Indigenous frameworks for data-intensive humanities: recalibrating the past through knowledge engineering and generative modelling.

Sydney Shep, Marcus Frean, Rhys Owen, Rere-No-A-Rangi Pope, Pikihuia Reihana, Valerie Chan

Journal of Data Mining and Digital Humanities

Indigenous frameworks for data-intensive humanities: recalibrating the past through knowledge engineering and generative modelling.

Sydney Shep[1], **Marcus Frean**[1], **Rhys Owen**[1], **Rere-No-A-Rangi Pope**[1], **Pikihuia Reihana**[1], **Valerie Chan**[1]

[1]Te Herenga Waka, Victoria University of Wellington, New Zealand

Corresponding author: Sydney Shep , sydney.shep@vuw.ac.nz

Abstract

Identifying, contacting and engaging missing shareholders constitutes an enormous challenge for Māori incorporations, iwi and hapū across Aotearoa New Zealand. Without accurate data or tools to harmonise existing fragmented or conflicting data sources, issues around land succession, opportunities for economic development, and maintenance of whānau relationships are all negatively impacted. This unique three-way research collaboration between Victoria University of Wellington (VUW), Parininihi ki Waitotara Incorporation (PKW), and University of Auckland funded by the National Science Challenge, Science for Technological Innovation catalyses innovation through new digital humanities-inflected data science modelling and analytics with the kaupapa of reconnecting missing Māori shareholders for a prosperous economic, cultural, and socially revitalised future. This paper provides an overview of VUW's culturally-embedded social network approach to the project, discusses the challenges of working within an indigenous worldview, shares some preliminary findings, and emphasises the importance of decolonising digital humanities.

Keywords

indigenous knowledge; semantic web; generative modelling; Bayesian record linkage; network analysis

I INTRODUCTION

Rere ki uta	*Fly inland*
Rere ki tai	*Fly coastward*
Tau mai te manu	*The bird settles*
Pitakataka ki to pae e	*And flits about its perch*

The impact of nineteenth-century Māori land confiscations is a lived experience in Aotearoa New Zealand today. Despite partial restitution and contemporary treaty settlements, identifying, contacting and engaging missing owners and shareholders of these lands constitutes an enormous challenge for Māori incorporations, iwi and hapū. Without accurate data or tools to harmonise existing fragmented or conflicting data sources, issues around land succession, opportunities for economic development, and maintenance of whānau [*kinship*] relationships are all negatively impacted. Kimihia te Matangaro - Finding the Missing is a multidisciplinary research project grounded in Indigenous frameworks that combines generative modelling and probabilistic thinking with culturally-tuned semantic web/linked open data (CIDOC-CRM) knowledge engineering to enable data interoperability and Bayesian record linkage. Victoria University of Wellington's (VUW) research journey is grounded in an understanding of the problem in the context of te ao Māori [*Māori worldview*] and te ao raraunga [*the world of Māori data*]. The interrelationship between whānau [*family*], whenua [*land*], and te reo [*language*] frames our

engagement with Parininihi ki Waitotara [2020] (PKW) and its shareholders, determines our research aims and objectives, and enables the co-design of technical solutions co-located in the social and cultural networked realities of mātauranga Māori [*Māori knowledge*]. This paper provides an overview of VUW's culturally-embedded social network approach to the project, discusses the challenges of working within an Indigenous worldview, shares some preliminary findings, and emphasises the importance of decolonising digital humanities.

