1: But it is interesting. It would be interesting to know how the IUCN decided to yeah, use the common names

P2: Which one because like you said they regional. Well, that's the other thing it's language specific. So I mean, yeah, what we might call the something frog. Someone else is going to call it that, so… language specificity, regionality. I don't know how it works, really.

P1: Really, I asked my students that last week I think I said I was talking about the mountain chicken and they all assumed it was obviously a bird species but it's actually a frog an amphibian species so that's a case where it's very confusing yeah.

P2: [inaudible]

P1: So yeah, they're kind of cases like that where it can be very confusing, but actually the name which is commonly used that is the kind of the main common name used in English anyway. But that is strange.

P2: Also where you have a common name in the local language but not in international not in non-local places. Even I was someone like chicken frog in Monsterrat??

P1:   Montserrat…..

P2: Yeah, they might call it the MonPoulet or…

SY: Yeah. Yeah. All right. Well, so I guess, with species that are very regional, then they might, they might I mean, they might have a vernacular variant in that language. Because it doesn't exist outside their area.

P2: Absolutely. It's not communicated elsewhere. If it doesn't appear outside the literature. And there's no reason for it to. Yeah.

SY: Parasalmo mykiss and Kamchatka. Limited data. Both corpora only two linked to Kamchatka steelhead and this scientific name. Maybe geographical – Russian articles or referencing the Russian article.

P2 It may be author specific to some extent. Yeah. They like to call it that and they've written multiple papers? Or group specific entry might refer to it's just that lab. We could decide we wanted to talk about something and call it the Brighton newt. It'd be wrong but…. I take that off the record that's a stupid statement.

P1: This obviously isn't the, the internationally recognised scientific name. So, it's interesting, we'd be interested to know what the where, you know the etymology work. Where did it come from this particular name?

SY: The Parasalmo mykiss?

P1: Yeah. Or am I getting that wrong? Is it that you are saying that this particular species, it is the steelheads, is this the official scientific name? Or is this…

SY: Parasalmo mykiss appeared as a recognised variant (scientific) in I think it came up in both the VTO and CoL as a synonymic variant and then the Kamchatka steelhead comes up as a connected common name vernacular that we've identified, but it only comes up in this context.

P2: Maybe, maybe this paper was written before the internet, and it was just wrong.

SY: Steelhead and rainbow trout vernacular. Sea or freshwater. So they're not they're not exact synonyms. Like there's other information that they're basically I don't know if that's something that happens in any of the species you work with

P2: I think anglers, fisherman probably are just doing strange things.

SY: These are in ecology papers have been when, I don't know. So you think you think that it's a anglers and they’re talking…

P2: As an example, they're probably, actually I have no idea.

P1: I'm trying to think of bird related examples where the name kind of changes come in top of my head.

P2: You got things that metamorphosise so when you got insects, something Caterpillar might have a different name to something butterfly, for example, and all the other things that do that species that have got an aquatic stage, and then a terrestrial stage, I imagine that a lot of that were some is a nymph then it's a or marine stuff too. Where you have a planktonic stage and then adult stage where they're very different morphologically. And then the common name would of course be different to go with that. But I can't think about why it would be otherwise.

P1: I guess you can kind of maybe extend that to perhaps different common names in terms of if you've got a migratory species, you know, where they winter and where they breed and will have different regional common names for sure. Yeah. So yeah.

SY: Um, so the another thing that I found and this is why I was interested in asking you about when they what you what you interpret as synonym to be and what's a misspelling? Because one of the things they found was turn in the resources that I looked at, then Salmo gairdneri with two Is was an accepted spelling, a former accepted name I think. But the more frequent term was with a spelling with one I.

P2: In terms of Latin it depends on the journals they're putting on me pretty good but if their editor or peer reviewers haven't got a knowledge of how you say Latinise a toponym and you get genders in Latin and things you get like neuter, masculine, feminine that can slip through and it does in a lot of papers where you get stuff grammatically misnamed but then that gets stuck because that's what it's called. And then someone may try to fix that because they've got a knowledge of Latin and yeah… that's my opinion of that. There are very specific rules for how you say take certain things like, say, like a species named after a toponym, a place name would be neuter normally and so that ending that you take with end in a certain syntax or suffix. And that can create complications if it is done wrong. Because it is a different language. A lot of people yeah, a lot of people just like it is about adding a double I or IS on the end of this but it is actually much more complicated than that. And if your editor most of goes through pretty rigorous peer review, but if it is not. Sometimes stuff gets missed and it goes strange, and then someone will spell it right, which is wrong. I can't give a specific example, but it definitely does happen.

P1: Yeah, sure. Yeah. I mean, most, I guess it depends on the nature of the paper as well. I mean, if it's a single species, paper, perhaps it's easier to patrol and check. But if you've got another study, which is as a supplementary material, got this massive species dataset, I mean, who's going to go through and check that the spelling that they've been using is correct. I mean, normally you don't peer review or you don't normally kind of proofread or proof-edit any supplementary material you would take a look at it, but you wouldn't be scrutinising it with a fine tooth comb. So yeah, there will be definitely issues with spelling that slip through because of that as well.

P2: 1:51:16

Just typing people can... Yeah. Yeah. When I submitted a paper recently on rhinos, I told you this P1, I misspelled the Latin name in the title. This was something I had been working on for four years. I noticed it, but I don't know if anyone else would have noticed it. And I submitted it. And so human error.

P1: Yeah for sure. So it's just like Chinese whispers isn't that you know, you copy it from a source. And it's wrong. Then it gets carried across, doesn't it? And that’s the thing I guess a lot of a lot of scientists and I'm including myself because I didn't you know, study Latin at school or anything I started using scientific names at university and but I was never really trained up in you know how scientific names are properly constructed and decided upon, P2 you probably having been involved in species level descriptions, you'll have more experience definitely than I do, but it's just it's a tool, that we use. I wouldn't say I was at all knowledgeable regarding how scientific names are properly constructed apart from just, you know, the binomial process of it. But yeah, I think a lot of, a lot of mistakes would slip through because of that lack of understanding of Latin too.

SY: Yeah definitely.

P2: I mean, it's not meant to be calling it the wrong Latin name. It just might be grammatically incorrect. Some specific rules but you can see them take a weird spelling if you want, but it would be very strange. Yeah. There's no, whatever the zoological nomenclature says goes basically the guideline. We have guidelines be like one of the guidelines is you're not supposed to name a species after yourself, you can name it after anyone else, I don't know that is a guideline and I don't think it's rule.

P1: Yes, my advice. Good Practice Yeah.

SY: Yeah and here I am looking at how people apply or not the according names. Here looking at where there are variants more frequently than accepted names. I think we've discussed why that might be. And then the fact that it does happen because of differences of opinion, or because the people I have there have the name that they prefer to use. I mean, yeah, yeah.

SY: Contradiction with knowledge resources. And with Oncorhynchus mykiss in the CoL the species was actually linked to brown trout. But then when we looked at the data, then it was like brown trout was only linked to Salmo trutta. And it was something that I was interested in because as you said that like sometimes one or the inconsistent usage of common names or using for multiple different species, but is that like? I mean, I guess it's something that I could speak to Neil a bit more about, but that is something like that. Is it likely is it this, this would be used or do you think the data is likely to be corrected in the assumptions being made? That I don't know. I don't know if that's something you come across or if people will just use common names for different species.

P1: Well, I think I mean, I'm hoping that people wouldn't use the term brown trout to describe a rainbow trout. It might happen in terms of, you know, case of missing of misidentification, but I'm sure. Yeah, it does. It does happen. People kind of saying that they are using a particular vernacular name, when perhaps another one be more appropriate. But then, you know, as we said, as we said before, should always be kind of associating a vernacular name with a scientific name. But it's interesting that that came up in your findings. Yeah. But no just saying it's an interesting and interesting finding. I just gonna say Neil would probably be able to comment on that.

P2: I don't know I think, some vernacular names you have to be sceptical of them, not rely on them because that kind of stuff does happen if they did that with the Latin name, use the wrong Latin name, that's when you run into problems.

SY: I guess. Now I mean I was interesting because the inclusion of brown trout was within the database, the Catalogue of Life it was it was it was the database doing it wasn't what was it wasn't people actually using it necessarily?

P1: Mmm hmm. That's I don't know what I don't know why that would be the case. I don't know.

P2: I'm not sure.

SY: Showing the graphs.

What we're looking at now just the bits with the contradiction between knowledge resources, with the data and knowledge resources for saying that in the case of the CoL, then brown trout was identified as a common name variant for Onc. mykiss. But then in the data so with the, the academic corpus and here though you can see that it doesn't appear to be linked with Onc mykiss, it's linked to Salmo trutta. The thickness of the arrows represents the relative number of relations identified. So there was one with Onc.mykiss with brown trout but that was an error in my methodology when I went in to look into the data. And there you can see again that there's a strong link between brown trout and Salmo trutta and not between Onc my and brown trout. They're just discussing the way that my method shows this difference and also the way that common names can be used. I don't know if you've come across anything like that?

P3: I am not sure I have to be honest. I mean we just don’t really work with just vernacular common names. Yeah, no, I haven't seen so much of that. So yeah I haven’t seen so much of that. But like I said before, I think it's probably more common in citizen science and public derived data.

SY: lake trout and brook trout [ask Neil about]

EVALUATION

P3: I’d also be interested in seeing the data. If you’re sharing the data, definitely.

P2: Yeah, thanks. No big comments from me. Hopefully I was helpful. My limited knowledge. Yeah, maybe, yeah, try and get hold of some taxonomists I reckon they'd have even more strange knowledge of things about looking and yeah, also looking at the temporal differences in the dataset as well. Or the linear thing with variants or whether they're all in there at once. I think that'd be quite interesting, but nothing, and see if that helps get them out of the mess or whatever the changing or whether it is just going 1212 Okay, thanks.

P1: Yeah. And yeah, I guess. Yeah, just like I was just gonna say, was there anything from the the initial questionnaire that we filled in that isn't clear at all to you, or does it all make sense?

SY: All queries answered throughout the group. And then just before you go, I just like, I'd like to ask a couple of questions of just about a feedback from what I've shown you today is we'd consider the network graph that the network representations - Do you think that they accurately reflect or help to disambiguate any, um, big ambiguities that actually exist in the data. So if you look at the way that they can identify links there, I mean, obviously these are looking at linguistic links between the between the two different terms. Do you think that it accurately does that?

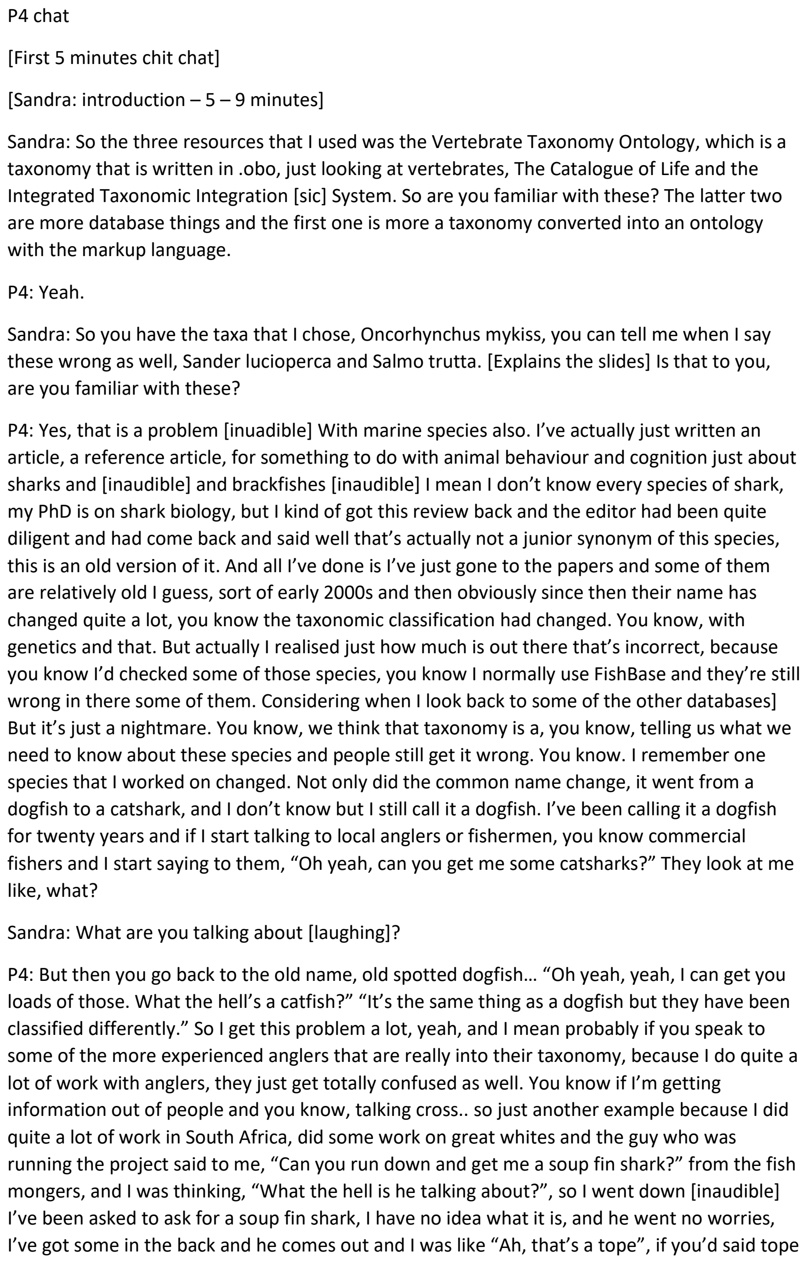
P1: I think anything that can help visualise quite a complex topic is always a good thing. And so, though I think it's certainly got a utility to it to be able to explore and to breakdown and zoom in and zoom out on the different, the different levels, so to speak, and I think yeah, like sounds better. That is a possible way of kind of getting a temporal element integrated into that would be, that'd be fantastic as well, more kind of context. But yeah, I'm all for visualisations of complex, complex data.

