GI_Forum 2023, Issue 1
Page: 3 - 21
Full Paper
Corresponding Author:
k.stock@massey.ac.nz
DOI: 10.1553/giscience2023_01_s3

The BioWhere Project: Unlocking the Potential of Biological Collections Data

Kristin Stock[1], Kalana Wijegunarathna[1], Christopher B. Jones[2], Hone Morris[1], Pragyan Das[1], David Medyckyj-Scott[3], Brandon Whitehead[3]

[1]Massey University, New Zealand
[2]Cardiff University, United Kingdom
[3]Manaaki Whenua - Landcare Research, New Zealand

Abstract

Vast numbers of biological specimens (e.g. flora, fauna, soils) are stored in collections globally. Many of these have only a natural-language location description, such as '*200ft above and south of main highway, 1.1 miles west of Porters Pass*', and numerical coordinates are unknown. The BioWhere project is pioneering methods to automatically determine the geographic coordinates (georeferences) of complex location descriptions. Particular challenges are posed by the variable accuracy of recent and historical data that might be used to train models to predict geographic coordinates from the natural-language descriptions; by the presence of historical place names in the descriptions that are not stored in existing gazetteers; and by the vague and context-sensitive nature (e.g. *above*, *on*, *south of*) of the descriptions. We are addressing these challenges by extending the latest transformer-based deep learning models to parse locality descriptions, and to build models for specific spatial terms that incorporate geographic context and data quality to more accurately predict georeferences. We also describe a gazetteer that contains enriched cultural content to support georeferencing of historical records, and to serve as a store of New Zealand Māori cultural knowledge for future generations.

Keywords:

georeferencing, biological collections, machine learning, gazetteers

1 Introduction

Knowledge of the current and historical locations of biological species is crucial to support environmental decision making, particularly in the face of climate change. However, much information is missing or incomplete. Globally, museums, libraries and government agencies hold billions of biological specimens, journal papers and reports. In all of these, the locations of species are often described using natural language (e.g. '*On bench 200ft above and south of main highway, 1.1 miles west of Porters Pass*'). Specimens can be plants, animals, fungi, soil, bacteria or geological samples. Currently only a fraction of these records can be mapped, because manual

interpretation of the often complex descriptions is costly, and existing automated tools are limited in their ability to accurately georeference such descriptions, focusing mainly on place names (also known as toponyms), and ignoring or using over-simplified models to interpret spatial language terms (e.g. *above*, *south of*, *near*). They also depend upon gazetteers that may omit many local names and their variations.

The BioWhere project is pioneering methods to georeference descriptions that include vague and often context-sensitive spatial terms. While some models for vague spatial relation terms have been developed, referred to variously as spatial templates, spatial acceptance models or spatial applicability models, they have not incorporated the range of contextual variations in the use of these terms, yet these have been shown to be important for accurate interpretation (Vasardani et al., 2013). The work described here is advancing the georeferencing of text by developing recent trends towards learning from 'language in use', rather than theoretical models of spatial relation terms. This approach allows context and vagueness to be accommodated, and enables more complex spatial location expressions to be georeferenced than has previously been possible. In addition to modelling prepositions (e.g. *on*, *north of*), on which some work has been done (Hall & Jones, 2021), we also consider verbs (e.g. *crosses*), multi-word expressions (*in line with*), parts of objects (*the centre of)* and geographic feature types (*main highway*). We are applying regression versions of deep learning methods that learn directly from text. We are also exploiting features derived from text descriptions, including linguistic (e.g. adverbial qualifiers, geographical features), place name (e.g. feature type) and physical (e.g. population density, habitat etc.) aspects.

With a focus on collections that store specimens from New Zealand (NZ), the project is also addressing the current lack of comprehensive gazetteers that include historical names. Many of the early specimens (going back to Cook's voyages to map NZ starting in 1769) reference place names that are not stored in current gazetteers, including some of the historical indigenous names used by New Zealand Māori (the indigenous people of Aotearoa/New Zealand). The project is developing a rich gazetteer that will store cultural and historical knowledge about Māori place names. These toponyms often provide details about the landforms and cultural practices in the places they name. Another innovation is the development of a self-learning gazetteer that is reverse-engineering the location of place names extracted from biological specimen location descriptions that are already georeferenced.

The project is funded by the New Zealand Ministry of Business, Innovation and Employment Endeavour research fund and began in late 2021. The project team involves computer science and geospatial researchers, specialists in indigenous place names, and biological collections experts. While there is a focus on the geographical area of NZ, the methods are intended to be generic and specimens from biological collections from around the world, including the Natural History Museum UK, Kew Gardens UK, and the collection of the international network known as the Global Biodiversity Information Facility (GBIF), are being marshalled to train and test the models.

In this paper, we highlight the challenges involved in georeferencing complex spatial location descriptions. After summarizing the current state of the art, we describe a set of AI-based, spatial, natural-language processing methods for georeferencing that are being developed in the BioWhere project.

2 Potential Impacts

The methods developed in the project have the potential to enable mapping of vast collections of biological resources, improving management of biodiversity, pests and diseases, and monitoring in the face of climate change. In the NZ context, the specific challenges that the methods developed in this project will help address include the monitoring and management of:

- exotic predators (rats, stoats etc.) that are endangering native birdlife (Department of Conservation, 2021);
- myrtle rust, which arrived in NZ in 2017, attacks native myrtle and dependent species, and severely impacts NZ industries (e.g. manuka honey, feijoa) (Biosecurity New Zealand, 2019);
- Kauri dieback (including reducing its impacts on tourism) (Mau, 2018).

The methods that are being developed will also have applications beyond the georeferencing of biological specimen descriptions, such as mapping current and historical phenomena, exploiting social media to map disaster impacts (Hameed et al., 2022), identifying the location of phenomena from journal papers (Acheson & Purves, 2021; Kordjamshidi et al., 2015; Scott et al., 2021), and locating other types of scientific information such as the site of soil surveys from digitized archival records.