II BACKGROUND TO THE RESEARCH PROBLEM

> **Me titiro whakamuri kia haere whakamua**
> *To understand where you are going, you must understand where you've been*

Despite Te Tiriti o Waitangi [*The Treaty of Waitangi*] being signed between the Crown and a number of rangatira [*chiefs*] across the country beginning on 6 February 1840, the mid 1800s in Aotearoa New Zealand was characterised by bloody skirmishes between Imperial Britain and Māori. The reason for war was simple; Britain wanted to acquire Māori land by any means possible to expand European settlement in the new South Pacific colony, whereas Māori wanted to remain on their ancestral lands, which had been inhabited for over a millennium. Fighting ensued, with millions of hectares of Māori land confiscated by the Crown to punish the rebellious natives. Although the war had largely come to an end, the period of the late 1800s to the early 1900s marked even more significant land confiscation and alienation for Māori, this time through forms of legislative and bureaucratic colonisation; the pen was indeed mightier than the sword (Fyers and Hartevelt [2018]).

Before European land ownership models were introduced, Māori land was held collectively by the iwi [*tribe*] or hapū [*clan*] and rights to occupy such lands were determined by the kinship group. Whakapapa [*genealogical*] ties to the original occupiers of said lands provided such rights. The establishment of the Native Land Court Act in 1862 set out to "encourage the extinction of native proprietary customs" in favour of an individualisation of property title similar to that of private property, in order to free up Māori land for European settlers to purchase. This process of having to establish "titles" for land that had been previously occupied for centuries resulted in widespread land loss and alienation since many Māori would often use sections of their land as down payments for food and travel costs to get to court hearings across the country. Since the certificate of title was not allowed to be issued to more than 10 people, there were many land disputes that persist still to this day, and absentee ownership is common.

Confiscation, commodification and individualisation of Māori land has created the need for management structures that ameliorate the problems of fragmented title and absentee ownership (Kingi [2008]). Since individual title contrasts with traditional Māori methods of collective enterprise, these entities also tend to emulate Māori social structures and maintain tikanga [*traditional custom*] whilst aiming to provide economic development to their shareholders. According to the NZ Institute of Economic Research [2003], Māori land administered by incorporations and trusts is estimated to be worth $NZ 1.5 billion and contributes around $NZ 700 million a year to the NZ economy.

In 1963, 22,000 hectares of Māori land originally confiscated by the Crown and absorbed into the West Coast Settlement Reserves were amalgamated into the Parininihi ki Waitotara Mega Reserve making all land owners now shareholders in a single, large portfolio. This absorption

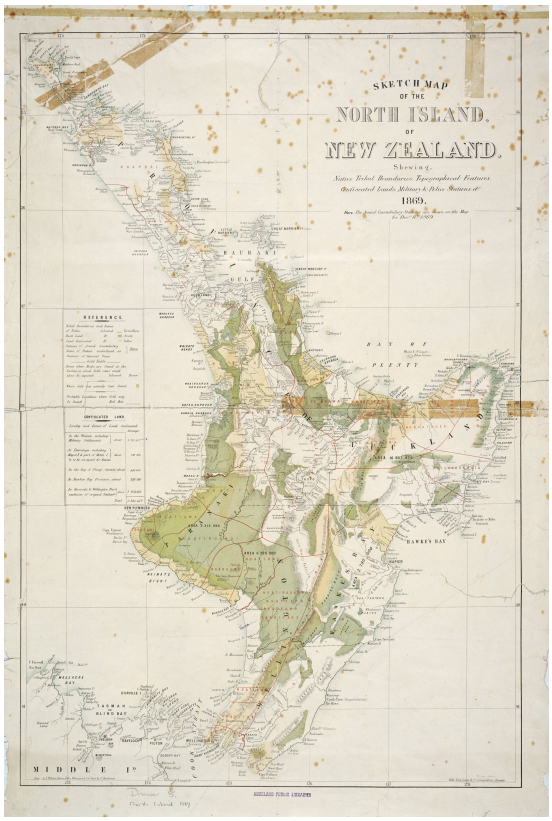


Figure 1: Te Ika a Maui | North Island, New Zealand (Palin [1869]).

