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 I 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.