3 The Challenges

3.1 Data Challenges

The task of georeferencing biological collections presents several challenges. Firstly, many organizations hold vast collections, submitted by members of the public or professional data collectors, only a small proportion of which are digitized. These may consist of the actual specimen, along with field notes or a record card, usually including some indication of location, often in the form of a place name or locality description, or sometimes a map grid reference. More recently, GPS coordinates may have been recorded. The process of digitizing a specimen involves carefully photographing it and recording its metadata in a computerized data structure. Many museums and herbaria have ongoing digitizing programmes (e.g. Kew

Gardens in the UK, the world's largest collection of plants and fungi, is currently digitizing its holdings that exceed 8 million records[1]).

The BioWhere project is confining its attention to locality description records that are already digitized. A further subset of these records are already georeferenced, and these are being used as training data for the models being developed. However, there are substantial challenges regarding the quality of the coordinates, largely resulting from the methods used to assign them (in some cases many years ago and not fully recorded). The difficulties include:

- The coordinates of a point in the map grid square (e.g. centroid) supplied by the collector. The accuracy is thus dictated by the size of the grid square.
- The coordinates of a place name within a locality description supplied by the collector, retrieved from a gazetteer. The accuracy depends on the degree to which the place name represents the actual specimen location.
- For recently collected specimens, GPS coordinates obtained by the collector, which can be of variable and unknown quality, as the equipment and skill of the operator are unknown. Specimens may be collected by members of the public, volunteer groups or professionals.

Coordinates may also be manually determined through a time-consuming process involving examination of maps and aerial photographs to try to identify the location referred to in the locality description. Many museums and other agencies have ongoing programmes for manual georeferencing, but at an optimistic estimate of 4.5 minutes per record (based on our experience of the task), georeferencing of the 5 million currently digitized but not georeferenced records held in 26 NZ collections (Nelson et al., 2015) would take 298 person-years, and the equivalent global figure for GBIF (>100 million un-georeferenced records) would take 5,952 person-years.

Furthermore, some of the coordinates in the collections' data contain gross errors, including transposed figures and missing negative signs on latitude and longitude (thus referring to the incorrect hemisphere), as well as rounded latitudes and longitudes (e.g. to the nearest degree, and thus potentially >100 km from the correct location).

3.2 Locality Description Challenges

Most existing methods of georeferencing documents (though with some notable exceptions discussed in Section 4) are based entirely on the detection and geocoding of place names. The accurate georeferencing of specimen locality descriptions, however, needs to take into account not just place names, but also the multiple other spatial terms used to describe a location relative to specific named places. For example, accurate georeferencing of the description '*On*

[1] https://www.gov.uk/government/news/historic-kew-gardens-collection-to-go-digital-in-major-boost-for-climate-change-research

bench[2] *200ft above and south of main highway, 1.1 miles west of Porters Pass*' could require us to complete the following steps (the generic task required for each step is shown in square brackets):

1. Find the location of *Porters Pass*. [***Toponym georeferencing: detecting and determining coordinates for place names***] Current toponym georeferencing methods, including those that use sophisticated language modelling disambiguation techniques, are dependent on high-quality detailed local gazetteers for precise geocoding; they do not exploit all relevant contextual data such as collector itineraries, habitat and other environmental data.

2. Determine the correct nesting of the different phrases within the locality description. [***Determining dependencies between spatial relational phrases***] It is not clear whether *1.1 miles west of Porters Pass* refers to the specimen location independently, or whether it is relative to the bench or the main highway (see Table 1) (Khan et al., 2013; Kordjamshidi et al., 2011). One approach is to consider that a comma signifies a new, unnested prepositional phrase (Option 1 in Table 1), but the use of punctuation is far from consistent in spatial descriptions in the English language. Examination of a map of the area (Figure 1) suggests that Option 3 in Table 1 is a likely interpretation, since the highway runs approximately west from Porters Pass.

Table 1: Alternative Interpretations of Sample Locality Description

(1)	The specimen was collected: → On bench 200ft above and south of main highway AND → 1.1 miles west of Porters Pass
(2)	The specimen was collected: → On bench 200ft above and south of main highway WHICH IS → 1.1 miles west of Porters Pass
(3)	The specimen was collected: → On bench WHICH IS → 200ft above and south of main highway AND → 1.1 miles west of Porters Pass

[2] In this case, "*bench*" refers to the small terrace from which the specimen was collected.

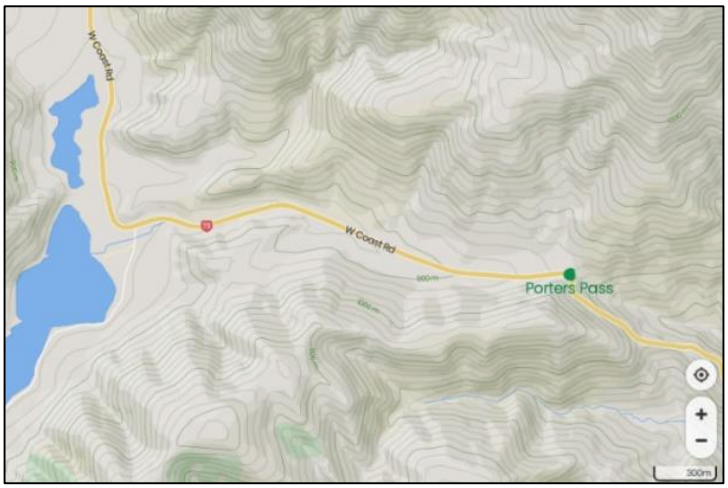

Figure 1: Porters Pass area (© OpenStreetMap contributors)

3. Determine a location relative to *Porters Pass* that takes into account the specified distance (*1.1 miles*) and direction (*west*). [***Spatial relation acceptance region (applicability) modelling: determining locations that a spatial relation refers to***] Natural-language cardinal directions are essentially imprecise (Hall et al., 2011, 2015), as indeed can be distances. Furthermore, the ways in which these terms are applied may depend on the context. For example, one end of a beach that runs north-north-west may be referred to as 'the northern end', even though the beach does not run directly north–south. The shape of the landscape thus biases the interpretation of the terms. As can be seenin Figure 1 the highway near Porters Pass runs approximately west (but slightly north), and considering our selected interpretation of the nesting of the prepositional phrases, it is likely that the bench referred to is close to the highway, rather than precisely west of Porters Pass.