P3: Yeah, I completely agree. I think it's a nice way to visualise it. Anything you can do to visualise that's good.

P2: Good, I think adding extra dimensions, like time.

SY: That's something that I've got put in for future work because it was always adding more and more and more and more things in time and being able to do everything, but it's definitely possible to do. Thank you all so much.

F.7 Informal chat transcript



P4 chat

[First 5 minutes chit chat]

[Sandra: introduction – 5 – 9 minutes]

Sandra: So the three resources that I used was the Vertebrate Taxonomy Ontology, which is a taxonomy that is written in .obo, just looking at vertebrates, The Catalogue of Life and the Integrated Taxonomic Integration [sic] System. So are you familiar with these? The latter two are more database things and the first one is more a taxonomy converted into an ontology with the markup language.

P4: Yeah.

Sandra: So you have the taxa that I chose, Oncorhynchus mykiss, you can tell me when I say these wrong as well, Sander lucioperca and Salmo trutta. [Explains the slides] Is that to you, are you familiar with these?

P4: Yes, that is a problem [inuadible] With marine species also. I've actually just written an article, a reference article, for something to do with animal behaviour and cognition just about sharks and [inaudible] and brackfishes [inaudible] I mean I don't know every species of shark, my PhD is on shark biology, but I kind of got this review back and the editor had been quite diligent and had come back and said well that's actually not a junior synonym of this species, this is an old version of it. And all I've done is I've just gone to the papers and some of them are relatively old I guess, sort of early 2000s and then obviously since then their name has changed quite a lot, you know the taxonomic classification had changed. You know, with genetics and that. But actually I realised just how much is out there that's incorrect, because you know I'd checked some of those species, you know I normally use FishBase and they're still wrong in there some of them. Considering when I look back to some of the other databases] But it's just a nightmare. You know, we think that taxonomy is a, you know, telling us what we need to know about these species and people still get it wrong. You know. I remember one species that I worked on changed. Not only did the common name change, it went from a dogfish to a catshark, and I don't know but I still call it a dogfish. I've been calling it a dogfish for twenty years and if I start talking to local anglers or fishermen, you know commercial fishers and I start saying to them, "Oh yeah, can you get me some catsharks?" They look at me like, what?

Sandra: What are you talking about [laughing]?

P4: But then you go back to the old name, old spotted dogfish… "Oh yeah, yeah, I can get you loads of those. What the hell's a catfish?" "It's the same thing as a dogfish but they have been classified differently." So I get this problem a lot, yeah, and I mean probably if you speak to some of the more experienced anglers that are really into their taxonomy, because I do quite a lot of work with anglers, they just get totally confused as well. You know if I'm getting information out of people and you know, talking cross.. so just another example because I did quite a lot of work in South Africa, did some work on great whites and the guy who was running the project said to me, "Can you run down and get me a soup fin shark?" from the fish mongers, and I was thinking, "What the hell is he talking about?", so I went down [inaudible] I've been asked to ask for a soup fin shark, I have no idea what it is, and he went no worries, I've got some in the back and he comes out and I was like "Ah, that's a tope", if you'd said tope

I'd have known what you mean, and yeah so yeah because this this whole thing is so crazy, then you fall back on the Latin term, and then that's wrong, Salmo trutta that's a common one. It's just a common one, the one globally that's probably confused… but certainly in the UK that's not a problem because that's one of our sort of main important commercial economic species, so that's not a problem. But certainly when I talk to other people across the world about this, or sort of have papers back, that have kind of questioned my naming of that species.

Sandra: Any why is that?

P4: Just because they call it something different. Or their taxonomy of it is slightly different. So it's a bit weird, it's a weird thing. We're always fighting this kind of… you know and then also we talk quite a lot, I mean I don't know all the Latin names of all the fish, but you know, fair enough there are 1000s of them…. But you know, we, kind of… there are lots of funny names that people use for there, you know they are very kind of, they are sort of abbreviated names, and also, it's just random, you know you have to fight to get the information out. And say right, what species are you talking about. So yeah, I do get it.

Sandra: So you weren't surprised about the number of variants. Concentration of variants was actually more stable than this picture suggests. But lots of variants used – both on the scientific and vernacular level.

P4: Yeah, yeah, definitely. To be honest with you, quite often when I've been writing papers, or often when I've been writing proposals. I mean for proposals it's not so bad, because it usually goes to a generalist board, so I am not so concerned. But I'm looking at a paper, and I don't know exactly what that fish is. Then I go back and I go OK, hold on a minute and find something that I'm not actually aware of. And normal with the species I know quite well and you're reading a paper and you are like, what's that and then you find out it's actually the species you are working on or that it's something totally different has been called the same thing. And that's a real issue for us. It's just getting to the bottom, I mean to be honest with you, most people are working on such a select number of species. So they just become familiar with those species. It's like what I was saying, I have been writing this paper about these two shark species. Just how diverse the nomenclature is for these, that surround this species. [inaudible] I was actually quite embarrassed when he came back with so many that I got wrong. And, you know I'm a shark biologist… so…

Sandra: But it's got to be really hard, because if you have all these different resources, and everything is constantly changing…these all need to be kept up to date to, I found it all fascinating.

P4: yeah, yeah. This is also a problem, I guess, in that when you are citing literature, if you go, if you go with the most up to date name in the actual piece that you're working on, but then you reference back, they are two different species names. So the title of the article, this is the problem I had, I went back to the editor, to ask them what to do. Because you're telling me that that species name is now wrong, but the citation, or the reference is now telling me that that was that species. How to reference that back, I've actually gone back to author's previously, when I have seen this, where people have, and I've thought, oh, that's funny, that paper isn't about that species, and you get back to them, and they say, oh no, I did check it but the nomenclature has changed and basically that's no longer valid. So that's really bloody confusing now because you know, I'm working off an older name or not marrying the two up and I've found that quite a challenge when I'm writing.

Sandra: Yeah, it's actually something that I saw quite a lot here. When the instances of variants that weren't the accepted variant, according to the resources I looked at anyway, they would appear in the references, rather than in the main part of the document. [Thinking from different perspective – I had the variants – hadn't thought of the difficulty in matching]

P4: Yes, it's very confusing. And yeah you think, that's interesting, I'll have to read more about that, that species, you know? But that threw me a little bit. I'm a bit more clued up on it now, but a year or so ago that was a bit like, ooh. So I started actually emailing people and asking the questions, you know, what is this, but what then you do a bit of research and you're like, well that's an old name or… because sometimes it gets really difficult to find. For example I had this paper from like the 70s and things have changed quite a lot from there it was one of the first feeding, one of the first proper feeding studies that had been carried out on this group of sharks and it's actually really difficult to find the name of the shark in any literature because it had changed back in the 70s, you know? But that was quite a challenge. And when the guy told me I was a bit, I don't know what this is really, because then you've got a Latin name, that's linked with a vague common name, in Australia. And that means that it's actually very difficult to marry those two species together. So, the fun of taxonomy.

Sandra: Yeah.

P4: You can see why my students struggle so much, can't you?

Sandra: Yeah, I've got a friend studying marine biology now, and she has a lot of fun with taxonomy.

P4: [illegible] and get blown away by it. I mean they think we're amazing when you start spouting these Latin names. And you say it's only because you work with them every week. And they're all like how do you know all this stuff? It's like calling your kids' names, you know them because you use them a lot, you know? But let them be fooled.

Sandra: Concentration of hits (frequency) – more stable than the plain frequencies would indicate [explaining the sheets]

P4: So that Salmo… Salmo gairdneri, that's the one that caught me out when doing my first trout paper. It caught me out. That's the one that I thought was a different species and I thought it was a different species, because it's not a Salmo and the editor came back to me and said, this is the same species. So I felt very stupid. But I'd given is as a separate case of another species. So that one always jumps out to me when I see that.

Sandra: Interesting thing – spelling. Resources state: Salmo gairdnerii – but more frequent with one i.

P4: Yeah it's mad isn't it, when you look at that the amount of people getting confused, you know, and you're talking about a species. And but I think you know, there's quite a lot of discussion among scientists about you know the naming of this, of where it's derived from and can we not just get rid of it, you know, can we not agree on a specific name, you know, certainly in the trout Salmonid world, where you're talking about salmon and trout is that they are so commercially valuable and, um, I've had lots of conversations with people from trout trusts and you know everything, because people can recognise what it is, but actually we're not talking about the same thing. So when you're, we had an instance when we were talking with a Japanese company about getting some rainbow trout in and they were talking about different species, or they had a different species in mind. And it thoroughly confused us. We were like, ummm, no we want Oncorhynchus mykiss, and they are, "but these are rainbow trout", but "are you calling rainbow trout what we are calling rainbow trout?" because the name is different and then when we looked at it it was the same species but I can't remember what they called them, this Japanese company, but it was just a nightmare, you know. So normally, you know yourself, we would fall back on the Latin names to know what we were going to buy, and when the Latin name is not the same, you know, you say, that's not right, and "yeah yeah they're rainbow trout", "no, no, no".

Sandra: That's interesting. So you are saying that even with Latin names, there is a geographical element to what Latin names people might use?

P4: Yeah, and also it means that you sometimes don't know if they are just different species. So yeah, you know, we were just very confused about that. You know, because this guy was actually flying over from Japan to come and work with us. And we were looking at these farmed fish, because we farm rainbow trout here quite widely and we were looking at the impacts of those farmed fish and how they survived in the wild and how they kind of survived downstream of fish farms and he was going to bring some eggs over, like some different strain of eggs over for us, he was going to bring what we thought was Oncorhynchus mykiss and it was but not by name. So the more work in it, the more you try and talk to people, the more you try and purchase things… you know the um, it's not fish but just going back to the shrimp [that he had had to remove from the lab because of the coronavirus outbreak] and yeah, they're a nightmare. We still don't know whether we've got the shrimp that we wanted. They are called cherry shrimps, but the variation we have got across the shrimp that we got, we're not convinced that they are all cherry shrimps, that they are all the same species of shrimp, because they are so different. But because they have this red pigment, but I mean they varied from anything to almost translucent with a red hue to you know, a very vibrant, deep deep deep red, and my colleague and I just said, you know, they don't even look the same, you know, but these are classed as cherry shrimp, and I can't remember the name of them, we've only just got them in. I can't remember the name of them. And we were like, are these even the same species? I mean they don't look like it but they have been classed as that, because, you know, they're red.

Sandra: Yeah, so do you have this issue of people in different areas using the common name, but you are not sure exactly what it is, and then sometimes you ask them and…?

P4: I mean I still don't know what species it is. I mean, these shrimp came through an aquatic wholesaler, like an importer, who deals with like, I mean they are very well-known shrimp wholesaler and we've used them for years, but I'm still convinced that they don't know what they are. I think they just get shrimp that are a red colour.

Sandra: Yeah and they go right and say, that goes in that one.

P4: Yeah, it's crazy so we're still not convinced, unless we do genetic testing on all of them, which, you know, we're not going to do. But it is a worry, because actually, we are publishing stuff that probably has an element of error in it. We're not convinced, we are reporting poorly on a species and that's kind of held us back from some of the publication because it's not always clear, but you know. But that doesn't surprise me having looked at is seeing all these configurations and variations of the trout. I think the biggest thing with the rainbow, is the change from the Oncorhynchus to the Salmo. It's the biggest thing, because I've seen them call Salmo all sorts of things and actually they're rainbows. But, yeah, they're not, they're not a Salmo species. [inaudible]

Sandra: People do what they will.

P4: Yeah, indeed.

Sandra: Sander lucioperca: infrequent in the corpus. But inverted frequency of accepted name and variant.

P4: I do know zander, I have not done a lot of work on them, but I do know that the pikeperch is the one that sticks out as the most, that's what we call them most, that's what we call them, to be honest I call them zander. But then again I think it's the typical thing, people call them pikeperch and then get totally confused, because I don't know if you looked but there are another two species, pike and perch and people get very confused about what they are. They think they are hybrids of those two species, and yeah, that gets very confusing. Because I used to teach on a fisheries degree, over in Hampshire and these [inaudible] anglers, who basically wanted to become fisheries managers and had an interesting conversation around these types of naming, you know, sort of activities of pikeperch and whatever, because I just know of Sander lucioperca as the Latin name, but I would normally call them zander or pikeperch, that other bit, I'm not sure what that other the St…

Sandra: Stizostedion… or however you say it.

P4: I don't know, I need my glasses, or I need some glasses, it's tiny on my screen, but um, I've never heard of that.

Sandra: I think, if I remember rightly, it was classed as a previous name that then they decided that the genus wasn't the correct genus.

P4: Yeah, they reclassified it..[inaudible]

Sandra: They reclassified it. Yeah, but I mean… it's interesting what you said about the pikeperch, I hadn't thought about that confusion.

P4: Yeah, it's funny the naming. I mean, you know, just going back to the sharks, because I've done more work on those, you know, when you're talking about Great Whites and they start calling them Bronze Whaler Sharks, you know, and everyone is calling them White Sharks, White Death, Great White Shark, and then somebody, one country comes out with Bronze Whaler. And that confuses everybody, because in your mind, why is it bronze and a whaler? And it doesn't make sense. So where that's come from, is just totally bizarre. I mean whether that's a translation that's just gone wrong, and that isn't quite right or something else, but it's that kind of naming, the naming of pikeperch, it's just weird. And that's where the confusion comes from. I mean quite a lot of my students in the past have been [inaudible], they are big predatory fish, and they want to catch pike or perch and there is this perception that these individuals are not zander, it's crazy.