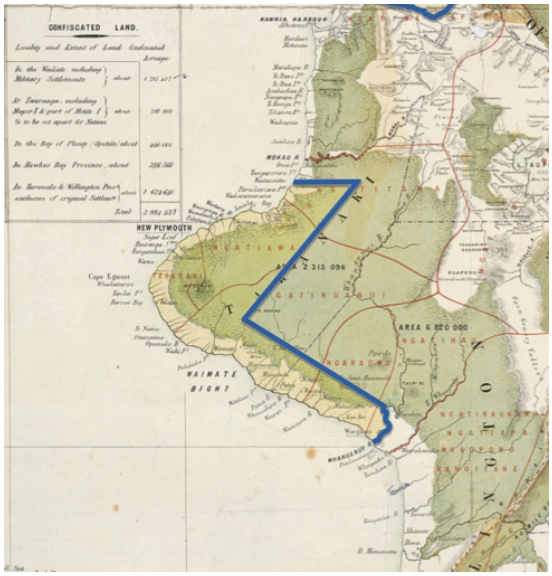


Figure 2: Taranaki confiscation line (Boast [2015]).

and, in effect, alienation of traditional, whānau-based communal land rights helped to further extinguish individual title and allowed land in the Taranaki region to be sold more easily since there was a greater pool of potential sellers, regardless of whether the original owners of those blocks were opposed to selling. Thirteen years later in 1976, Taranaki-based incorporation Parininihi ki Waitotara (PKW) was created to administer these lands and derive benefit for its shareholders, past present and future.

Succession is the legal process by which a whānau proves its historic claim to a specific land block and, in effect, inherits title to that block. There are 43 different ownership types with some blocks having over a thousand owners or held in a whānau trust. The evidential bar and the stigma of having to go to the Māori Land Court to have one's ancestral rights validated by the Crown means successions are often left in abeyance with generations missing out on benefits or even knowing their entitlements. The other effect of amalgamation was that it severed whakapapa links with a shareholder's ancestral lands. As a result, it is not uncommon for people to contact PKW wondering if they are a shareholder, and if so, where their original block of land might be. PKW's task is to assemble enough information to enable a person to connect back to their land and participate in the Crown's succession process.

The Incorporation also has a suite of strategic cultural, social, health, educational, and economic engagement initiatives to strengthen those connections once re-established. Currently, much of this work to find and connect with missing shareholders is reliant upon manual methods conducted by a shareholder officer who is an experienced historical researcher with fifteen years of front-line service processing successions for the Māori Land Court. However, the scale of the problem is enormous: only about 40% of the PKW shareholders are known and contactable

out of a register of 10,000. Dividends cannot be dispersed, and collective decision making is compromised. This kind of scenario is repeated daily across Aotearoa New Zealand.

Finding these missing community members is a complex problem requiring collection and processing of data from multiple disparate information sources and using analytics to infer connections. Our research challenge is to develop computational tools and techniques to complement and accelerate our expert's analogue work. A key step in this process is matching names through computational linguistics, a work stream undertaken by our research collaborators at the University of Auckland. However, Māori names never stand in isolation: they expose contested histories, embrace Indigenous worldviews rooted in deep time, and embed an intimate connection to whenua. As Ross [2020] notes, "names, in a society with an unwritten language prior to the arrival of Europeans, were used to retain important information for families." They were, in effect, tribal kete mātauranga [*knowledge baskets or repositories*]. Rather than focusing exclusively on narrow searches from the information available for each named individual, researchers at Victoria University of Wellington are capturing information about the community to which all the missing shareholders belong. We may call these people 'missing' shareholders, but they may not know they are lost: they also may not wish to be found. If we remain looking for individuals, then we are overlooking a whole range of opportunities to investigate how an individual is related to a larger collective, be it the whānau, the marae [*meeting house*], the hapū, the iwi, the rūnanga [*tribal authority*], or the incorporation. In research terms, then, we are shifting the unit of analysis from the single person to the whānau. Such a network approach is key to understanding the problem in the context of mātauranga Māori and te ao Māori.