4. Identify the location of the unnamed *main highway* that the description refers to. [***Determining locations of generic topographic features***] This requires the use of additional data sources (containing highway locations) to identify highways that qualify as *main* in the vicinity of the location referenced in the last prepositional phrase (*1.1 miles west of Porters Pass*).

5. Determine a set of locations that are *south of* the *main highway*. [***Spatial relation acceptance region (applicability) modelling***] While similar to step 2, there is the additional challenge that the distance from the main highway is unspecified. If we assume that the final prepositional phrase '*1.1 miles west of Porters Pas*' refers to the location of the bench, we could determine an approximate location along the main highway from which to search southwards, but significant uncertainty would still be present.

6. Identify benches that are *200ft above* the main highway. **[*Determining locations of generic topographic features*]** If we assume that the last prepositional phrase refers to

the location of the bench and the direction southwards gives us an approximate location, then a terrain model could indicate heights that are 200 ft above the highway, but we are left with the challenge of identifying the geomorphological feature (bench), which will not usually be explicitly identified on a topographic map.

This example indicates some of the challenges that are posed for automated georeferencing of locality descriptions found in biological collections. Additional challenges include:

- abbreviated terms (e.g. SE for south east, cnr for corner);
- coreferences or anaphora (Manzoor & Kordjamshidi, 2018) (e.g. in the expression '*Bethells Beach, mouth of Waitakere River where it runs over sand flats*', the automated process must understand that *it* refers to the *Waitakere River*);
- adjectives to specify a particular instance of a geographic feature type (e.g. the expression '*Titahi Bay, in first valley N of Mount Cooper, and just S of the steep ridge leading to summit*' requires identification of the **first** valley and a **steep** ridge – see Figure 2);
- adverbs that specify the spatial relation term more precisely (e.g. '*small gully on low slopes of Whitehorn Ridge immediately south of Frosty Creek*' specifies that the location is *immediately* south, rather than approximately south, as is often the case for natural-language descriptions – see Figure 3).

Figure 2:
Titahi Bay area (©OpenStreetMap contributors)

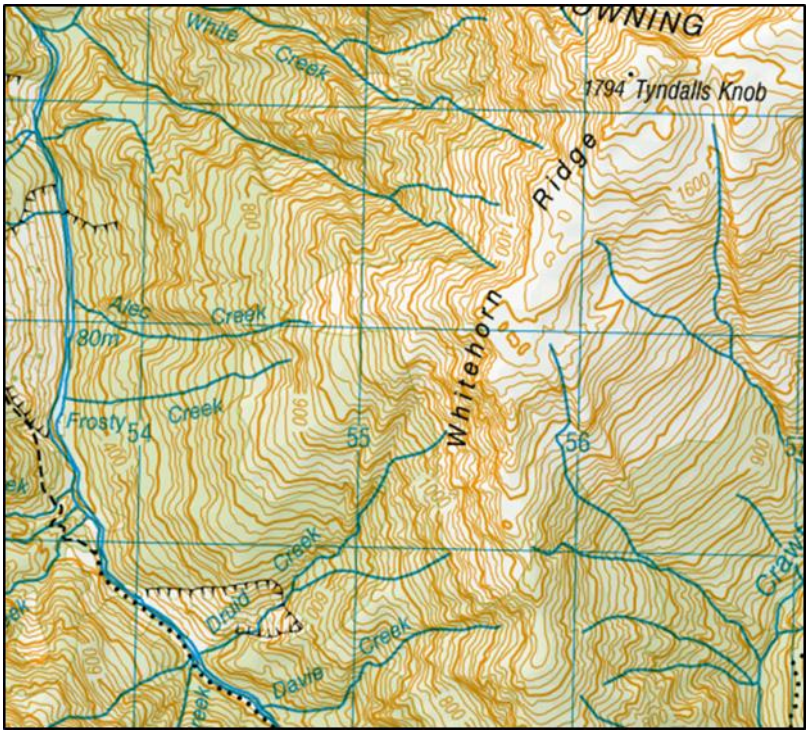

Figure 3: Whitehorn Ridge area (©NZ Geographic Board under CC-BY-4.0)

Finally, taking into account the uncertainty associated with automatically generated coordinates is key to the effective use of georeferenced data. We are estimating the uncertainty resulting from the input data sources, and are modelling and recording it to enable end-users of the coordinates to use them appropriately (Guo et al., 2008; Wieczorek et al., 2004).

3.3 Gazetteer Challenges

Historical records that are up to 200 years old present a challenge for gazetteers available in NZ. While some gazetteers store historical names (for example, the New Zealand Geographic Board Gazetteer[3]), these gazetteers are often relatively recent. Place names may have changed over time, may be misspelt, may have multiple valid versions, or may be abbreviated in the location descriptions. Many of these challenges have been recognized (Hill, 2006), but it is often not feasible to conduct the necessary historical investigations to fully record historical place names.

Table 2 provides two examples of toponyms that have changed over time. The location known as Ahuriri has changed its name multiple times since the name 'Ahuriri' was first used in 1851. To further confuse the situation, the modern-day town of Napier is now known as Ahuriri

[3] https://gazetteer.linz.govt.nz/

(many NZ towns and cities have both European and Māori names), even though Ahuriri is also a particular part of the town of Napier (the port area). The Houhou example illustrates the multiple (mis)spellings and transliterations used for a Māori name. Automated georeferencing of location descriptions in historical collections must be able to cater for all of these possible cases, and to disambiguate different locations that share the same name, which may also have changed over time.