Sandra: Very funny, I think Rachel mentioned one which was a mountain chicken that's actually a frog?

P4: [inaudible] with some of these names you are like, “What?” [inaudible] Weird, weird.

Sandra: Same sort of thing with Salmo trutta. [I explain the graph of the nodes, the separate of entities rather than words, and the arrows for parent to child relations] A lot of the ambiguities I identified were linked to the vernacular usage. Sea trout: parent versus child.

P4: Basically with brown trout, nobody knows why they turn into sea trout at all. Scientists are still unsure but they think it is to do with competition in the riverine system, this is just one theory, I don’t know if it’s right [inaudible] but, um, the thing is that they are basically pushed out and they transform to go out to sea. There’s no real reason to do that…

Sandra: So nothing to do with breeding, or anything like that?

P4: Yeah, I know they call them anadromous but they are not truly anadromous because they don’t need to do it, but they basically… so to me, what I was going to say to you, in relation to your graph, I would say that brown trout is Salmo trutta as a pair and this is an actual offshoot of the same species. They don’t speciate, they are still essentially the same species genetically they are the same but it so happens that some do go to sea and some don’t. So that’s where… but again they look very different, if you look at a picture of a trout, of a brown trout and a sea trout, they actually look very different. So they are very silver, it’s like salmon silver up to go out to sea, and brown trout silver up to go out to sea as well. And actually they change shape slightly to [inaudible] so you kind of get this very bizarre, you can understand why people get confused, because physically they are very different, but genetically they are still brown trout. They come back into the rivers, still breed the same…

Sandra: But they change quite a lot physically when they are out… ahhh.

P4: Yeah, Google sea trout and brown trout and you’ll see the difference, they are quite physically different but yeah. There again, this is pure confusion [inaudible] are these, they’re a Salmo species, like Salmo salar, salmon are they a typically anadromous fish that have come back in and basically found a niche in a river and now they don’t need to migrate and go back out to sea? But a few of them do? Or is this an adaptation of a riverine species, you know. And I guess that’s the debate, no one really understands the reason why these fish go out to sea.

Sandra: That’s fascinating, I didn’t know any of that. Because I noticed the same thing with the rainbow trout and the use of steelhead.

P4: Yeah, steelheads, yeah, a similar thing.

Sandra: It’s the same thing?

P4: [inaudible] There’s a really, really interesting case actually the walking… the catfish. The riverine catfish, where they originally thought there were 30.. I think it was something like 32 species of catfish they found, but then when they actually narrowed it down, there were only about 9 different species. I mean they look different, because some of them were darker, bigger, blacker, and then when they actually did the molecular testing on them, they whittled it down to about 9 species, they reckon. And you know this is a typical case of that. You’ve just got species that just behave differently in different habitats. You see that a lot in other areas. You’ve got crayfish that in normal riverine systems they are quite clear and light coloured and then you get these random black ones that live in silt and they just grow bigger because they have more detritus and it’s kind of black and then people think they are different species. So they call them something different and … [inaudible]

Sandra: then it sticks

P4: yeah, then that's what they are, it's really weird. Like you say, it's funny, like the mountain chicken, [inaudible] I mean apparently to us, when we're talking about these species, apparently to us, there's no real reason why they are called these things [inaudible] but you know with some of these fish you're like, why is it called that? Why has anyone ever come up with that name. You know, it's weird.

Sandra: Why, yeah? Cos if, for example, there's something about the appearance of the species that helps you to identify it through that name, then it seems to be a useful communication tool in that, but when it doesn't have anything to do with it..

P4: Yeah, exactly it's totally random. Because, you know we were talking at the [inaudible] and the guy who was running it was a really well-known fish expert [inaudible] and he was chatting away about these things and um, fat-head minnows, they're called. And basically all it is is that, because I was like, why are they called fat-head minnows, weird, but apparently they have a little fatty lump on the top of their head. You can't see it, so some of them you can [inaudible] but if you are just looking at them, why? But they have a little lump of fat on their head. And you're thinking it should have a big fat head, you know. But no. [inaudible]

Sandra: This was just one, this might be a sort of example, with the Japanese. [Parasalmo mykiss and Kamchatka, geographical specificity]

P4: Yeah, Yeah.

Sandra: Here I had identified steelhead trout and rainbow trout were both very frequent common names used in relation to Oncorhynchus mykiss. Synonymous exactly or not? Looked at whether they appeared in the same documents, same parts of documents. Saw they were often used in combination. Also, looking at other descriptions (future work).

P4: So are you saying that in your lit review you found a combination of those names in the same document?

Sandra: Yes. A lot. Let's see if I can go back.

P4: So that's the number of times that steelhead trout was found with rainbow trout in the same document.

Sandra: Yes, out of 47, where's the 47 from, yes, out of the 47 which would have been the number of times that steelhead trout appeared, because rainbow trout, rainbow trout was by far the most common, across the whole thing…

P4: Maybe part of that, may be that if you are working on a population and if the rainbows are doing the same as the brown trout and the majority of them are staying and some are migrating out, it may be that they are classifying what proportion of that is actually seaward and trying to identify which of those they have sampled actually go to sea, which are steelheads and which are rainbows maybe. I was just thinking actually as you were talking if I was writing a paper on brown trout and I mean I have actually written reports back to different organisations when I've done [inaudible] fishing with students and go out and catch fish and do surveys, I would probably identify if I found sea trout by calling them sea trout. So that would be an assumption that the people I was reporting back to knew that they were the brown trout that had gone to sea. So it's possible that maybe that's what they are doing in the literature. I guess, I mean I don't know because I have worked with rainbow trout a lot but in aquaculture. [inaudible] But it took me years to realise that steelhead when they went seawards. So that is probably, what I imagine that is what is happening in the literature.

Sandra: Definitely. And so with things like sea trout, is sea trout exclusively used in relation to brown trout?

P4: Uh, yes. Well, as far as I know, that's the only time we ever use sea trout is when brown trout have gone to sea. Because we only have two Salmonids, well we have other Salmonid species here that we work with, the two main ones are Salmo salar, which is the Atlantic salmon and Salmo trutta which is brown trout. And basically, because we use that distinction to distinguish between, because they both go silver, the salmon and the trout, so we use that to distinguish between the salmon and the trout basically. So if you look up on Google you just put Salmo trutta and seatrout, sorry you look up Salmo salar and sea trout, you will get a images of the two together where they look very similar and there are diagrams of the differences and stuff so…[inaudible] my thought would be that, just trying to distinguish between [inaudible]. I might be wrong, I might be wrong.

Sandra: That makes sense. And just, like, with salmon and trout, as common names, are there specific differences? Because I know that looking at the genera and things then within a genus then you have lots of names which are called salmon or trout, it doesn't seem to be split by genus.

P4: So yeah [inaudible], everything is further back. So on the mouth, if you look at the side of a trout, the maxilla extends beyond the eye. And on the trout it doesn't. And also on the tail, on a trout it's sort of pushed back straight, can you see, it's pushed back straight. Hold on my camera's here. And on the trout, I mean the salmon, the fins are more like forked. So yeah, you've got a fork in them and basically [inaudible] the lateral line, the line that runs down the middle of the fish, the spots on the trout actually extend, they are usually much more defined, but they extend below the lateral line quite a way and the salmon doesn't. So I always tell the students, you know everything is further back on a trout. So yeah down that way the maxilla comes back beyond the eye, the tail pushes out and everything on the salmon is further forwards. So yeah that's the main way that we identify usually between salmon and trout. I mean, there are other things, I mean fin sizes and shapes, and that's what defines the genus. And normally trout are quite fuller bodied whereas salmon are much sleeker. They have a much more aerodynamic-shaped head. And basically, but I mean you know, for all intents and purposes, I mean we had somebody, I mean I was doing some salmon tagging when I first started working with salmon and I had someone from the environment agency come down and say oh look that's a nice trout. And she told us she'd worked on a salmon farm for two years before she'd come into the environment agency. I mean these are farmed salmon but for all intents and purposes they're not trout. So, you know. And I can get it wrong. I have just got a PhD student's report who is handed in today and he has got a picture of what I am sure is a salmon but he's called it a trout. And I'm like, that looks like a salmon but until, because his imaging was so bad [inaudible] collapsed. I can't really, it's hard to tell I mean even we're struggling [inaudible]. I mean he's working with them all the time, I'm sure he knows what it is but yeah so it's really weird.

Sandra: Nothing's clear is it?

P4: No. Nothing's clear. [inaudible] To be honest with you, quite often when I've done these things, and I have reported on surveys for environmental work, we did some work on a stocking thing and you're always a bit nervous that you got it wrong and that you are actually working with the wrong species. So you always hold back some papers where you're not quite sure what it is we've got, you know. It's a bit tricky.

Sandra: And then there's just.. there were a couple of contradictions. They were all related to the Catalogue of Life and the inclusion of common names, that when I looked at the data it didn't look like they should be included in the taxa that they had been included in. So with brown trout, it was included in Salmo trutta but it was also included in the Oncorhynchus mykiss, which like, I mean, obviously in the data then it showed that it wasn't linked, but it that something that surprises you or not?

P4: No, at all. It's what I said to you before, I think the rainbows and the browns get confused quite a lot, like what I was saying about the Oncorhynchus mykiss is not a Salmo. And often in the literature it is referred to as a Salmo something. And that's a huge sense of confusion for people because we are working with a totally different genus here. So certainly, I mean I'm not going to talk about the work in the UK that we have done, but we do have farmed rainbow trout that are Oncorhynchus mykiss but basically they escape into the wild and then people catch them and people will think they are, well in the literature they will go and find Salmo and they will think they are a Salmonid, well a Salmo genus and they are not, so… yeah it doesn't surprise me because I've seen it, I've seen it before. And certainly if you are talking about literature from countries you know don't have rainbow trout in them. Sorry, they have rainbow trout but not brown trout and they're calling them Salmo something, you're thinking they don't have those species there. So it's no surprise.

Sandra: And that was just in the other corpus just to show the same thing came up. And here if you can see it is, so also, so in the Salmo trutta then terms both brook trout and lake trout were included, but then I saw that these and in both of the corpora they actually linked brook trout with Salvelinus fontinalis and lake trout with Salvelinus namaycush.

P4: Yeah, I give a lecture on invasive species and there is a really famous case in Yellowstone lake, where the lake trout, I think they introduced lake trout and then they pretty much wiped out the brook trout.

Sandra: Oh wow, yeah.

P4: And when I was reading it to give the lecture, the two names were used interspersely. Basically there was so much confusion about those two species, so it was actually quite difficult to get a feel for what [inaudible] which one was invading which. So, um, and again, I've got, one of them came up as the Salmo gardinium, or whatever it is, and there was another one that was… I can't see the Latin name there… Salmo… clarkii? So that came up as the lake trout species as well. So I was like, bloody hell how many species are in this lake, you know? So yeah, I'm familiar with those species and that kind of thing. So yeah, one paper called it brown trout and they are not brown trout. Not as we know them. They are not even Salmo genus, so…

Sandra: Such a confusion.

P4: Yes, it is.

Sandra: So that is just the same: you have the lake trout linking with the Salvelinus namaycush and the brook trout with the Salvelinus fontinalis. And it not being linked to Salmo trutta, which is what the Catalogue of Life… what I haven't check is, well the Catalogue of Life mainly, well they have people curating it but also there are links to where they've found these references to, because the Catalogue of Life is very all inclusive, well trying to look at all of the times that this species has been mentioned but I don't have access to a lot of the things. I mean the one with brown trout coming up in the Oncorhynchus mykiss I think was a paper in Nepal, which matches up with what you just said about how…

P4: Yeah, yeah yeah. Yeah, because I've had the same problem you know, where you are looking for species, you think you've found something and actually you are looking at a geographical location and you are thinking, that can't be right. I mean I know that we're, certainly for the brown trout, I know where they extend to. You're kind of thinking, this just can't be. So you end up sort of discounting literature because you're not sure if you would be including… For example the example that I gave with the paper and the guy came back to me and was like can you give me an example about the same species. So I had to change the discussion around that. But that's interesting, so. It creates somewhat of a headache, I can tell you.

Sandra: I bet. Because, a lot of the time there isn't actually a way round, because I mean if you don't have the physical things, it doesn't matter what research you do, you can't find out what they were talking about.

P4: Yeah. I mean also, the biggest problem and also it's something I've actually, when I've written papers previously, I've actually excluded literature from it, because I'm not sure, I've thought, actually I've got no evidence, I've got no proof, there's no images to be able to say actually that is Salmo trutta or that's Salmo salar or whatever. And actually what people are calling these things is just random. Random and quite often, I haven't got a nice example, ut I'm not convinced either, whatever you're talking about, where I know the species range, isn't there, because it wouldn't survive in that environment, it's too warm, you know. But that's a different trout species, but I don't know what that or sharks... I know geographically those sharks don't exist. Where they've called them something where you know it isn't actually capable of living in that environment, its range doesn't extend to that point. I mean it's not that often, and I not that great a fish biologist that I know where all the species exist but there are a few where I know enough about that species to know that that isn't right. And somebody much more aware of kind of, because I don't really look at taxonomy so much but in toxicology when we buy things in for the lab, to do tests in the lab, we get them from accredited suppliers if they're fish and we know 100% that the fish is Carpio, for carp or whatever, or Salmo trutta. So that way we sort of have that confirmation but I'm sure there are people that I've spoken to that, there are a few people, I mean one of my friends has described a few species of fish and he just knows naturally, he knows oh, that's wrong, that can't be that fish. You know, we've sort of written a few papers together and he's kind of said to me, look, that's not the right species. So yeah, it can be very difficult pinning down what people are talking about. And actually, because there's no control over it, with nomenclature across the world, people are pretty much freewheeling it, I mean mostly Salmo trutta is pretty much the accepted brand here but you across sort of the Atlantic and so many people are using it and confusing it with other species that are related to America. I mean we don't find brown trout, well we do find it there but not in the abundance that normally of other species. SO yeah, it's interesting, it's a really interesting thing. Headache for you, trying to work it all out.