III INDIGENOUS FRAMEWORKS

Indigenous identity is linked to place, articulated through language, and expressed through one's pepeha: the formulaic acknowledgement of connection to mountain, river, tribe as well as tūpuna [*ancestors*] and family. Identity is also fundamentally mutable. In acknowledging the complexity of Maori identities, Kukutai and Webber [2017] note, "they are simultaneously ever changing (because they are necessarily responsive to context, people, space, time) and sure and still (because our reo, tikanga, kawa and connectedness to our whenua, iwi and hapū will forever be the essence of what it means to be Māori)."

The kaupapa [*approach*] for our project weaves together whenua, whānau, and te reo — land, people, and language — into our kete mātauranga. Visualised by project kairuruku [*researcher*] Pikihuia Reihana, our knowledge triangle is derived from two whakaaro [*concepts*]:

Ngati Hine Puke Puke Rau. He Puke He Rangatira!
The myriad of hills of Ngati Hine. It is said that on every hill there lives a Rangatira, chief over all that he sees.

The design is based on the Niho Taniwha [*teeth of the taniwha*] and represents the historian, the keeper of knowledge. It also represents whakapapa from the Atua [*Gods*] to the Rangatira and their many uri [*descendants*]. The darker spaces represent existing knowledge and the lighter

spaces, new knowledge. The white spaces represent the unknowing. This is our maunga [*mountain*]. This is our bend in the landscape. Our knowledge triangle is embedded in an Indigenous framework that shapes our research practice and informs our engagement with PKW and its community. Te ao Māori is based on kaupapa [*a values system*] and likewise our kaupapa influences our tikanga [*our methodology, our practice*]. These are concepts which are fundamental to our mātauranga Māori framework.

Western science research paradigms begin by identifying a researchable problem out of which a research question is formed. Indigenous research begins with community engagement to build relationships, acknowledge authority and expertise, and create an environment of trusted communication, feedback and validation. Only once these core principles have been established and enacted is the process of research discovery and co-design initiated (Shedlock and Vos [2018]). Such a kaupapa Māori approach "gives full recognition to Māori cultural values and systems; challenges Western (dominant) constructions of research; determines the assumptions, values, key ideas, and priorities of the research, ensures that Māori maintain conceptual, methodological and interpretive control over the research, and is guided by Māori philosophical beliefs, traditions and values" (Kennedy [2010]).

There is no one pathway or method to embed mātauranga Māori in a research programme. All things are born Indigenous (Harmsworth and Awatere [2013]; it is non-Māori that require rationalisation of te ao Māori. A large corpus of literature exists that defines mātauranga Māori. Academics and researchers agree that the Indigenous paradigm for Māori is its own system and if Māori are to flourish as Māori living and developing as Māori, then mātauranga Māori must be accordingly prioritised (Byrom [2017]; Hikuroa [2017]; Mercier [2018]). Māori seek to understand the collective and its interdependencies not just parts in isolation (Harmsworth and Awatere [2013]; Winiata [2001]). Additionally, mātauranga Māori must not be dependent on its value to western science but instead its value to Māori. Mātauranga Māori is greater than science alone, it is a cultural system of knowledge about everything important in the lives of Māori (Broughton and McBreen [2015]). Durie [2004] agrees, explaining that researchers must give mutual respect to both Indigenous knowledge and science, that Indigenous knowledge cannot be verified by scientific criteria nor can science be adequately assessed according to the beliefs of Indigenous knowledge. Mercier [2018] posits that mātauranga revitalisation must be Māori-led and include recognition of tino rangatiratanga [*self-determination*]. The core functions of mātauranga Māori are at the forefront and interface of our research, which is values-based and a respecter of Indigenous knowledge and science.