Table 2: Examples of Historical Place Names

Ahuriri		**Houhou**	
Hawke's Bay		Westland	
Earliest mention 1851		Earliest mention 1866	
Industrial suburb 1.6km NW of Napier Central Post Office.		Gold mining settlement 5km E of Hokitika.	
The Spit	1869–1896	Ho Ho Creek	1866–1868
Port Ahuriri	1869–1896	Hau Hau	1868–1872
Spit	1896–1903	Hoho	1873–1893
Ahuriri	1903–	Ho Ho	1893–1944
		Houhou	1944–
		Alt names: Holo	

4 Previous Work

Good progress has been made on georeferencing place names (Purves et al., 2018; Wang et al., 2019), but the substantial limitations of existing methods have been highlighted (Gritta et al., 2018). Recognition of place names, which is required for successful toponym georeferencing, is still largely dependent on, but hindered by, the shortcomings of current place name gazetteers. Location descriptions may refer to place names no longer in use or which have changed. Particular domains may also pose problems. In the field of geology, for example, documents may refer to geological features that are not identified by standard named-entity recognition methods (e.g. shields, cratons), are not present in conventional gazetteers, and hence require access to supplementary lists of such feature names (Leveling, 2015). The challenges of working with historical place names have been recognized (Ardanuy et al., 2019), and the need for substantial manual work to add these names to gazetteers is a barrier. In the BioWhere project, we address this through the development of a self-learning gazetteer, which will learn place name locations from locality descriptions that are already georeferenced, analogous to Chen et al. (2018). Further work has been conducted on the creation of gazetteers that store rich cultural information, including for NZ Māori (Te Rūnanga o Ngāi Tahu, 2023). We extend this work and leverage the enriched gazetteer to assist with georeferencing.

In the biological collections literature, some limited progress has been made in georeferencing directional offsets (e.g. *3 km north of Lincoln*) (Guralnick et al., 2006; Rios & Bart Jr, 2014), but there has been little progress in georeferencing complex expressions (Leidner & Lieberman, 2011), despite widespread acknowledgment of the need (Doherty et al., 2011; Guo et al., 2008; Hill, 2006; van Erp et al., 2015; Wieczorek et al., 2004). The work of Y. Liu et al. (2009) is a notable exception in highlighting some of the problems of modelling complex location descriptions and proposing probabilistic solutions. These authors focus on evaluating uncertainty and present some default approaches to interpretation of different types of spatial relation, stressing the importance of context. However, context is not incorporated in their models, and there is no systematic evaluation of their methods. Nor do such studies the challenges of natural-language processing to identify individual spatial relations, or of parsing the interdependencies between multiple spatial relations in complex expressions.

While the challenge of georeferencing location descriptions may be relevant across a range of domains, studies applying methods in specific areas are rare. In addressing the more general (i.e. cross-domain) task, there have been advances in detecting the presence of relative geospatial expressions in text (F. Liu et al., 2014; Radke et al., 2019; Stock et al., 2013; Zhang et al., 2009) as a first step in the georeferencing of large bodies of text. Human subjects testing and data mining have been used to model applicability (acceptability) regions for specific vague spatial relation terms (e.g. Hall et al., 2011, 2015; Hall & Jones, 2021), but they take limited account of contextual factors. However, a vast range of research in linguistics has shown that acceptance areas of spatial relation terms vary with the situation of use, including the location of nearby objects, knowledge of how objects work, the purpose of the utterance, and the roles of the parties communicating (Stock & Hall, 2018; Stock & Yousaf, 2018). More recent work has incorporated context in limited ways, including object types (Lan et al., 2012; Malinowski & Fritz, 2015; Platonov & Schubert, 2018) or size (Collell et al., 2018), and text embeddings, which are a vector representation of the semantics of the terms in the description (Bisk et al., 2018; Collell et al., 2018; Malinowski & Fritz, 2015). These latter studies are all in so-called 'tabletop' indoor environments or describe locations in images that take no account of the geographical factors characterizing our own area of research. They address problems including the interpretation of natural-language instructions to robots to move objects (e.g. in a factory environment) (Bisk et al., 2018; Platonov & Schubert, 2018), and the retrieval of photographs in response to queries (e.g. *find me a photo that shows a boy on a horse*) (Collell et al., 2018; Malinowski & Fritz, 2015). In geographical environments, previous work has addressed the generic task of georeferencing, considering only broad urban vs rural contexts (Hall et al., 2011; Hall & Jones, 2021) and characteristics such as scale and geometry type (Stock & Yousaf, 2018), while place size and prominence were addressed in Chen et al. (2018). Our work goes beyond earlier methods in including a much broader set of linguistic, environmental and collection-based contextual factors and in applying more recent transformer-based approaches.

On the specific task of extracting spatial relations from text, which we address in this project in order to identify relevant terms to model (see Section "Parsing of Locality Descriptions"),

most work to date has not been specific to geo-spatial contexts, but rather in generic, biomedical and scientific domains (Datta & Roberts, 2020; Kordjamshidi et al., 2011, 2015; Mazalov et al., 2015; Pustejovsky et al., 2015). A recent exception is Qiu et al. (2022), who focus on the Chinese language. Relation extraction from specimen collection records is complicated in that the primary located object of a relation (the specimen itself) is not usually mentioned (Khan et al., 2013) and, as indicated, there can be ambiguous interdependencies between phrases (the correct interpretation of which might be learnt from coordinate-based training data). We are building on recent generic transformer-based relation extraction methods such as Zhong and Chen (2021) to address this challenge.

5 Methods

5.1 Gazetteer Model

As has been discussed, most locality descriptions describe location using spatial relation terms relative to named places (as reference objects, such as Porters Pass), and thus effective georeferencing relies on comprehensive gazetteer data. The logical model for the BioWhere gazetteer is based on the Alexandria Digital Library (ADL) Gazetteer Content Standard (Hill, 2006), which provides extensive consideration of many of the aspects that are required for a functional gazetteer. The model has been extended in two areas.

Firstly, Māori place names have a key role in the history of Aotearoa New Zealand, and many are used in the biological collections' locality descriptions. Since the arrival of Europeans in NZ around 250 years ago, knowledge of some of these names has been lost due to the dominance of European names, at least in official records. Although some are still remembered and used by *hapū* (subtribe social grouping) and *iwi* (tribe), many are not included in existing gazetteers. Māori place names often encapsulate information about the physical characteristics of the landscape and the cultural practices that were conducted there, and thus are themselves important historical records. The BioWhere gazetteer extends the etymology component of the ADL model to refer to *whakapapa* (lineage or descent, genealogy). This may include information about the stories, background, history, cultural practices or meaning associated with the place name, the details of where that information originated from (different *hapū* or *iwi* may use different names for the same place, or may attach different stories to them), as well as citations of source documents. We also add information about whether toponyms are transliterations (Māori transliterations of European names are sometimes used). The collection, storage and sharing of this information is an additional cultural dimension of the project that complements the georeferencing work, but in the future some of the knowledge about landscape and practices may also be useful as context to aid in more accurate georeferencing (see Section "Spatial Relation Models").