Sandra: I mean in a way, I've got the… I'm trying to work out how it's being used. I can't tell anything about what they actually meant, well the actual physical meaning of what it is so I mean, but yeah. But I mean from what I showed you, do you think the representations can go some way as to identify where there are ambiguities and clarifying some of the ambiguities in the usage of the terms?

P4: Yes, I mean in terms of what you're doing yeah. Yes, because I think there is real confusion. And quite often I've seen, because I mark so many pieces of students' work as well, I've seen that coming through from the student's perspective as well. And I think what really needs to be done, and I think part of the problem is coming back to the catfish scenario that I told you about where actually if you've got something that looks different it is different and actually quite often, they're not, they're just the same species. It's like us isn't it. I mean you go to Africa, you know, OK we've got the same appendages and so on but we look different. You know and if you go to Alaska, you know, and the inuits they look different from a Kenyan.

Sandra: Yep.

P4: You know, but we are the same species. And I think I try to get that across to my students. Their default is always colour and size. They are the two worst characteristics you could use, I mean look around the room, we're different colours, different sizes, all different shapes, and we're the same thing. But I think if you, I think that's the confusion. But what I think really needs to, some clarification is really needed on you know what… because there are mistakes being made. I mean there are people testing species out there that are reporting testing on species when they are not using the right species and that's actually misinformation, you know. We are talking about things, that effectively, you know, because fish do react different depending on where they are found. I mean you know, I'll give you an example, tilapia for example, they are really tolerant of all sorts of temperature but then there are different species, um, noroscus and mossambicus and actually they react slightly differently to different stimulus, one's more able to sustain itself with no/low oxygen and higher temperatures than the other. SO testing on those two species, they don't look similar but if you're testing on them you are going to get two totally different results. So yeah, I think there is definitely a need to try and bring this together in terms of you know people need to be more aware and I think we need a more reliable resource to go to. I mean, I say the one I tend to favour is FishBase because I'm familiar with it… how accurate it is I have no idea and actually if you look at each fish entry it can actually be quite confusing and if you go country by country sometimes you see, you're like wow even some of the Latin names are different. And there needs to be some clarification on, you know I mean these parent names are really important, the synonyms are probably less so but they are useful to go OK that's a species, but I have found that you know, where I've known about a that species has been used incorrectly, they don't appear on that FishBase. And that's a problem. So there we go back to where I have left out papers from the literature because I'm not convinced about what it is. Going back to the example of the paper where it came back and I'd given the example as another species and they said it's the same species, I felt like such a fool, I'm very careful about what I use in terms of species now. Yes, I think there's definitely scope to put something in place. I mean how you do it, I have no idea. I mean I don't know who controls all this.

Sandra: Many different people all over the place! One of the many issues I think… Oh and there was something just in your, in the questionnaire you filled out for me. You said that you used resources to annotate data.

P4: Did I?

Sandra: I think so…

P4: Oh did I? What do you mean, or what did I mean?

Sandra: I asked if when you use taxonomies or ontologies, taxonomic resources, if you use them to check name variant status or this and there was one about.

P4: Yeah, maybe it got misinterpreted but I do go back to FishBase and check all the synonyms, is that what you mean? And I just check back… I mean it depends what I am writing and what I am doing. But certainly for papers, that's how I got caught out with that shark paper, because I thought I'd checked it all, and I'd gone back and basically I'd checked, because there were some sharks, that I thought, oh, I think I know what that is, I think what highlights it is when you look at the genus of a species that you're working with and you think, I know where that shark should fit. [inaudible] aren't that many sharks out there, I'm dealing with lots of them. I go back and check the junior synonyms and some of them I've found are actually considered a synonym of this species. Some of them didn't exist and they're the ones I got wrong, but because they're so old, that…

Sandra: They're not there anymore?

P4: They're not there anymore, yeah. And they're the ones I got caught out on, so yes, I do tend to do that. I think my problem is I don't quite know what to do with it, do you know what I mean. So, do I just wrap them all up in the same name? Or do I just put something in there to say that this species is now referred to as…[inaudible] because I am always just nervous about, about, nobody ever tells us what to do. And every journal comes back with a different….

Sandra: I was going to say, everybody tells you something different? Because different people have different ideas and also…depending on the way, I don't know, this is just how I see it, but depending on what you're doing a different approach might be appropriate? But if you don't have a, a homogeneous approach, an approach across the board then you're still not going to be able to bring everything together.

P4: Yeah, yeah… depending on, I think it comes back to, we're doing lots of work on [inaudible] but the word we kept using was standardisation, standardisation. There's no standardisation. Everything is so different, you can't compare. People are using different methods, different even the same chemicals, they're using different tissue, different temperatures, different concentrations, different timings, you can't then compare that against, you know. And it's the same with this. You can't, it's actually very difficult to actually marry two papers up when they are using different terms. It's the same thing like with steelhead and rainbow, even for me. I mean I've worked with trout for what, 15, 20 years now, and, well it still confused me. It took me years to realise they're the same bloody thing. To start with I thought they were a different species. Not that I needed to worry because I don't really work with rainbows, but it was only when someone said they're just the equivalent of sea trout. Oh yeah, I thought they were a different species, and uh, yeah, until someone kind of gets that and there's that uniformity in what we do with the information then it's going to remain this kind of jumbled mess of confusion for everybody. I mean students always ask me about taxonomy, and get it wrong, and because of my colleague Anya, and she's very, she's German, she's very German, she knows she's very German and she's like – you have to know this, this, this, and I did this in my degree and I did this, and all that. And we're like, Anya, me and my colleague who's deputy head of school, Anya, just lay off them a little bit, you know, they don’t know taxonomy, it’s confusing for us, it’s confusing for them. You know. Anya’s like, these are the next generation of scientists… yeah but there’s nothing actually in place to allow you to grasp this with ease. You know, it’s only when you start working with it, that you get a better feel.

Sandra: Feel, because…

P4: And trying to get this into students’ head when they actually don’t give a shit whether it’s a Salmo trutta or a Salmo whatever else, but it becomes a bit more real when they use it. Nightmare. I mean, my view on it, is I submit it, and if it gets picked up, it gets picked up. And I’ve stopped worrying too much about getting it 100% because normally the journal will pick it up and…

Sandra: Yeah, you do your checking and leave out anything you’re unsure of in the actual thing and then see how…

P4: Yeah, and also, we’re submitting to quite a few American journals, or they’re international journals. We just put one into um, it’s a review paper, we’ve just put it into an Asian Fisheries Journal, from, they’re publishing in Asia. And their view on, I mean they sent some stuff back and we were just like, “what?” like, what is going on here? I mean we’re not comfortable, we’re not comfortable with that, but we’ve done it and resubmitted it. But the way they were asking us to present the Latin names, wasn’t quite how we would present, and we were like, no, we’re not doing that, but yeah, just the way they wanted it presented, and you know, it was just like really weird. Abbreviated species names as well they suggested, and…

Sandra: What? It’s the opposite, right?

P4: It was like, what are you doing? So yeah, [inaudible] it’s a minor journal, it’s a review for a MA thesis, so, yeah, he’s quite happy to get through. You know what country you’re working from and what the general requirements are and the way that they want this done and how diligent they are in checking the Latin names, and whatever so… so yeah, my view now is to stop worrying about it, I do what I can, and they can sort it out if they want to and they do generally. I mean every time I publish a paper, I don’t publish hundreds by any means, but I just sit and wait for the torrent of abuse. It’s never happened, it’s never happened yet, [inaudible]. Just one guy [inaudible], and he’s a bit of an arse, but he knows about sharks and said that’s the junior synonym and it needs to be changed so I changed it, but that’s the only time I’ve ever had something come back to me. I think to be honest with you most people aren’t sure anyway, unless they work with that species specifically. [inaudible]

Sandra: I was going to say with so many things you can only really be focused on the thing you know the best.

P4: Yeah, I say to the students, there are 32500 species of fish that are listed. And they are finding new species every week.

Sandra: All the time, yeah.

P4: I mean the value is predicted to go up to, I mean with all the deep-sea exploration, I was speaking to someone yesterday who predicted that there are 1 million new species down there.

Sandra: Wow.

P4: So how are we meant to know that. I mean my students always tell me, “wow you know so many Latin names!” Yeah, I know the Latin names of the things I work with, you know. If you ask me what’s the Latin name of a you know, of a [inaudible] crystal mouth, you know [inaudible] If you shout out a few Latin names occasionally they think you are a genius. [inaudible]

Sandra: But it’s the fact that everything is so different. So if you had something that could map at least the way that different terms in at least authoritative resources and seeing if geographically there were different patterns in the way things were being used, would things like that be useful?

P4: Yeah, very useful. I mean I think for me you know, coming back to it, always, I don’t care what common name they use now, all I want to know if the Latin name is sound, and actually the Latin name isn’t sound. And that’s the underpinning, and I’ll tell my students all the time. It doesn’t matter, why do we have taxonomy, why do we have taxonomic names, this nomenclature that basically tells us, because there’s such discrepancy in the common names.

Sandra: Yep.

P4: We should be confident when we are talking about Latin names and species. And I understand that not every species has been genetically linked, so there will be constantly changes as the genetic analysis of these animals goes on, but that’s fine. I can deal with that. It’s just when you see this massive discrepancy against Latin names globally. And actually, you know, that actually causes a lot of confusion. And I think misinformation out there on what those species are doing. You know, what what they are. And certainly because I’ve worked in aquaculture, you know there are a lot of discrepancies in aquaculture, as well in what species they are farming, unless they’ve been translocated out the country, um put somewhere else, there are very different techniques for different species. And that has caused some issues as where people have tried to pick up techniques for a species that they relate as the same species that have been farmed, and then gone back and it hasn’t worked.

Sandra: Because it’s not the same thing…

P4: It’s not the same thing, yeah, but they’ve called it the same thing. That can be common. Yeah it’s all fun. But yeah that would definitely do it, if you see how they derived those names that would be really useful, because that would reduce the work we would have to do working that down. But anything that would help to understand that.

Sandra: Understand that.

P4: It would be better and would make our lives a whole lot easier.

[END]

F.8 Evaluation questionnaire results

Please refer to Focus group folder, file name “responses_2_eval.csv”.

Bibliography

[1] IUCN 2020. The IUCN Red List of Threatened Species. Version 2020-2, https://www.iucnredlist.org.

[2] Mohammad Fikry Abdullah and Kamsuriah Ahmad. The mapping process of unstructured data to structured data. *International Conference on Research and Innovation in Information Systems, ICRIIS*, pages 151–155, 2013.

[3] Lakshmi Manohar Akella, Catherine N Norton, and Holly Miller. NetiNeti: discovery of scientific names from text using machine learning methods. *BMC bioinformatics*, 13(1):211, 2012.

[4] Abeer Al-Arfaj and Abdulmalik Al-Salman. Ontology construction from text: challenges and trends. *International Journal of Artificial Intelligence and Expert Systems (IJAE)*, 6(2):15–26, 2015.

[5] Najib M. Ali, Haris A. Khan, Amy Y-Hui Then, Chong Ving Ching, Manas Gaur, and Sarinder Kaur Dhillon. Fish Ontology framework for taxonomy-based fish recognition. *PeerJ*, 5:e3811, 2017.

[6] Vidhya Analytics. Analytics Vidhya.

[7] Erick Antezana, Martin Kuiper, and Vladimir Mironov. Biological knowledge management: The emerging role of the Semantic Web technologies. *Briefings in Bioinformatics*, 10(4):392–407, 2009.

[8] Laurence Anthony. AntFileConverter (Version 1.2.1) [Computer Software], 2017.

[9] Laurence Anthony. AntConc 3.5.8, 2019.

[10] Aristotle. Topics.

[11] Muhammad Nabeel Asim, Muhammad Wasim, Muhammad Usman Ghani Khan, Waqar Mahmood, and Hafiza Mahnoor Abbasi. A survey of ontology learning techniques and applications. *Database*, 2018(2018):1–24, 2018.

[12] Muhammad Nabeel Asim, Muhammad Wasim, Muhammad Usman Ghani Khan, Waqar Mahmood, and Hafiza Mahnoor Abbasi. A survey of ontology learning techniques and applications. *Database*, 2018(2018):1–24, 2018.

[13] Gerard Escudero Bakx. *Machine Learning Techniques for Word Sense Disambiguation.* PhD thesis, Universitat Politecnica de Catalunya, 2006.

[14] Caroline Barrière. Knowledge-Rich Contexts Discovery. *Seventeeth Canadian Conference on Artificial Intelligence (AI- 2004)*, 3060(May 2004):187–201, 2004.

[15] Steven J Baskauf and CO Webb. Darwin-SW: Darwin Core-based terms for expressing biodiversity data as RDF. *Semantic Web 7.6*, 7(6):629–643, 2016.

[16] W. G. Berendsohn. The concept of potential taxa in databases. *Taxon*, 44(2):207–212, 1995.

[17] Walter G Berendsohn and Marc Geoffroy. Networking Taxonomic Concepts — Uniting without 'Unitary-ism'. *Biodiversity databases - techniques, politics, and applications*, (April 2007):13–22, 2007.