3.1 Data Sources: challenges and affordances

Working within an Indigenous framework also means acknowledging the complex and often controversial political and ethical issues around te ao raraunga and Māori data sovereignty and stewardship: how has the data been collected, where is it stored, to whom does the data actually belong, who has the right to use it. As noted Māori researcher Linda Tuhiwai Smith observes, for too long Māori have been made the object of research with 'collaboration' consisting of "helicopter" researchers flying into a community, grabbing data, and using it for their own research programme, with little or no benefit to the community and, historically, generating much harm (Smith [2012]). Organisations such as Te Mana Raraunga [2016] have initiated calls to action to reclaim Māori data, store it in secure, local, Māori-governed clouds, and manage access for the collective benefit of Māori and to enable the fulfilment of contemporary aspirations.

Te Tiriti o Waitangi was intended to ensure Māori maintained sovereignty over their taonga (i.e. land, resources) and maintained tino rangatiratanga over their communities (i.e. whānau, hapū, iwi). The year 2020 marks 180 years since the signing of Te Tiriti but the issues of sovereignty continue to be debated. The matter of data sovereignty is a relatively new concept that has become a significant issue globally (Hudson et al. [2017]). Indigenous data sovereignty has also emerged as a significant issue (Kukutai and Taylor [2016]). Just as data is subject to management aligning to the laws, practices and customs of the nation in which it is located, so too should Indigenous data be subject to the practices and customs of the collective (Lovett et al. [2019]). Thus, Māori [Indigenous] data should align with sovereignty rights articulated in Te Tiriti o Waitangi further supported by the United Nations Declaration on the Rights of Indigenous Peoples. These are also the underpinnings of the Te Mana Raraunga [*Māori Data Sovereignty Network*] (Gifford and Mikaere [2019]; Te Mana Raraunga [2016]).

Much Māori data is held by the Crown and is an artefact of colonisation. It is unavailable (i.e. individual unit records), the product of deeply flawed data collection practices, kept behind government data walls (Integrated Data Infrastructure: Stats NZ [2018]), or too costly to access in the case of Births, Deaths, and Marriages. Many datasets are embedded in proprietary software or use completely different information systems which do not talk to each other. 'Open data' is not a term which fits comfortably within te ao Māori given much tribal information, including personal names - whether in the public domain or not - is locked in whakapapa, remains tapu [*sacred*], and is considered taonga [*treasure*]. This runs directly counter to the three pillars of digital humanities: open access publishing, open access/open source software development, and open data. Indigenous communities can work innovatively within the first two pillars, but is it "the rhetoric and practice of the open access data movement [that] obscures both Native agency in determining the use of community materials as well as the role of technical determinism in proliferating the violence of colonial archives on Native communities" (Guiliano and Heitman [2017]). As Gaertner [2017] observes, "In the realm of technology, the colonial drive to know, and the demand to have access to any and all forms of knowledge with the touch of a button, is repackaged as 'open access'. The idea that 'information wants to be free' is dependent on colonial structures of knowing that privilege the dissemination of knowledge over the rights, interests, and well-being of the people it is drawn from." Consequently, our project is constantly navigating the open imperative of our discipline whilst respecting and honouring the cultural protocols of our researchers and communities. For these reasons, we anonymise our data and do not make it publicly available. We also support our Maōri researchers to share their whakapapa data because they have the requisite cultural permissions and mana [*spiritual authority*] to do so.

In response to the reclamation of Māori data and evidence of tino rangatiratanga in action, the Iwi Leaders Group for Data (Data ILG) was established in 2016 to empower Iwi Māori to better harness the potential of data, including collection, protection, preservation, storage, and re-use. The kaupapa matua [*purpose*] is aspirational requiring that the Data ILG obtain full and free access and control over data about and for Iwi Māori with the goal of advancing Iwi Māori aspirations and data agenda. Starting with strategic relationships with the Ministry of Business Innovation and Employment (MBIE) and Statistics New Zealand (StatsNZ), two national Māori Data Futures Hui have also led the charge: the first in 2018 hosted at Te Herenga Waka Victoria University of Wellington (Science for Technological Innovation NSC et al. [2018]) laid the groundwork for the conversation; the second in 2019 at Te Aurere Marae, Taipa (Science for Technological Innovation NSC et al. [2019]) focussed on intellectual property, exploring