Secondly, it is common practice for biological collections to maintain gazetteers that are not confined to toponyms: they may also include records that are themselves complex locality

descriptions. This may be a result of collection practices, in which multiple specimens may be collected in the same place or where repeat collections may be made over time. Thus if a complex locality description has been georeferenced, it is useful to store it in case the same description is used again (or in case there is a duplicated record from a different collection, which is a common practice). We exploit this in our automated georeferencing methods by learning georeferences not only for identical locality descriptions, but also for those that are semantically similar. For example, '*beside the bridge, 3 km north of Ōrewa*', '*next to the bridge, 3 km north of Ōrewa*' and '*beside the ford, 3 km north of Ōrewa*' may all refer to the same location.

5.2 Automated Context-Sensitive Georeferencing

Our method for georeferencing locality descriptions involves three steps: first we parse the descriptions, second we build models to predict the interpretation of each spatial relation term, and finally we combine the different relational phrases found in the description to predict a location and calculate associated error.

Parsing of Locality Descriptions

We use Named Entity Recognition (NER) to detect not only place names within the locality descriptions, but also other kinds of term, including geographic features (e.g. *bench* and *highway* in our running example); spatial relation terms (terms that describe the location where the specimen was collected relative to a named place), and their subtypes (e.g. topological, proximal relations) as per the example in Figure 4; and adverbs that provide additional information about the spatial relation term (e.g. ***directly*** *opposite the church*). We are comparing a range of NER methods, including spaCy and several BERT-based transformer deep learning models (BERT, DistilBERT, ALBERT, RoBERTa, XLNet), which we are retraining using our own geographic corpora. In this we are following earlier work (Radke et al., 2022) that demonstrates that these methods outperform previous approaches to the detection of spatial relation terms (Kordjamshidi et al., 2011; Radke et al., 2019).

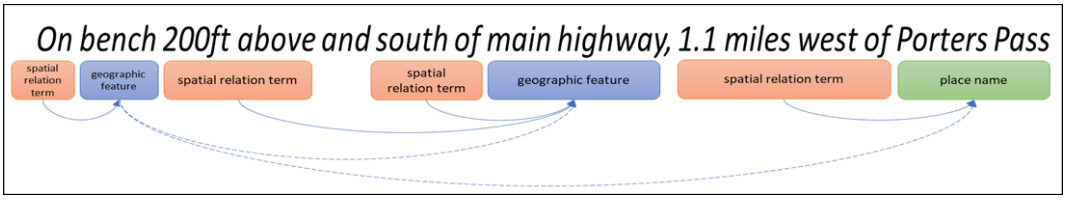

Figure 4: Example of Parsing of Locality Description

Following extraction of the entities, we are developing methods for *relation* extraction to incorporate relevant geographic context. First, we apply a generic baseline method (Zhong & Chen, 2021) to extract the syntactic relations between words in the locality description (as shown with arrows in Figure 4). This operation again uses BERT-based transformer methods. The method inserts markers to indicate the entity types in order to help extract the types of

relations between pairs of entities, which gives considerably better results than popular models that perform NER and relation extraction jointly. Before tackling the more complex locality descriptions found in our data, we applied this method to the IAPR TC-12 data set[4], which tags triples of <trajector, spatial indicator, landmark>, with 'spatial indicator' being a spatial relation term (e.g. *above*, *south of*), 'trajector' being the object that is being located, and landmark being a reference object (e.g. <trajector>bench</trajector><spatial indicator>above</spatial indicator><landmark>main highway</landmark>. We extract binary relations between each spatial indicator and its trajector and landmark, and identify their type (direction, distance or region/topological), giving a total of six relation types. We represent relations as a concatenation of the span embeddings of each pair of entities that are under consideration. Running the model for 10 epochs, we achieved the results shown in Table 3. As can be seen, this method provides promising results, predicting trajector and landmark, and their types, with F1 values of 0.95 and 0.93 respectively.[5] Our results are similar to those of Datta and Roberts (2020) in the biomedical domain and exceed those of Zhong and Chen (2021) for scientific relations (see Table 3).

In future work, we will evaluate this method for predicting triples (which we will derive from the binary predictions), in order to compare it with other methods in the spatial domain (Mazalov et al., 2015; Shin et al., 2020). We will tailor the method to the geographic context, including training the models with large geographic text corpora and including additional input information in the models to reflect geographic context.

Spatial Relation Models

Another key innovation of the project is the development of context-sensitive deep learning models for the prediction of locations referenced by spatial relation terms.

Georeferencing of individual toponyms: Our methods for individual toponym-georeferencing will combine location language models that characterize locations by the words that describe them (DeLozier et al., 2015; Melo & Martins, 2017) with novel exploitation of our own and public environmental and collections data. Certain species only exist in particular environments, and specimen collectors work within limited geographic areas with specific itineraries, hence restricting the number of possible locations.

[4] https://github.com/kolomiyets/sprl2013

[5] F1 is the harmonic mean of precision and recall, 'precision' indicating the proportion of predicted relations that were correct compared with the actual (ground truth) data, and 'recall' the proportion of actual relations that were correctly predicted.

Table 3: Relation Extraction Results

Reference	Dataset	Relations Detected	F1
Our implementation, BERT (Zhong & Chen, 2021)	IAPR TC-12	Trajector, Landmark and their respective types (direction, distance, region)	0.95, 0.93
(Mazalov et al., 2015), convolutional neural networks	IAPR TC-12	Trajector+Spatial Indicator+Landmark and type (direction, distance, region)	0.72
(Datta & Roberts, 2020), BERT, BiLSTM, XLNet	Biomedical domain	Trajector, Landmark	0.94, 0.96
(Zhong & Chen, 2021), BERT	ACE05 (Scientific documents)	Six relation types among scientific entities (e.g. method, task)	0.69
(Shin et al., 2020), BERT	ISO-Space	Three different schemes involving different spatial relations.	0.61

Georeferencing specimen records: One of the key advances of this project is the incorporation of contextual information into machine learning models that predict the location detailed in a specimen locality description. In order to improve the accuracy of georeferencing, we are building models that more closely reflect the ways in which humans use spatial relation language, and take both the geographic and the linguistic contexts into account to predict the distance and direction (between a reference location and the specimen) associated with individual spatial relation terms. Thus far, we have built models for 16 spatial relation terms, 8 of which are direction-related (*north*, *south-west* etc.), and 8 others (e.g. *near*, *at*, *above*, *below*, *end of*) (Liao et al., 2022). Input features for the machine learning models include:

- The embeddings, geometry types, and scale of the feature types (also feature types of the toponyms) used in the locality descriptions. These features capture some aspects of the semantics of the objects mentioned in the descriptions that are likely to affect the ways in which spatial relation terms are used (for example, *near* is likely to refer to a smaller distance in relation to a post office than to a mountain).
- Characteristics of the area surrounding the place names referred to in the description (e.g. population density).