[18] Tim Berners-Lee, James Hendler, and Ora Lassila. The Semantic Web will enable machines to COMPREHEND semantic documents. (May):1–5, 2001.

[19] Douglas. Biber, Susan. Conrad, and Randi. Reppen. *Corpus linguistics : investigating language structure and use.* Cambridge University Press, 1998.

[20] S. M. Billerman, B. K. Keeney, P. G. Rodewald, T. S. Schulenberg, and (Editors). Birds of the World., 2020.

[21] Robert Blumberg and Shaku Atre. The Problem with Unstructured Data. *DM Review*, 13:42–46, 2003.

[22] BNC Consortium. The British National Corpus, version 3 (BNC XML Edition), 2007.

[23] Tolga Bolukbasi, Kai Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *Advances in Neural Information Processing Systems*, (Nips):4356–4364, 2016.

[24] Virginia Braun and Victoria Clarke. Qualitative Research in Psychology Using thematic analysis in psychology Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2):77–101, 2006.

[25] Rosanna L. Breen. A practical guide to focus-group research. *Journal of Geography in Higher Education*, 30(3):463–475, 2006.

[26] Christopher Brewster, Harith Alani, Srinandan Dasmahapatra, Portebello Street, and C Brewster Y Wilks. Data Driven Ontology Evaluation. In *International Conference on Language Resources and Evaluation (LREC 2004)*, 2004.

[27] Christopher Brewster and Yorick Wilks. Ontologies, Taxonomies, Thesauri: Learning from Texts. In *The Use of Computational Linguistics in the Extraction of Keyword Information from Digital Library Content Workshop*, pages 1–32, 2004.

[28] V. Brezina, M. Timperley, and T. McEnery. #LancsBox v. 4.x, 2018.

[29] Vaclav Brezina. Collocation Graphs and Networks: Selected Applications. In Pascual Cantos-Gómez and Moisés Almela-Sánchez, editors, *Lexical Collocation Analysis: Advanced and Applications*, chapter 4, pages 59–84. Springer International Publishing, 2018.

[30] Vaclav Brezina. *Statistics in Corpus Linguistics: A Practical Guide*. Cambridge University Press, 2018.

[31] Pier Luigi Buttigieg, Norman Morrison, Barry Smith, Christopher J Mungall, Suzanna E Lewis, and ENVO Consortium. The environment ontology: contextualising biological and biomedical entities. *Journal of biomedical semantics*, 4(1):43, 2013.

[32] Pier Luigi Buttigieg, Evangelos Pafilis, Suzanna E. Lewis, Mark P. Schildhauer, Ramona L. Walls, and Christopher J. Mungall. The environment ontology in 2016: bridging domains with increased scope, semantic density, and interoperation. *Journal of Biomedical Semantics*, 7(1), 12 2016.

[33] University of California. AmphibiaWeb, https://amphibiaweb.org.

[34] D. Carson, A. Gilmore, C. Perry, and K. Gronhaug. Focus Group Interviewing. *Qualitative Marketing Research*, (October):113–131, 2011.

[35] Naturalis Biodiversity Center. Catologue of Life: Oncorhynchus mykiss.

[36] Rathachai Chawuthai, Hidaeki Takeda, Vilas Wuwongse, and Utsugi Jinbo. Presenting and preserving the change in taxonomic knowledge for linked data. *Semantic Web*, 7:589–616, 2016.

[37] Ling Chen, Jian Shao, Zhou Yu, Jianling Sun, Fei Wu, and Yueting Zhuang. RAISE: A Whole Process Modeling Method for Unstructured Data Management. *Proceedings - 2015 IEEE International Conference on Multimedia Big Data, BigMM 2015*, pages 9–12, 2015.

[38] Laura Chiticariu and Frederick R Reiss. Rule-based Information Extraction is Dead ! Long Live Rule-based Information Extraction Systems ! In *ENLP 2013 - 2013 Conference on Empirical Methods in Natural Language Processing*, number October, pages 827–832, 2013.

[39] Jerry G. Chmielewski and David M. Krayesky. *General botany laboratory manual.* 2013.

[40] J. F. Clements, T. S. Schulenberg, M. J. Iliff, S. M. Billerman, T. A. Fredericks, B. L. Sullivan, and C. L. Wood. *Clements Checklist of Birds of the World v2019.* Cornell University Press, 2019.

[41] Willem Coetzer, Deshendran Moodley, and Aurona Gerber. A knowledge-based system for generating interaction networks from ecological data. *Data and Knowledge Engineering*, 112:55–78, 2017.

[42] CoL. Catalogue of Life - Standard Dataset, https://www.catalogueoflife.org/content/contributing-your-data#standard.

[43] CoL. Catalogue of Life, https://catalogueoflife.org/col/info/about.

[44] Ronan Collobert and Jason Weston. A Unified Architecture for NLP: Deep Neural Networks with Multitask Learning. In *Association for Computing Machinery (ACM)*, pages 160–167, 2008.

[45] Merce Crosas. A Data Sharing Story. *Journal of eScience Librarianship*, 1(3):173–179, 2012.

[46] Hong Cui, Dongfang Xu, Steven S Chong, Martin Ramirez, Thomas Rodenhausen, James A Macklin, Bertram Ludäscher, Robert A Morris, Eduardo M Soto, and Nicolás Mongiardino Koch. Introducing Explorer of Taxon Concepts with a case study on spider measurement matrix building. *BMC Bioinformatics*, 17(1):471, 2016.

[47] DataCamp. Data Camp, 2019.

[48] DCMI Usage Board. Dublin Core Standard, 2012.

[49] Yde de Jong, Juliana Kouwenberg, Louis Boumans, Charles Hussey, Roger Hyam, Nicola Nicolson, Paul Kirk, Alan Paton, Ellinor Michel, Michael Guiry, Phillip Boegh, Henrik Pedersen, Henrik Enghoff, Eckhard von Raab-Straube, Anton Güntsch, Marc Geoffroy, Andreas Müller, Andreas Kohlbecker, Walter Berendsohn, Ward Appeltans, Christos Arvanitidis, Bart Vanhoorne, Joram Declerck, Leen Vandepitte, Francisco Hernandez, Róisín Nash, Mark Costello, David Ouvrard, Pascale Bezard-Falgas, Thierry Bourgoin, Florian Wetzel, Falko Glöckler, Günther Korb, Caroline Ring, Gregor Hagedorn, Christoph Häuser, Nihat Aktaç, Ahmet Asan, Adorian Ardelean, Paulo Borges, Dhimiter Dhora, Hasmik Khachatryan, Michael Malicky, Shaig Ibrahimov, Alexander Tuzikov, Aaike De Wever, Snejana Moncheva, Nikolai Spassov, Karel Chobot, Alexi Popov, Igor Boršić, Spyros Sfenthourakis, Urmas Kõljalg, Pertti Uotila, Gargominy Olivier, Jean-Claude Dauvin, David Tarkhnishvili, Giorgi Chaladze, Michael Tuerkay, Anastasios Legakis, László Peregovits, Gudmundur Gudmundsson, Erling Ólafsson, Liam Lysaght, Bella Galil, Francesco Raimondo, Gianniantonio Domina, Fabio Stoch, Alessandro Minelli, Voldermars Spungis, Eduardas Budrys, Sergej Olenin, Armand Turpel, Tania Walisch, Vladimir Krpach, Marie Gambin, Laurentia Ungureanu, Gordan Karaman, Roy Kleukers, Elisabeth Stur, Kaare Aagaard, Nils Valland, Toril Moen, Wieslaw Bogdanowicz, Piotr Tykarski, Jan Węsławski, Monika Kędra, Antonio M. de Frias Martins, António Abreu, Ricardo Silva, Sergei Medvedev, Alexander Ryss, Smiljka Šimić, Karol Marhold, Eduard Stloukal, Davorin Tome, Marian Ramos, Benito Valdés, Francisco Pina, Sven Kullander, Anders Telenius, Yves Gonseth, Pascal Tschudin, Oleksandra Sergeyeva, Volodymyr Vladymyrov, Volodymyr Rizun, Chris Raper, Dan Lear, Pavel Stoev, Lyubomir Penev, Ana Rubio, Thierry Backeljau, Hannu Saarenmaa, and Sandrine Ulenberg. PESI - a taxonomic backbone for Europe. *Biodiversity Data Journal*, 3:e5848, 9 2015.

[50] Adanma Cecilia Eberendu. Unstructured Data: an overview of the data of Big Data. *International Journal of Computer Trends and Technology*, 38(1):46–50, 2016.

[51] EBI. European Bioinformatics Institute, https://www.ebi.ac.uk.

[52] Paul N. Edwards, Matthew S. Mayernik, Archer L. Batcheller, Geoffrey C. Bowker, and Christine L. Borgman. Science friction: Data, metadata, and collaboration. *Social Studies of Science*, 41(5):667–690, 2011.

[53] EMBL-EBI. European Bioinformatics Institute, https://www.ebi.ac.uk/.

[54] Encylopedia.com. ”phylogenetic species concept”, 0.

[55] EOL. Encyclopedia of Life, http://eol.org.

[56] Jane Evison. What are the basics of analysing a corpus? In Anne. O'Keeffe and Michael McCarthy, editors, *The Routledge Handbook of Corpus Linguistics*, pages 122–135. Routledge, 2010.

[57] Pamela Faber, Pilar León-Araúz, and Juan Antonio Prieto. Semantic relations, dynamicity, and terminological knowledge bases. *Current Issues in Language Studies*, 1(1):1–23, 2009.

[58] Roberta Facchinetti. *Corpus linguistics 25 years on.* Rodopi, 2007.

[59] Jung Wei Fan and Carol Friedman. Deriving a probabilistic syntacto-semantic grammar for biomedicine based on domain-specific terminologies. *Journal of Biomedical Informatics*, 44(5):805–814, 2011.

[60] Christiane. Fellbaum. *WordNet : an electronic lexical database.* MIT Press, 1998.

[61] P Flemons, J Tann, L Kelly, and D Hobern. Uses for biodiversity data - the Atlas of Living Australia user needs analysis. In A.L. Weitzman and Belbin, editors, *The Proceedings of TDWG*, 2008.

[62] Thierry Fontenelle. Lexicography. In James Simpson, editor, *The Routledge Handbook of Applied Linguistics*. Routledge, 2011.

[63] Nico Franz, Edward Gilbert, Bertram Ludäscher, and Alan Weakley. Controlling the taxonomic variable: Taxonomic concept resolution for a southeastern United States herbarium portal. *Research Ideas and Outcomes*, 2:e10610, 2016.

[64] Nico Franz and Robert Peet. Perspectives : Towards a language for mapping relationships among taxonomic concepts. *Systematics and Biodiversity*, (March), 2009.

[65] Nico M. Franz, Mingmin Chen, Parisa Kianmajd, Shizhuo Yu, Shawn Bowers, Alan S. Weakley, and Bertram Ludäscher. Names are not good enough: Reasoning over taxonomic change in the Andropogon complex. *Semantic Web*, 7(6):645–667, 2016.

[66] Nico M. Franz and David Thau. Biological taxonomy and ontology development: scope and limitations. *Biodiversity Informatics*, 7(1):45–66, 1 2010.

[67] R. Froese, D. Pauly, and (editors). FishBase. World Wide Web electronic publication.

[68] Daniel Fuentes and Nicola Fiore. The LifeWatch approach to the exploration of distributed species information. *ZooKeys*, 463(463):133–48, 1 2014.

[69] Pedro A. Fuertes-Olivera, editor. *The Routledge handbook of lexicography.* Routledge, 2017.

[70] GBIF: The Global Biodiversity Information Facility. What is GBIF?, https://www.gbif.org/what-is-gbif, 2019.

[71] GBIF.org. Global Biodiversity Information Facility, https://www.gbif.org.

[72] S. Geetha and G. S. Anandha Mala. Effectual extraction of data relations from unstructured data. *IET Conference Publications*, (624 CP):58–61, 2012.

[73] Saurav Ghosh, Prithwish Chakraborty, Emily Cohn, John S. Brownstein, and Naren Ramakrishnan. Characterizing Diseases from Unstructured Text: A Vocabulary Driven Word2vec Approach. In *Association for Computing Machinery (ACM)*, page 1129–1138, 2016.

[74] Juan Carlos Gil-Berrozpe, Pilar León-Araúz, and Pamela Faber. Specifying Hyponymy Subtypes and Knowledge Patterns : A Corpus-based Study. In *eLex*, pages 63–92, 2014.

[75] F Gill, D Donsker, and (Eds.). IOC World Bird Names List (v6.2), 2016.

[76] Georgios V. Gkoutos, Paul N. Schofield, and Robert Hoehndorf. The anatomy of phenotype ontologies: principles, properties and applications. *Briefings in bioinformatics*, 19(5):1008–1021, 2018.

[77] Global Biodiversity Informatics Facilities. Delivering Biodiversity Knowledge in the Information Age. *Global Biodiversity Informatics Conference*, (October), 2012.

[78] Ajda Gokcen, Ethan Hill, and Michael White. Madly Ambiguous: A Game for Learning about Structural Ambiguity and Why It's Hard for Computers. In *Association of Computational Linguistics*, pages 51–55, 2018.

[79] Graciela H. Gonzalez, Tasnia Tahsin, Britton C. Goodale, Anna C. Greene, and Casey S. Greene. Recent advances and emerging applications in text and data mining for biomedical discovery. *Briefings in Bioinformatics*, 17(1):33–42, 2016.

[80] C. Gray, A. Ma, D. Perkins, L. Hudson, D. Figueroa, and G. Woodward. Database of trophic interactions [Data set]., 2015.