how raraunga and Mātauranga Māori might be protected, and how Māori might start capturing the benefits of data. Moreover, under the mantle of Te Ara Takatu, StatsNZ have provided customised data services for iwi and iwi-related groups. An agreed programme of work that aims to mitigate some of the effects of the 2018 Census on iwi data is one example of how Iwi Māori are advancing the data agenda (Kukutai and Cormack [2018]). In the spirit of implementing 'open science' within Iwi Māori and increasing availability and access to scientific research information and data, this work programme has adopted a cultural licence for Māori data sovereignty and a social licence for trusted data use. It is also working in partnership with Māori interest organisations, iwi, and Māori to find real and relevant solutions to Māori data needs for Aotearoa New Zealand. This includes supporting other government agencies to collect and provide good quality iwi affiliation data, supporting iwi to build their data capability, and co-designing specific data initiatives such as our current research project.

One of Kimihia te Matangaro's key Crown datasets, Māori Land Online [2020], is the main public-facing portal for documenting and managing Māori land succession information; it also represents the richest dataset of current Māori landowners. However, it is a complex system that has serious legacy issues stretching over 150 years which impact on opportunities to harvest, analyse, and visualise the rich whānau, whenua, and te reo cultural data held by Te Kooti Whenua Māori [*Māori Land Court*] and Toitū te Whenua [*Land Information New Zealand*]. The current digital practice of updating by overwriting historical records makes tracking data provenance tricky if not impossible. Similarly, another key Crown data source, Births, Deaths, and Marriages Historical (BDM [2020]), has variable access dates for each dataset based on New Zealand's privacy legislation, historically separate record-keeping systems for Māori, and serious anomalies within the data. Yet it is still regarded by the Crown as the record of authority for all identity documents and remains the basis of Crown decision-making.

By contrast, tribal genealogies have a different system of access and validation, relying on oral rather than written tribal knowledge held by kaumātua and kuia [*tribal elders*], verified by the collective. For younger generations, this information is often shared through private social media channels or public-facing tribal genealogy websites. Moreover, the land speaks volumes about identity and these kōrero [*stories*] are increasingly part of iwi-led cultural mapping projects. Accessing these data sources relies on tribal contacts and underpins our project's commitment to employing Māori researchers who can use their own whakapapa as ground truth and who can bring the project's mahi [*work*] into their own iwi or hapū contexts.

Since we are connecting two data sources by some sort of causal arrow, the authoritativeness of these records are not part of the calculation. We do not accept the Crown to be the authoritative voice for what constitutes ground truth. Māori identity, in all of its fluidity should not be fixed to one piece of data. Therefore different Crown datasets used in the research all hold equal weight regardless of their flaws or merits. Our positioning of ground truth is similar to that of philosopher Mikhail Bakhtin's suggestion that "truth is not born nor is it to be found inside the head of an individual person, it is born between people collectively searching for truth, in the process of their dialogic interaction" (Bakhtin and Emerson [1984]). The question of "ground truth" ordinarily used as a measure of validity and reliability in a computational research setting is also complicated by our third data source, PKW's in-house, confidential share register. Like the Crown's authority records or tribal whakapapa, the register is an approximation of "truth" at any given moment in time based on available information and its status as verified evidence. Its truth-value is argued by PKW's in-house historian who functions as a lawyer using the available

evidence to achieve a determination. Like snapshots of a person at important stages in their life, we know the person exists, but depending on the time, place and people, that person may appear differently in each photo. We are, in effect, serving up a photo album. Consequently, the project has adopted an understanding of ground truth as always already negotiated, manufactured, constructed. As such, we employ the term "relative ground truth" as a way of combining big data with thick data (Siodmok [2020]). Given the aim of the research is to provide the infrastructure, tools, and prototype applications to help PKW weave their kete mātauranga about their shareholders, our fusion of public and confidential data sources, the creation of a secure, trusted, and local data repository (Mātauranga cloud), and meaningful engagement with the community are critical.