This approach achieved improvements in the mean absolute error of the predicted *distance* between the specimen collection location and the reference toponym location ranging from 15% (for *at*) to 60% (for *south-west of*[6]) over the baseline. For prediction of *direction*, the

[6] Although 'south-west of' is a spatial relation that specifies direction, it is used to describe objects within specific distance ranges, and thus it is possible to use cardinal direction relations to predict distance as well as direction.

maximum improvement over the baseline was 30% (for *south-west of*). No improvement was found for some spatial relation terms, including *north of* (Liao et al., 2022).

We are further improving these methods by creating and combining two forms of transformer model that we will fine-tune to learn the referenced location of a specimen, including a wider set of contextual factors. As input, one form takes the entire expression concatenated with associated contextual data. The other, similar to Liao et al. (2022), predicts relative locations from individual spatial relation adverbial phrases (e.g. '*1.1 miles west of Porters Pass*') extracted from the descriptions, before combining them. Using regression, we can predict the distance and/or direction offsets from the reference location (e.g. Porters Pass). We are also developing classification versions of the models that will predict acceptability of spatial grid squares, again relative to the location of a reference object (see Collell et al., 2018; Malinowski & Fritz, 2015). Additional contextual factors that we plan to include in the models are:

a. semantics of the spatial relation terms and their senses (from Aflaki et al., 2022);
b. geomorphological, climate and habitat data;
c. collection features (date of collection, region, collection type, collector, habitat type, geology, collector itinerary) (Chapman & Wieczorek, 2020; Nicolson & Tucker, 2017);
d. relation-specific features (e.g. orientation relative to road centreline or other prominent axis is important for relations such as left and right).

The approach models each spatial relation term in a locality description individually, then combines the results to identify the region that the locality description refers to, and the associated error. The simplest approach will use the intersection of the different spatial relation models, but in later work we plan to develop more advanced approaches that take into account complex inter-relations between the prepositional phrases (such as those described in Section 3.2), and apply semantic masks to exclude areas where the specimen could not have been collected (e.g. a lake if the species of interest is terrestrial).

6 Conclusions

Mapping the wealth of unexplored data in biological collections globally has great value for the study of changes in biota over time and in space, with the potential to provide essential data for climate-change monitoring and management. Other applications include pest and disease management. However, many specimens in these vast collections are currently georeferenced only through natural-language descriptions. The georeferencing of these descriptions presents significant challenges:

- the quality of available training data in existing collections varies widely due in part to historical processes used for georeferencing, use of grid references (which are very approximate), consideration of place names without any associated further descriptions, unreliable GPS coordinates, or manual interpretation of the locations;

- the vague and context-sensitive nature of locational (human) language, which often depends on human understanding of the situation in which the terms are used;
- the reliance of many location descriptions on historical toponyms for which georeferences are not available.

This paper has described the challenges, presented early results, and set out an agenda for future research directions, including the creation of a culturally rich gazetteer, and methods for parsing location descriptions and predicting georeferences by incorporating context. There is great potential to marshal advances in natural-language processing approaches, combined with the latest developments in geospatial science, to enable this vast store of data to be used to address current and future environmental challenges.

References

Acheson, E., & Purves, R. S. (2021). Extracting and modeling geographic information from scientific articles. PLOS ONE, 16(1), e0244918. https://doi.org/10.1371/journal.pone.0244918

Aflaki, N., Stock, K., Jones, C. B., Guesgen, H., & Morley, J. (2022). An empirical study of the semantic similarity of geospatial prepositions and their senses. Spatial Cognition & Computation, 0(0), 1–45. https://doi.org/10.1080/13875868.2022.2111683

Ardanuy, M. C., McDonough, K., Krause, A., Wilson, D. C. S., Hosseini, K., & van Strien, D. (2019). Resolving places, past and present: Toponym resolution in historical british newspapers using multiple resources. Proceedings of the 13th Workshop on Geographic Information Retrieval, 1–6. https://doi.org/10.1145/3371140.3371143

Biosecurity New Zealand. (2019). New Zealand Myrtle Rust Strategy 2019-2023. https://www.myrtlerust.org.nz/assets/Uploads/Myrtle-Rust-Strategy-web3.pdf

Bisk, Y., Shih, K., Choi, Y., & Marcu, D. (2018). Learning Interpretable Spatial Operations in a Rich 3D Blocks World. Proceedings of the AAAI Conference on Artificial Intelligence, 32(1), Article 1. https://ojs.aaai.org/index.php/AAAI/article/view/12026

Chapman, A. D., & Wieczorek, J. R. (2020). Georeferencing best practices. Version 1.0. GBIF Secretariat. https://docs.gbif.org/georeferencing-best-practices/1.0/en/

Chen, H., Winter, S., & Vasardani, M. (2018). Georeferencing places from collective human descriptions using place graphs. Journal of Spatial Information Science, 2018(17), 31–62. https://doi.org/10.5311/JOSIS.2018.17.417

Collell, G., Gool, L. V., & Moens, M.-F. (2018, April 27). Acquiring Common Sense Spatial Knowledge Through Implicit Spatial Templates. Thirty-Second AAAI Conference on Artificial Intelligence. Thirty-Second AAAI Conference on Artificial Intelligence. https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16232

Datta, S., & Roberts, K. (2020). Spatial Relation Extraction from Radiology Reports using Syntax-Aware Word Representations. AMIA Joint Summits on Translational Science Proceedings. AMIA Joint Summits on Translational Science, 2020, 116–125.