[81] Stefan Th Gries. Introduction. In Anatol Stefanowitsch Stefan Thomas Gries, editor, *Corpora in Cognitive Linguistics: Corpus-Based Approaches to Syntax and Lexis*, chapter Introducti, pages 1–18. Walter de Gruyter GmbH, 2007.

[82] Stefan Th Gries. Dispersions and Adjusted Frequencies in Corpora. *International Journal of Corpus Linguistics*, 11(4):403–437, 2008.

[83] Stefan Th Gries. What is Corpus Linguistics ? *Language and Linguistics Compass 3*, 3:1–17, 2009.

[84] Thomas R. Gruber. Toward Principles for the Design of Ontologies used for Knowledge Sharing. *International Journal of Human-Computer Studies*, 43(5-6):907–928, 1995.

[85] Nancy E. Gwinn and Constance Rinaldo. The Biodiversity Heritage Library: sharing biodiversity literature with the world. *IFLA Journal*, 35(1):25–34, 3 2009.

[86] Aric Hagberg, Dan Schult, and Pieter Swart. NetworkX Reference, 2019.

[87] M.A.K. Halliday, Wolfgang Teubert, Colin Yallop, and Anna Cermakova. *Lexicology and Corpus Linguistics: An Introduction.* Continuum, London, 2 edition, 2005.

[88] Yongqun He, Zuoshuang Xiang, Jie Zheng, Yu Lin, James A. Overton, and Edison Ong. The eXtensible ontology development (XOD) principles and tool implementation to support ontology interoperability. *Journal of Biomedical Semantics*, 9(1):1–10, 2018.

[89] University of Michigan HE Group. OntoBee.

[90] Marti A. Hearst. Automatic Acquisition of Hyponyms from Large Text Corpora Lexico-Syntactic for Hyponymy Patterns. *Proceedings of the 14th Conference on Computational Linguistics*, page 539, 1992.

[91] Christian Herzog, Christian Handke, and Erik Hitters. Analyzing Talk and Text II: Thematic Analysis. *The Palgrave Handbook of Methods for Media Policy Research*, (April):385–401, 2019.

[92] Alan R Hevner. A Three Cycle View of Design Science Research A Three Cycle View of Design Science Research. *Scandinavian Journal of Information Systems*, 19(2):87–92, 2007.

[93] Alan R Hevner, Salvatore T March, Jinsoo Park, and Sudha Ram. DESIGN SCIENCE IN INFORMATION SYSTEMS RESEARCH. *MIS Quarterly*, 28(1):75–105, 2004.

[94] Cody E. Hinchliff, Stephen A. Smith, James F. Allman, J. Gordon Burleigh, Ruchi Chaudhary, Lyndon M. Coghill, Keith A. Crandall, Jiabin Deng, Bryan T. Drew, Romina Gazis, Karl Gude, David S. Hibbett, Laura A. Katz, H. Dail Laughinghouse, Emily Jane McTavish, Peter E. Midford, Christopher L. Owen, Richard H. Ree, Jonathan A. Rees, Douglas E. Soltis, Tiffani Williams, and Karen A. Cranston. Synthesis of phylogeny

and taxonomy into a comprehensive tree of life. *Proceedings of the National Academy of Sciences*, 112(41), 9 2015.

[95] Andrew Hippisley, David Cheng, and Khurshid Ahmad. The head-modifier principle and multilingual term extraction. *Natural Language Engineering*, 11(2):129–157, 2005.

[96] Donald Hobern, Alberto Apostolico, Elizabeth Arnaud, Juan Carlos Bello, Dora Canhos, Gregoire Dubois, Dawn Field, Enrique Alonso García, Alex Hardisty, Jerry Harrison, Bryan Heidorn, Leonard Krishtalka, Erick Mata, Roderic Page, Cynthia Parr, Jeff Price, and Selwyn Willoughby. Global Biodiversity Informatics Outlook. page 41, 2013.

[97] Trevor R. Hodkinson and John A. N. Parnell. *Reconstructing the tree of life : taxonomy and systematics of species rich taxa.* CRC/Taylor & Francis, 2007.

[98] Dr Dave Hone. "How species are identified", 6 2013.

[99] Matthew Horridge, Holger Knublauch, Alan Rector, Robert Stevens, Chris Wroe, Simon Jupp, Georgina Moulton, Nick Drummond, and Sebastian Brandt. A Practical Guide To Building OWL Ontologies Using Protégè 4 and CO-ODE Tools Edition 1.3, 2011.

[100] Zhisheng Huang, Frank Van Harmelen, and Annette Ten Teije. Reasoning with inconsistent ontologies. *IJCAI International Joint Conference on Artificial Intelligence*, pages 454–459, 2005.

[101] Ignacio Iacobacci, Mohammad Taher Pilehvar, and Roberto Navigli. Embeddings for word sense disambiguation: An evaluation study. *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Long Papers*, 2(2003):897–907, 2016.

[102] International Commission on Zoological Nomenclature. *International Code of Zoological Nomenclature adopted by the International Union of Biological Sciences.* The International Trust for Zoological Nomenclature, 4 edition, 1999.

[103] ITIS. Integrated Taxonomic Information System, https://www.itis.gov.

[104] ITIS. ITIS - Background, https://www.itis.gov/info.html.

[105] ITIS. ITIS - Data access, https://www.itis.gov/access.html.

[106] ITIS. ITIS - Data Development History and Data Quality, https://itis.gove/itis_primer.html.

[107] ITIS. ITIS Data Model: Entity Relationships and Element Definitions, https://www.itis.gov.uk/pdf/ITIS_conceptualModelEntityDefinition.pdf, 2019.

[108] Vindula Jayawardana, Dimuthu Lakmal, Nisansa De Silva, Amal Shehan Perera, Keet Sugathadasa, Buddhi Ayesha, and Madhavi Perera. Semi-supervised instance population of an ontology using word vector embeddings. In *17th International Conference on Advances in ICT for Emerging Regions, ICTer 2017*, volume Jan-2018, 2018.

[109] Andrew C. Jones, Richard J. White, and Ewen R. Orme. Identifying and Relating Biological Concepts in the Catalogue of Life. *Journal of biomedical semantics*, 2(1), 10 2011.

[110] Clement Jonquet, Anne Toulet, Elizabeth Arnaud, Sophie Aubin, Esther Dzalé Yeumo, Vincent Emonet, John Graybeal, Marie-angélique Laporte, Mark A Musen, Valeria Pesce, and Pierre Larmande. AgroPortal : A vocabulary and ontology repository for agronomy. *Computers and Electronics in Agriculture*, 144(October 2017):126–143, 2018.

[111] Sin-Jae Kang and Jong-Hyeok Lee. Semi-automatic practical ontology construction by using a thesaurus, computational dictionaries, and large corpora. *HLTKM '01 Proceedings of the workshop on Human Language Technology and Knowledge Management*, 2001(6):1–8, 2001.

[112] Cathleen Kantner, Amelie Kutter, Andreas Hildebrandt, and Mark Puettcher. How to get rid of the noise in the corpus: Cleaning large samples of digital newspaper texts. 2011.

[113] Amye Kenall, Simon Harold, and Christopher Foote. An open future for ecological and evolutionary data? *BMC ecology*, 14(1):10, 2014.

[114] A. Kilgarriff, V. Kovář, and S. Krek. A quantitative evaluation of word sketches. *Proceedings of the 14th EURALEX International Congress, Leeuwarden, The Netherlands*, pages 372–79, 2010.

[115] Adam Kilgarriff. "I don't believe in word senses". *Computers and the Humanities*, 31(2):91–113, 1997.

[116] Adam Kilgarriff, Pavel Rychly, Pavel Smrz, and David Tugwell. The Sketch Engine. *Proceedings of the Eleventh EURALEX International Congress*, pages 105–116, 2004.

[117] Adam Kilgarriff and David Tugwell. WORD SKETCH : Extraction and Display of Significant Collocations for Lexicography for Lexicography. In *Collocations workshop*. Association of Computational Linguistics, 2001.

[118] Almut Koester. Building small specialised corpora. In Anne. O'Keeffe and Michael McCarthy, editors, *The Routledge Handbook of Corpus Linguistics2*, chapter 6, pages 66–79. Routledge, Abingdon, 1 edition, 2010.

[119] Christian König, Patrick Weigelt, Julian Schrader, Amanda Taylor, Jens Kattge, and Holger Kreft. Biodiversity data integration–The significance of data resolution and domain. *PLOS Biology*, 3 2019.

[120] Drew Koning, Indra Neil Sarkar, and Thomas M. Moritz. TaxonGrab: extracting taxonomic names from text. *Biodiversity Informatics*, 2:79–82, 2005.

[121] Agnieszka Konys. Knowledge Systematization for ontology learning methods. *Procedia Computer Science*, 126:2194–2207, 2018.

[122] M Koperski, M Sauer, W Braun, and S R Gradstein. Referenzliste der Moose Deutschlands. In *Schriftenreihe für Vegetationskunde*, volume 34, page 519. 2000.

[123] Richard A Krueger and Marry Anne Casey. Focus groups: A practical guide for applied research 5th Edition. *Focus Groups: A Practical Guide for Applied Research*, pages 63–84, 2015.

[124] John La Salle, Kristen J Williams, and Craig Moritz. Biodiversity analysis in the digital era. *Philosophical Transactions of the Royal Society of London B: Biological Sciences‘*, 371(1702), 2016.

[125] Mathieu Lafourcade and Lionel Ramadier. Semantic Relation Extraction with Semantic Patterns : Experiment on Radiology Report. In *LREC 2016 Conference on Language Resources and Evaluation*, 2016.

[126] Anne Lauscher and Goran Glavaš. Are We Consistently Biased? Multidimensional Analysis of Biases in Distributional Word Vectors. *Association for Computational Linguistics (ACL)*, pages 85–91, 2019.

[127] Geoffrey Leech. Corpora and Theories of Linguistic Performance. In Jan Svartvik, editor, *Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82*, pages 105–122, 1992.

[128] Pilar León-Araúz and Antonio San Martín. The EcoLexicon Semantic Sketch Grammar: from Knowledge Patterns to Word Sketches. *Proceedings of the LREC 2018 Workshop "Globalex 2018 – Lexicography & WordNets"*, 2013(2016):94–99, 2018.

[129] Lexical Computing Ltd. CQL - Corpus Query Language, https://www.sketchengine.eu/documentation/corpus-querying/.

[130] "Lexical Computing Ltd.". Sketch Engine - documentation, https://www.sketchengine.eu/documentation/.

[131] "Lexical Computing Ltd.". Statistics used in the Sketch Engine, https://www.sketchengine.eu/documentation/statistics-used-in-sketch-engine/, 2015.

[132] Hans Lindquist. *Corpus linguistics and the description of English.* Edinburgh University Press, Edinburgh, 1 edition, 2009.

[133] Feifan Liu, Jinying Chen, Abhyuday Jagannatha, and Hong Yu. Learning for Biomedical Information Extraction: Methodological Review of Recent Advances. 2016.

[134] Steve Lohr. The Age of Big Data, 2 2012.

[135] Richard K. Lomotey and Ralph Deters. Topics and terms mining in unstructured data stores. *Proceedings - 16th IEEE International Conference on Computational Science and Engineering, CSE 2013*, pages 854–861, 2013.

[136] Bertram Ludäscher, Kai Lin, Shawn Bowers, Efrat Jaeger-Frank, Boyan Brodaric, and Chaitan Baru. Managing scientific data: From data integration to scientific workflows. *Special Paper of the Geological Society of America*, 397:109–129, 2006.

[137] Joshua S. Madin, Shawn Bowers, Mark P. Schildhauer, and Matthew B. Jones. Advancing ecological research with ontologies. *Trends in Ecology and Evolution*, 23(3):159–168, 2008.

[138] Alexander Maedche and Steffen Staab. Ontology Learning for the Semantic Web. *IEEE Intelligent Systems*, 16(2):72–79, 2001.

[139] James Mallet. What are species? *TREE*, 12(11):453–454, 1997.

[140] "Max Plank Institute for Informatics". Network Analyzer Help.

[141] R.L. Mayden. A hierarchy of species concepts: The denouement in the saga of the species problem. In M. F. Claridge, H. A. Dawah, and M. R. Wilson, editors, *Species: The units of biodiversity*, pages 381–424. Chapman and Hall, London, 1997.

[142] E. Mayr. *Principles of Systematic Zoology.* McGraw–Hill, New York, 1969.

[143] Diana Mccarthy, Marianna Apidianaki, and Katrin Erk. Word Sense Clustering and Clusterability. In *Computational Linguistics*, number 42, pages 245–275, 2016.

[144] Diana McCarthy, Adam Kilgarriff, Miloš Jakubíček, and Siva Reddy. Semantic Word Sketches. *8th International Corpus Linguistics Conference (CL 2015)*, 2015.

[145] David D Mcdonald. Lexical Inference and the Problem of the Long Tail. In *Association for the Advancement of Artificial Intelligence*, pages 71–73, 2008.

[146] Wes McKinney. pandas: a Foundational Python Library for Data Analysis and Statistics. In *PyHPC*, 2011.

[147] Olena Medelyan, Ian H. Witten, Anna Divoli, and Jeen Broekstra. Automatic construction of lexicons, taxonomies, ontologies, and other knowledge structures. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 3(4):257–279, 2013.

[148] Olena Medelyan, Ian H. Witten, Anna Divoli, and Jeen Broekstra. Automatic construction of lexicons, taxonomies, ontologies, and other knowledge structures. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 3(4):257–279, 2013.

[149] William K. Michener and Matthew B. Jones. Ecoinformatics: Supporting ecology as a data-intensive science. *Trends in Ecology and Evolution*, 27(2):88–93, 2012.