DeLozier, G., Baldridge, J., & London, L. (2015). Gazetteer-independent toponym resolution using geographic word profiles. Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, 2382–2388.

Department of Conservation. (2021, April 21). Predator Free 2050. Predator Free 2050. https://www.doc.govt.nz/nature/pests-and-threats/predator-free-2050/

Doherty, P., Guo, Q., Liu, Y., Wieczorek, J., & Doke, J. (2011). Georeferencing Incidents from Locality Descriptions and its Applications: A Case Study from Yosemite National Park Search and Rescue. Transactions in GIS, 15(6), 775–793. https://doi.org/10.1111/j.1467-9671.2011.01290.x

Gritta, M., Pilehvar, M. T., Limsopatham, N., & Collier, N. (2018). What's missing in geographical parsing? Language Resources and Evaluation, 52(2), 603–623. https://doi.org/10.1007/s10579-017-9385-8

Guo, Q., Liu, Y., & Wieczorek, J. (2008). Georeferencing locality descriptions and computing associated uncertainty using a probabilistic approach. International Journal of Geographical Information Science, 22(10), 1067–1090. https://doi.org/10.1080/13658810701851420

Guralnick, R. P., Wieczorek, J., Beaman, R., Hijmans, R. J., & BioGeomancer Working Group. (2006). BioGeomancer: Automated georeferencing to map the world's biodiversity data. PLoS Biology, 4(11), e381. https://doi.org/10.1371/journal.pbio.0040381

Hall, M., & Jones, C. B. (2021). Generating geographical location descriptions with spatial templates: A salient toponym driven approach. International Journal of Geographical Information Science. https://doi.org/10.1080/13658816.2021.1913498

Hall, M., Jones, C. B., & Smart, P. (2015). Spatial Natural Language Generation for Location Description in Photo Captions. In S. I. Fabrikant, M. Raubal, M. Bertolotto, C. Davies, S. Freundschuh, & S. Bell (Eds.), Spatial Information Theory (pp. 196–223). Springer International Publishing.

Hall, M., Smart, P. D., & Jones, C. B. (2011). Interpreting spatial language in image captions. Cognitive Processing, 12(1), 67–94. https://doi.org/10.1007/s10339-010-0385-5

Hameed, S., Stock, K., Francis, S., Li, D., Yandamuri, H., Prasanna, R., Jones, C. B., & Liberatore, F. (2022). A Pipeline for Geospatial Situational Awareness of Disasters Through Social Media Text. ISCRAM Asia Pacific, Melbourne, Australia.

Hill, L. L. (2006). Georeferencing: The geographic associations of information. Mit Press.

Khan, A., Vasardani, M., & Winter, S. (2013). Extracting Spatial Information From Place Descriptions. Proceedings of The First ACM SIGSPATIAL International Workshop on Computational Models of Place, 62:62-62:69. https://doi.org/10.1145/2534848.2534857

Kordjamshidi, P., Roth, D., & Moens, M.-F. (2015). Structured learning for spatial information extraction from biomedical text: Bacteria biotopes. BMC Bioinformatics, 16(1), 129. https://doi.org/10.1186/s12859-015-0542-z

Kordjamshidi, P., Van Otterlo, M., & Moens, M.-F. (2011). Spatial Role Labeling: Towards Extraction of Spatial Relations from Natural Language. ACM Trans. Speech Lang. Process., 8(3), 4:1-4:36. https://doi.org/10.1145/2050104.2050105

Lan, T., Yang, W., Wang, Y., & Mori, G. (2012). Image Retrieval with Structured Object Queries Using Latent Ranking SVM. In A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, & C. Schmid (Eds.), Computer Vision – ECCV 2012 (pp. 129–142). Springer. https://doi.org/10.1007/978-3-642-33783-3_10

Leidner, J. L., & Lieberman, M. D. (2011). Detecting Geographical References in the Form of Place Names and Associated Spatial Natural Language. SIGSPATIAL Special, 3(2), 5–11. https://doi.org/10.1145/2047296.2047298

Leveling, J. (2015). Tagging of temporal expressions and geological features in scientific articles. Proceedings of the 9th Workshop on Geographic Information Retrieval, 1–10. https://doi.org/10.1145/2837689.2837701

Liao, R., Das, P. P., Jones, C. B., Aflaki, N., & Stock, K. (2022). Predicting Distance and Direction from Text Locality Descriptions for Biological Specimen Collections. In T. Ishikawa, S. I. Fabrikant, & S. Winter (Eds.), 15th International Conference on Spatial Information Theory (COSIT 2022) (Vol. 240, p. 4:1-4:15). Schloss Dagstuhl – Leibniz-Zentrum für Informatik. https://doi.org/10.4230/LIPIcs.COSIT.2022.4

Liu, F., Vasardani, M., & Baldwin, T. (2014). Automatic Identification of Locative Expressions from Social Media Text: A Comparative Analysis. Proceedings of the 4th International Workshop on Location and the Web, 9–16. https://doi.org/10.1145/2663713.2664426

Liu, Y., Guo, Q. H., Wieczorek, J., & Goodchild, M. F. (2009). Positioning localities based on spatial assertions. International Journal of Geographical Information Science, 23(11), 1471–1501. https://doi.org/10.1080/13658810802247114

Malinowski, M., & Fritz, M. (2015). A Pooling Approach to Modelling Spatial Relations for Image Retrieval and Annotation. ArXiv:1411.5190 [Cs]. http://arxiv.org/abs/1411.5190

Manzoor, U., & Kordjamshidi, P. (2018). Anaphora Resolution for Improving Spatial Relation Extraction from Text. Proceedings of the First International Workshop on Spatial Language Understanding, 53–62. https://doi.org/10.18653/v1/W18-1407

Mau, A. (2018, July 3). Tāne Mahuta could soon be infected with fatal Kauri dieback disease. Stuff. https://www.stuff.co.nz/national/105184309/tne-mahuta-could-be-dead-within-a-year-from-kauri-dieback

Mazalov, A., Martins, B., & Matos, D. (2015). Spatial role labeling with convolutional neural networks. Proceedings of the 9th Workshop on Geographic Information Retrieval, 1–7. https://doi.org/10.1145/2837689.2837706

Melo, F., & Martins, B. (2017). Automated Geocoding of Textual Documents: A Survey of Current Approaches. Transactions in GIS, 21(1), 3–38. https://doi.org/10.1111/tgis.12212

Nelson, W., Breitwieser, I., Fordyce, E., Bradford-Grieve, J., Penman, D., Roskruge, N., Trnski, T., Waugh, S., & Webb, C. (2015). National Taxonomic Collections in New Zealand. Royal Society of New Zealand+ Appendices. Royal Society of NZ. https://www.royalsociety.org.nz/assets/Uploads/Report-National-Taxonomic-Collections-in-New-Zealand-2015.pdf

Nicolson, N., & Tucker, A. (2017). Identifying Novel Features from Specimen Data for the Prediction of Valuable Collection Trips. In N. Adams, A. Tucker, & D. Weston (Eds.), Advances in Intelligent Data Analysis XVI (pp. 235–246). Springer International Publishing. https://doi.org/10.1007/978-3-319-68765-0_20

Platonov, G., & Schubert, L. (2018). Computational Models for Spatial Prepositions. Proceedings of the First International Workshop on Spatial Language Understanding, 21–30. https://doi.org/10.18653/v1/W18-1403

Purves, R. S., Clough, P., Jones, C. B., Hall, M. H., & Murdock, V. (2018). Geographic Information Retrieval: Progress and Challenges in Spatial Search of Text. now. https://ieeexplore.ieee.org/document/8311405

Pustejovsky, J., Kordjamshidi, P., Moens, M.-F., Levine, A., Dworman, S., & Yocum, Z. (2015). SemEval-2015 Task 8: SpaceEval. Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), 884–894. https://doi.org/10.18653/v1/S15-2149

Qiu, Q., Xie, Z., Ma, K., Chen, Z., & Tao, L. (2022). Spatially oriented convolutional neural network for spatial relation extraction from natural language texts. Transactions in GIS, 26(2), 839–866. https://doi.org/10.1111/tgis.12887

Radke, M., Das, P., Stock, K., & Jones, C. B. (2019). Detecting the Geospatialness of Prepositions from Natural Language Text (Short Paper). In S. Timpf, C. Schlieder, M. Kattenbeck, B. Ludwig, & K. Stewart (Eds.), 14th International Conference on Spatial Information Theory (COSIT 2019) (Vol. 142, p. 11:1-11:8). Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik. https://doi.org/10.4230/LIPIcs.COSIT.2019.11

Radke, M., Gupta, A., Stock, K., & Jones, C. B. (2022). Disambiguating spatial prepositions: The case of geo-spatial sense detection. Transactions in GIS, 26(6), 2621–2650. https://doi.org/10.1111/tgis.12976

Rios, N., & Bart Jr, H. (2014). Geolocate Users's Manual—Now out of date. http://www.museum.tulane.edu/geolocate/standalone/manual_ver2_0.pdf

Scott, J., Stock, K., Morgan, F., Whitehead, B., & Medyckyj-Scott, D. (2021). Automated Georeferencing of Antarctic Species. In K. Janowicz & J. A. Verstegen (Eds.), 11th International Conference on Geographic Information Science (GIScience 2021)—Part II (Vol. 208, p. 13:1-13:16). Schloss Dagstuhl – Leibniz-Zentrum für Informatik. https://doi.org/10.4230/LIPIcs.GIScience.2021.II.13

Shin, H. J., Park, J. Y., Yuk, D. B., & Lee, J. S. (2020). BERT-based Spatial Information Extraction. Proceedings of the Third International Workshop on Spatial Language Understanding, 10–17. https://doi.org/10.18653/v1/2020.splu-1.2

Stock, K., & Hall, M. (2018). The Role of Context in the Interpretation of Natural Language Location Descriptions. In P. Fogliaroni, A. Ballatore, & E. Clementini (Eds.), Proceedings of Workshops and Posters at the 13th International Conference on Spatial Information Theory (COSIT 2017) (pp. 245–254). Springer International Publishing.

Stock, K., Pasley, R. C., Gardner, Z., Brindley, P., Morley, J., & Cialone, C. (2013). Creating a Corpus of Geospatial Natural Language. 279–298. https://doi.org/10.1007/978-3-319-01790-7_16

Stock, K., & Yousaf, J. (2018). Context-aware automated interpretation of elaborate natural language descriptions of location through learning from empirical data. International Journal of Geographical Information Science, 32(6), 1087–1116. https://doi.org/10.1080/13658816.2018.1432861

Te Rūnanga o Ngāi Tahu. (2023). Homepage—Cultural Mapping Project—Te Rūnanga o Ngāi Tahu. https://www.kahurumanu.co.nz/

van Erp, M., Hensel, R., Ceolin, D., & Meij, M. van der. (2015). Georeferencing Animal Specimen Datasets. Transactions in GIS, 19(4), 563–581. https://doi.org/10.1111/tgis.12110

Vasardani, M., Winter, S., & Richter, K.-F. (2013). Locating place names from place descriptions. International Journal of Geographical Information Science, 27(12), 2509–2532.

Wang, X., Ma, C., Zheng, H., Liu, C., Xie, P., Li, L., & Si, L. (2019). DM_NLP at SemEval-2018 Task 12: A Pipeline System for Toponym Resolution. Proceedings of the 13th International Workshop on Semantic Evaluation, 917–923. https://doi.org/10.18653/v1/S19-2156

Wieczorek, J., Guo, Q., & Hijmans, R. (2004). The point-radius method for georeferencing locality descriptions and calculating associated uncertainty. International Journal of Geographical Information Science, 18(8), 745–767. https://doi.org/10.1080/13658810412331280211

Zhang, C., Zhang, X., Jiang, W., Shen, Q., & Zhang, S. (2009). Rule-based extraction of spatial relations in natural language text. CiSE 2009. International Conference on Computational Intelligence and Software Engineering, 1–4.

Zhong, Z., & Chen, D. (2021). A Frustratingly Easy Approach for Entity and Relation Extraction (arXiv:2010.12812). arXiv. http://arxiv.org/abs/2010.12812