[150] Peter E. Midford, Thomas Alex Dececchi, James P. Balhoff, Wasila M. Dahdul, Nizar Ibrahim, Hilmar Lapp, John G. Lundberg, Paula M. Mabee, Paul C. Sereno, Monte Westerfield, Todd J. Vision, and David C. Blackburn. The vertebrate taxonomy ontology: A framework for reasoning across model organism and species phenotypes. *Journal of Biomedical Semantics*, 4(1):2–7, 2013.

[151] Peter E. et al Midford. Vertebrate Taxonomy Ontology.

[152] Tomáš Mikolov, Anoop Deoras, Stefan Kombrink, Lukáš Burget, and Jan Honza Černocký. Empirical evaluation and combination of advanced language modeling techniques. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pages 605–608, 2011.

[153] Jeremy A. Miller, Torsten Dikow, Donat Agosti, Guido Sautter, Terry Catapano, Lyubomir Penev, Zhi-Qiang Zhang, Dean Pentcheff, Richard L. Pyle, Stanley Blum, Cynthia S. Parr, Chris Freeland, Tom Garnett, Linda S. Ford, Burgert S. Muller, Leo Smith, Ginger Strader, Teodor Georgiev, and Laurence Bénichou. From taxonomic literature to cybertaxonomic content. *BMC Biology*, 10(87), 1 2012.

[154] Andriy Mnih. A fast and simple algorithm for training neural probabilistic language models. In *29th International Conference on Machine Learning*, pages 1751–1758, 2012.

[155] ITIS: Oncorhynchus Mykiss. https://www.itis.gov/servlet/SingleRpt/SingleRpt?search_topic=TSN&search_value=

[156] Tom Narock and Adam Shepherd. Semantics all the way down : the Semantic Web and open science in big earth data. *Big Earth Data*, 1(1-2):159–172, 2017.

[157] Roberto Navigli. Word sense disambiguation: A survey. *ACM Computing Surveys*, 41(2), 2009.

[158] Mike Nelson. Building a written corpus. In Anne. O'Keeffe and Michael McCarthy, editors, *The Routledge Handbook of Corpus Linguistics*, chapter 5, pages 53–65. Routledge, Abingdon, 1 edition, 2010.

[159] Vincent Ng and Claire Cardie. Identifying anaphoric and non-anaphoric noun phrases to improve coreference resolution. *Proceedings of the 19th international conference on Computational linguistics* -, 1(1987):1–7, 2002.

[160] Farhad Nooralahzadeh, Lilja Øvrelid, and Jan Tore Lønning. Evaluation of Domain-specific Word Embeddings using Knowledge Resources. *International Conference on Language Resources and Evaluation*, pages 1438–1445, 2018.

[161] Andreia Novelli and Jose Oliveira. Simple Method for Ontology Automatic Extraction from Documents. *International Journal of Advanced Computer Science and Applications*, 3(12):44–51, 2012.

[162] OBO Foundry. NCBI organismal classification: An ontology representation of the NCBI organismal taxonomy, http://www.obofoundry.org/ontology/ncbitaxon.html.

[163] Anne. O'Keeffe and Michael McCarthy, editors. *The Routledge handbook of corpus linguistics*. Routledge, 2010.

[164] Tobias O.Nyumba, Kerrie Wilson, Christina J. Derrick, and Nibedita Mukherjee. The use of focus group discussion methodology: Insights from two decades of application in conservation. *Methods in Ecology and Evolution*, 9(1):20–32, 2018.

[165] Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, and Ideker T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, 13(11):2498–504, 2003.

[166] Evangelos Pafilis, Sune P. Frankild, Lucia Fanini, Sarah Faulwetter, Christina Pavloudi, Aikaterini Vasileiadou, Christos Arvanitidis, and Lars Juhl Jensen. The SPECIES and ORGANISMS Resources for Fast and Accurate Identification of Taxonomic Names in Text. *PLoS ONE*, 8(6):2–7, 2013.

[167] Evangelos Pafilis, Sune P. Frankild, Julia Schnetzer, Lucia Fanini, Sarah Faulwetter, Christina Pavloudi, Katerina Vasileiadou, Patrick Leary, Jennifer Hammock, Katja Schulz, Cynthia Sims Parr, Christos Arvanitidis, and Lars Juhl Jensen. ENVIRONMENTS and EOL: Identification of Environment Ontology terms in text and the annotation of the Encyclopedia of Life. *Bioinformatics*, 31(11):1872–1874, 2015.

[168] Cynthia S. Parr, Robert Guralnick, Nico Cellinese, and Roderic D M Page. Evolutionary informatics: Unifying knowledge about the diversity of life. *Trends in Ecology and Evolution*, 27(2):94–103, 2012.

[169] Cynthia S Parr and Anne E Thessen. Biodiversity Informatics. In *Ecological Informatics*, pages 375–399. 2018.

[170] Cynthia S. Parr, Nathan Wilson, Patrick Leary, Katja Schulz, Kristen Lans, Lisa Walley, Jennifer Hammock, Anthony Goddard, Jeremy Rice, Marie Studer, Jeffrey Holmes, Robert Corrigan, Jr., Jr., and Robert J. Corrigan Jr. The Encyclopedia of Life v2: Providing Global Access to Knowledge About Life on Earth. *Biodiversity Data Journal*, 2, 5 2014.

[171] D. J. Patterson, J. Cooper, P. M. Kirk, R. L. Pyle, and D. P. Remsen. Names are key to the big new biology. *Trends in Ecology and Evolution*, 25(12):686–691, 2010.

[172] David Patterson, Dmitry Mozzherin, David Shorthouse, and Anne Thessen. *Challenges with using names to link digital biodiversity information*, volume 4. 2016.

[173] İzzet Pembeci. Using Word Embeddings for Ontology Enrichment. *International Journal of Intelligent Systems and Applications in Engineering*, 4(3):49, 2016.

[174] Lyubomir Penev, Donat Agosti, Teodor Georgiev, Terry Catapano, Jeremy Miller, Vladimir Blagoderov, David Roberts, Vincent Smith, Irina Brake, Simon Ryrcroft, Ben Scott, Norman Johnson, Robert Morris, Guido Sautter, Vishwas Chavan, Tim Robertson, David Remsen, Pavel Stoev, Cynthia Parr, Sandra Knapp, W. John Kress, Frederic Thompson, and Terry Erwin. Semantic tagging of and semantic enhancements to systematics papers: ZooKeys working examples. *ZooKeys*, 50:1–16, 6 2010.

[175] Lyubomir Penev, Mariya Dimitrova, Viktor Senderov, Georgi Zhelezov, Teodor Georgiev, Pavel Stoev, and Kiril Simov. OpenBiodiv: A Knowledge Graph for Literature-Extracted Linked Open Data in Biodiversity Science. *Publications*, 7(2), 5 2019.

[176] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global Vectors for Word Representation. *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

[177] A Townsend Peterson, Jorge Soberón, and Leonard Krishtalka. A global perspective on decadal challenges and priorities in biodiversity informatics. *BMC ecology*, 15(1):15, 1 2015.

[178] Jorrit H. Poelen, James D. Simons, and Chris J. Mungall. Global biotic interactions: An open infrastructure to share and analyze species-interaction datasets. *Ecological Informatics*, 24:148–159, 2014.

[179] Richard L. Pyle. Towards a Global Names Architecture: The future of indexing scientific names. *ZooKeys*, 550(550):261–281, 1 2016.

[180] "Python Core Team". Python: A dynamic, open source programming language, 2015.

[181] QSR International. NVivo Qualitative Data Analysis Software [Software]. Available from https://qsrinternational.com/nvivo/nvivo-products/, 1999.

[182] Radio 4. "Hybrids", In Our Time, 2019.

[183] Alok Ranjan Pal and Diganta Saha. Word Sense Disambiguation: A Survey. *International Journal of Control Theory and Computer Modeling*, 5(3):1–16, 2015.

[184] Randi Reppen. Building a corpus: what are key considerations? In Anne O'Keefe and Michael McCarthy, editors, *The Routledge Handbook of Corpus Linguistics*, pages 31–37. Routledge, Abingdon, 1 edition, 2010.

[185] Arpita Roy, Youngja Park, and SHimei Pan. Learning Domain-Specific Word Embeddings from Sparse Cybersecurity Texts. *CoRR*, 2017.

[186] Pavel Rychlý. A Lexicographer-Friendly Association Score. *Proceedings of Second Workshop on Recent Advances in Slavonic Natural Languages Processing, RASLAN 2008*, pages 6–9, 2008.

[187] Guido Sautter, Klemens Böhm, and Donat Agosti. A combining approach to Find All taxon names (FAT) in legacy biosystematics literature. *Biodiversity Informatics*, 3:46–58, 2006.

[188] Stefan Schlobach, Zhisheng Huang, Ronald Cornet, and Frank Van Harmelen. Debugging incoherent terminologies. *Journal of Automated Reasoning*, 39(3):317–349, 2007.

[189] Ida Schomburg, Antje Chang, and Dietmar Schomburg. Standardization in enzymology: data integration in the world's enzyme information system BRENDA. *Perspectives in Science*, 1(1-6):15–23, 2014.

[190] Marc Schrieber. *Towards Effective Natural Language Application Development.* PhD thesis, Kassel University, 2019.

[191] Stefan Schulz, Holger Stenzhorn, and Martin Boeker. The ontology of biological taxa. *Bioinformatics*, 24(13):313–321, 2008.

[192] Viktor Senderov, Kiril Simov, Nico Franz, Pavel Stoev, Terry Catapano, Donat Agosti, Guido Sautter, Robert A Morris, and Lyubomir Penev. OpenBiodiv-O: ontology of the OpenBiodiv knowledge management system. *Journal of Biomedical Semantics*, 9(5), 2018.

[193] Mehrnoush Shamsfard and Ahmad Abdollahzadeh Barforoush. The state of the art in ontology learning: A framework for comparison. *Knowledge Engineering Review*, 18(4):293–316, 2003.

[194] Mehrnoush Shamsfard and Ahmad Abdollahzadeh Barforoush. Learning ontologies from natural language texts. *Int. J. Human-Computer Studies*, 60:17–63, 2004.

[195] P Shannon. What is Cytoscape?, https://cytoscape.org/what_is_cytoscape.html.

[196] Abraham Solomonick. Towards a Comprehensive Theory of Lexicographic Definitions. In *EuraLex*, pages 481–488, 1996.

[197] Steffen Staab and Rudi Studer. *Handbook on Ontologies.* 2007.

[198] Roger Alan Stein, Patricia A. Jaques, and João Francisos Valiati. An Analysis of Hierarchical Text Classification Using Word Embeddings. *Information Sciences*, (471):216–232, 2019.

[199] Brian J Stucky, Rob Guralnick, John Deck, Ellen G Denny, Kjell Bolmgren, and Ramona Walls. The Plant Phenology Ontology: A New Informatics Resource for Large-Scale Integration of Plant Phenology Data. *Frontiers in Plant Science*, 9(May):1–12, 2018.

[200] Taxonomic Names and Concepts Group. Taxonomic Concept Transfer Schema, 2005.

[201] Carol Tenopir, Suzie Allard, Kimberly Douglass, Arsev Umur Aydinoglu, Lei Wu, Eleanor Read, Maribeth Manoff, and Mike Frame. Data sharing by scientists: practices and perceptions. *PloS one*, 6(6):e21101, 1 2011.

[202] David Thau and Bertram Ludäscher. Reasoning about taxonomies in first-order logic. *Ecological Informatics*, 2(3 SPEC. ISS.):195–209, 2007.

[203] The Plant List. The Plant List Version 1.1., http://www.theplantlist.org/1.1/cite/, 2013.

[204] Anne Thessen, Jenette Preciado, Payoj Jain, James Martin, Martha Palmer, and Riyaz Bhat. Automated Trait Extraction using ClearEarth, a Natural Language Processing System for Text Mining in Natural Sciences. *Biodiversity Information Science and Standards*, 2:e26080, 2018.

[205] Anne Thessen, Jenette Preciado, and Chris Jenkins. Collaboration between the Natural Sciences and Computational Linguistics : A Discussion of Issues. Technical Report December, 2018.

[206] Anne E. Thessen, Daniel E. Bunker, Pier Luigi Buttigieg, Laurel D. Cooper, Wasila M. Dahdul, Sami Domisch, Nico M. Franz, Pankaj Jaiswal, Carolyn J. Lawrence-Dill, Peter E. Midford, Christopher J. Mungall, Martín J. Ramírez, Chelsea D. Specht, Lars Vogt, Rutger Aldo Vos, Ramona L. Walls, Jeffrey W. White, Guanyang Zhang, Andrew R. Deans, Eva Huala, Suzanna E. Lewis, and Paula M. Mabee. Emerging semantics to link phenotype and environment. *PeerJ*, 3:e1470, 12 2015.

[207] Anne E. Thessen, Hong Cui, and Dmitry Mozzherin. Applications of natural language processing in biodiversity science. *Advances in Bioinformatics*, 2012, 2012.

[208] Anne E. Thessen and Cynthia Sims Parr. Knowledge extraction and semantic annotation of text from the encyclopedia of life. *PLoS ONE*, 9(3):e89550, 3 2014.

[209] Jouni Tuominen, Nina Laurenne, and Eero Hyvonen. Biological Names and Taxonomies on the Semantic Web – Managing the Change in Scientific Conception. In *The Semanic Web: Research and Applications: 8th Extended Semantic Web Conference, ESWC 2011*, number Part II, pages 255–269, 2011.

[210] N. Turland, J. H. Wiersema, F. R. Barrie, W. Greuter, D. L. Hawksworth, P. S. Herendeen, S. Knapp, W. H. Kusber, D.-Z Li, K. Marhold, T. W. May, J. McNeill, A. M. Monro, J. Prado, M. J. Price, and G. F. Smith, editors. *International Code of Nomenclature for algae, fungi, and plants (Shenzhen Code) adopted by the Nineteenth International Botanical Congress Shenzhen, China, July 2017.* Koeltz Botanical Books, Glashütten, regnum veg edition.

[211] Gaurav Vaidya, Denis Lepage, and Robert Guralnick. The tempo and mode of the taxonomic correction process: How taxonomists have corrected and recorrected North American bird species over the last 127 years. *PLoS ONE*, 13(4):1–19, 2018.

[212] Edward Vanden Berghe, Gianpaolo Coro, Nicolas Bailly, Fabio Fiorellato, Caselyn Aldemita, Anton Ellenbroek, and Pasquale Pagano. Retrieving taxa names from large biodiversity data collections using a flexible matching workflow. *Ecological Informatics*, 28:29–41, 2015.

[213] Machhindra Govind Varpe. The Traditional , Structural And Cognitive Approach To Linguistics. *IOSR Journal Of Humanities And Social Science (IOSR-JHSS)*, 22(12):39–43, 2017.

[214] Violeta Seretan. Bridging Collocation and Syntactic Analysis. In Pascual Cantos-Gómez and Moisés Almela-Sánchez, editors, *Lexical Collocation Analysis: Advanced and Applications*, chapter 2, pages 23–38. Springer International Publishing, 2018.

[215] Smith V.S., Rycroft S., Scott B., Baker E., Livermore L., Heaton A., Bouton K., Koureas D.N., and Roberts D. Scratchpads 2.0: a virtual research environment infrastructure for biodiversity data, 2012.

[216] Ramona L. Walls, John Deck, Robert P. Guralnick, Steven J. Baskauf, Reed S. Beaman, Stanley Blum, Shawn Bowers, Pier Luigi Buttigieg, Neil Davies, Dag Endresen, Maria Alejandra Gandolfo, Robert Hanner, Alyssa Janning, Leonard Krishtalka, Andréa Matsunaga, Peter Midford, Norman Morrison, Éamonn Ó. Tuama, Mark P. Schildhauer, Barry Smith, Brian J. Stucky, Andrea K. Thomer, John R. Wieczorek, Jamie Whitacre, and John Wooley. Semantics in Support of Biodiversity Knowledge Discovery: An Introduction to the Biological Collections Ontology and Related Ontologies. *PLoS ONE*, 9(3), 3 2014.

[217] Ramona L Walls, Robert Guralnick, John Deck, Adam Buntzman, Pier L Buttigieg, Neil Davies, Michael W Denslow, Rachel E Gallery, J J Parnell, David Osumi-Sutherland, Robert J Robbins, Philippe Rocca-Serra, John Wieczorek, and Jie Zheng. Meeting report: advancing practical applications of biodiversity ontologies. *Standards in Genomic Sciences*, 9(1):17, 12 2014.

[218] John R. Wieczorek, David Bloom, Robert P. Guralnick, Stanley Blum, Markus Döring, Renato Giovanni, Tim Robertson, and David Vieglais. Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. *PLoS ONE*, 7(1):e29715, 1 2012.

[219] Wikipedia contributors. WikiData, https://www.wikidata.org/wiki/Wikidata:Main_Page, 2020.

[220] ”Wiley Online Library”. Wiley Text and Data Mining Agreement - Wiley Online Library, 2019.

[221] Don E. Wilson and DeeAnn M. Reeder, editors. *Mammal Species of the World: A Taxonomic and Geographic Reference.* JHU Press, 3 edition, 2005.

[222] Adrian Wong, Joseph M. Plasek, Steven P. Montecalvo, and Li Zhou. Natural Language Processing and Its Implications for the Future of Medication Safety: A Narrative Review of Recent Advances and Challenges. *Pharmacotherapy*, 38(8):822–841, 2018.

[223] Maosong Sun Xinxiong Chen, Zhiyuan Liu. A Unified Model for Word Sense Representation and Disambiguation. *EMNLP2014*, pages 1025–1035, 2014.

[224] Hu Xu, Bing Liu, Lei Shu, and Philip S. Yu. Lifelong Domain Word Embedding via Meta-Learning. In *IJCAI International Joint Conference on Artificial Intelligence*, pages 4510–4516, 2018.

[225] Roskov Y., Abucay L., Orrell T., Nicolson D., and Et Al. Species 2000 & ITIS Catalogue of Life 2017 Annual Checklist, 2017.

[226] Colin Yallop. Words and Meaning. In *Lexicology and Corpus Linguistics: An Introduction*, chapter 2, pages 25–71. Continuum, London, 2 edition, 2005.

[227] Wei Zhao, Haiyun Peng, Steffen Eger, Erik Cambria, and Min Yang. Towards Scalable and Reliable Capsule Networks for Challenging NLP Applications. *Association for Computational Linguistics (ACL)*, pages 1549–1559, 2019.

Glossary

adjusted frequency

Adjusted frequency is a frequency that has been adjusted to balance various qualities of a word's presence in a dataset, such as frequency across different parts of the corpus.

bag of words

Approach in NLP that disregards all features (grammar, syntax, order) about the word, but looks at random groups of words that occur within the same data to make assumptions about them.

coherence

Coherence can have multiple meanings depending on context. In the context of ontologies, it refers to the logical coherence of the structure, or there not being any conflicts, or logical issues which would undermine the logic of the ontology. In linguistics, coherence refers to the logical and semantic consistency throughout a text, which means that it can be followed and understood by the reader..

conceptual stability

Conceptual stability refers to where the meaning behind the terminology remains constant. This is used in taxonomy when seeing if different circumscriptions are congruent, inclusive or exclusive in meaning. However, in this thesis the reference is specifically to whether the terms used are being used to mean the same thing across different corpora, or in comparison with an authoritative resource.

concordance

A search method in corpus linguistics in which you can search for a word, string of words or other linguistic phenomena. The results come out in a list of sentences that can be ordered and analysed to look for patterns..

corpora

Plural of corpus.

corpus

A large body of natural language texts, compiled to be representative of a domain.

data framing

Refers to the fact that data can be presented in different ways. In the case of this thesis, nomenclature is the data in question. Nomenclature includes many multi-word terms, so in the case of this thesis the framing relates to whether these terms are considered with each word as a separate unit or of each term is unified for the nomenclature term to be considered as one..

dispersion

Dispersion is the relative homogeneity or heterogeneity of the distribution of occurrences of a word across a dataset.

gold standard

Used in ontology creation, corpus linguistics and NLP as a resource which can be used as a baseline, as representative of a domain, for example.

informatics

The field of informatics is a branch of information engineering that focuses on information processing and systems. In this thesis it will be used to describe the domains of research specifically involved in the automation of many of the processes in extracting information from data and accessing it in an integrated way..

JEFF corpus

Corpus comprised of articles from Journal of Freshwater Fish Ecology, used in Phases 1 to 4 of the research.

knowledge representation

Knowledge representation is used in artificial intelligence to refer to the design of computer representations that capture information about the world that can be used to solve complex problems..

knowledge representation resource

Knowledge representation resource is used in this thesis to refer to data structures such as standardised vocabularies, taxonomies and ontologies..

lempos

Lempos is used in Sketch Engine to amalgamate information about a lemma by combining it with its part-of-speech (POS) (lemma + pos). It is formed by the lemma, plus a hyphen, plus a code for the POS. It is found in the third column of the WPL file as produced by Sketch Engine. Word Sketches output the lemma but the Sketch Grammars that produce the Word Sketches use the lempos instead of the lemma and POS separately..

named entity disambiguation

Named entity disambiguation is a task in NLP dedicated to disambiguating between different possible meanings of specific named entities.

named entity normalisation

Named entity normalisation is a task in NLP dedicated to identifying named entities and grouping them when they refer to the same thing..

named entity recognition

Named entity recognition is a task in NLP dedicated to identifying named entities in text for further processing..

narrative text

Narrative text refers to documents such as newspaper or journal articles, instead of tabular data.

natural language processing

Natural language processing is the term given to the process of making natural language computer-readable.

nomenclatural stability

Terminological stability of the use of nomenclature, in that the words used remain constant.

nomenclature pair

Pairs of words from the nomenclature that have been identified as being related. e.g. a species' name (genus and species - Salmo trutta), or a pair further up the hierarchy such as Salmonidae Salmo.

nomenclature profile study

Study performed as part of this PhD research to map characterisation of the scientific nomenclature and vernacular variants linked to a taxonomic entity.

nomenclature reference

Any word or group of words that forms part of the nomenclature for any rank in the hierarchy.

normalised frequency

Normalised frequency is a frequency adjusted to occurrences per million (or other number) to permit comparison between datasets of different sizes. In this research occurrences per million words has been used..

ontology

In the context of computer science and informatics, ontologies are formal, explicit descriptions of a specific domain or area. They are used for data representation. They come in various levels of formality: from full description logic, defined ontologies to more simple structures that simply follow a taxonomic hierarchy..

ontology population

Ontology population is the name given to creating entries for an ontology (concepts/classes and relations between them). This can be manual, semi-automatic or automatic..

original corpus

Version of the corpus which requires no prior processing before uploading to Sketch Engine for preliminary processing (lemmatisation, tokenisation, etc.).

range

Range (referred to as Range2) in the context of corpus linguistics is the result of a dispersion calculation which looks at the percentage of documents a word appears in within a dataset. It is referred to as Range2 to distinguish it from range in wider statistics, which describes the distance between the highest and lowest values of a variable..

raw frequency

Raw frequency is a straight frequency: the exact number of hits or occurrences of a word (or phrase) in a given dataset.

real life text

Real life text refers to texts that have been produced for real, as in during the course of real life, not created especially for research or a dictionary, for example.

relation network graph

Network relation graphs are produced from the totality of taxonomic entity mention pair relations identified in a corpus or for a specific taxonomic entity.

scientific nomenclature

Scientific naming system of species and the taxonomic hierarchy, which is the linguistic representation of the biological taxonomy.

semi-structured data

Semi-structured data comprises, as the name suggests, data which possesses some form of structure. Semi-structured data has some form of markup that instructs computers as to the meaning of sections of the data. Further examples will be given in subsequent sections..

standardised vocabulary

Standardised vocabulary refers to a set of instructions with descriptions as to how to describe a domain. These usually accompany a standard, which also sets out the different sections to be completed..

structured data

Structured data is the easiest form of data for computers to handle. Structured data usually comes in the form of tables or databases, in which all the information is clearly categorised for processing..

taxon

Taxon is scientific term to denote a group of organisms which are classified together. This is frequently used to refer to species, although it can refer to groups up the taxonomic ranking..

taxon concept

Taxon concept was devised to overcome issues where multiple taxonomists had described a specimen and assigned the specimen the same scientific name. To differentiate between the different circumscriptions, the taxon concept rules that the name should be followed by the author of the circumscription and the date of said description..

taxonomy

A taxonomy is by definition a means of classification. In the case of this thesis, any taxonomies will refer to either a general, non-identified taxonomy usually based on a hierarchy of things. These are used frequently in informatics and also in lexicography. The other taxonomy that features highly in the thesis is the biological taxonomy, or the hierarchy of species (and their ranking). This will be referred to throughout as the biological taxonomy..

term unification

Term unification refers to the joining of multi-word terms, in this case in the nomenclature and vernacular variants, to consider the term as one, not as multiple parts.

token

Token is used in corpus linguistics to denote the number of separate units in a corpus. Units comprise the occurrence of a word form (each mention is counted separately), and also usually punctuation (commas, full stops, etc.) . Spaces between are not counted..

type

In corpus linguistics the term type is used often synonymously with the term word: a string of letters which has a meaning and may have multiple forms (plural, singular, past, present, etc.)..

unified term corpus

Version of the corpus in which multi-word scientific nomenclature are joined with an underscore before the first stage of processing by Sketch Engine, so that multi-word scientific nomenclature is analysed as a single unit.

unstructured data

Unstructured data comprises all other data. It can include natural, narrative language or formats such as videos or pictures. In the case of this thesis it only refers to natural language texts. This data is the most difficult and expensive data for computers to process..

vernacular

Common variant of species' names.

WEB corpus

Corpus created through a web-scrape using seed words identified in the JEFF corpus, used in Phases 2 to 4 of the research.

word

In corpus linguistics the term word is often replaced with the word type: a string of letters which has a meaning and may have multiple forms (plural, singular, past, present, etc.)..

word embedding

Word embeddings are the name given to vectors created to identify the position of a word in context, according to processing large amounts of data using statistical means. They are also called word vectors in some contexts..

word sense disambiguation

Word sense disambiguation is the name for the area of research that looks at how to identify the different senses a specific word can be used to mean. It can take many different forms, from manual human-led forms to now many statistical and machine learning models..

Word Sketches

Feature of Sketch Engine corpus query tool that provides a page summary of a word's grammatical and collocational behaviour..

Acronyms

API

Application Programming Interface.

CASSPC

California Academy of Sciences fish species table ID.

CoL

Catalogue of Life.

CSV

comma-separated value.

ITIS

Integrated Taxonomic Information System.

JEFF

Journal for the Ecology of Freshwater Fish.

NCBI

National Center for Biotechnology Information.

NED

named entity disambiguation.

NEN

named entity normalisation.

NER

named entity recognition.

NLP

natural language processing.

obo

Open Biomedical Ontology.

RDF

Resource Description Framework.

VTO

Vertebrate Taxonomy Ontology.

XML

Extensible Markup Language.