IV WEAVING THE KETE

Our approach to the socio-technical wero [*challenge*] of finding 'missing' shareholders focuses on networked relationships to people and, critically, to land. We suggest that single individuals can be found because, rather than being lost, they are part of larger networks as yet unidentified. By mapping entire networks, we can plot the links between people and groups and find out who is most likely to be related to whom, and thus to know someone either directly or indirectly. As relationships change over time and people move around, this network becomes a dynamic, complex system that may throw up surprising links and hitherto unknown inter-group affiliations. To achieve this kaupapa, our team has focussed on two methods: knowledge engineering and generative modelling.

4.1 Knowledge Engineering

Underpinning our whānau or network approach, is a culturally-tuned, linked data architecture or "macroscope" (Graham et al. [2016]) developed as a interoperability framework to knit together and explore disparate datasets, enable data fusion, build analytics tools, and create interactive visualisations for and with the PKW community. The linked data ontology CIDOC-CRM [2015] was selected as being robust enough to represent and comprehend the complexity of our historical data sources and responsive enough, at least initially, to our Indigenous cultural context. Successfully used in the cultural heritage sector particularly for big linked data sets also ensured access to an international and experienced community of practice.

Building the schema became a way of cleaning thick but messy data to render it in a principled computational form for our data scientists and to flag questions for our PKW subject expert. The schema was reviewed, tested, and fine-tuned iteratively against specific real world examples taken from our researchers' own whānau histories while we deepened our understanding of the idiosyncracies of the Māori Land Court systems and data as they changed over time.

The iterative development of this conceptual reference model was only made possible through building a relationship with PKW and developing a mutually respectful cultural environment that allowed for knowledge exchange across subject boundaries. Our understanding of the complex nature of Māori land succession data was served up to staff who had years of experience with Māori Land Court legal processes which helped fine tune how we were conceptualising the complex data environment and thus guide us to better represent the data in a schema. As Siodmok [2020] explains, "ultimately it is the prototype that is the acid test for new ideas." In

this case, the prototype sitting at the intersection of big and thick data was the most recently updated version of the ontology, providing for PKW staff a window into the complex and messy data. Such an interface between two worlds was crucial in utilising feedback to gain a more intimate understanding of the complex historic legal processes that created this data, and model those processes accurately. It also reinforced the dynamic and reciprocal nature of ontology building.

4.1.1 Technical workbench

In technical terms, the project uses python programs to harvest to data in json format from two sources: Māori Land Online and Births Deaths Marriages Historical. Harvesting the entire MLO corpus rather than just the Aotea judicial district in which the Taranaki region is placed is a necessary step towards recreating the entirety of Maōri land ownership in Aotearoa New Zealand. It is common for an individual title in Māori land to be passed down from whānau who descend from different lands around the country. It is therefore more often than not that an owner will have their interests geographically spread out and as such, mapping this data will provide a more accurate picture of the scale of their interests.

This data contains considerably detailed information about current owners, each tightly clustered around a Māori Land Court minute book reference, land block identifier, and a consistent number of shares. These "m-groups" can be thought of as a possible family grouping of names of siblings present at the time of the succession, although with a high likelihood that unrelated people's names may occasionally be included, along with some unknown probability that some valid members may be missed out. Similarly, pre-1920 birth records ("b-groups") were harvested from Births, Deaths, Marriages Historical and clustered into sibling groups based on birth entries which shared the same surname and exact same parents' names. It is these sibling groups that are the current unit of analysis for current and future Bayesian record linkage work.

From MLO, we obtain land information using a wfs query, and owner information with scripted form submissions. From BDM we obtain summary birth information using form submission then html parsing and xpath querying using html5lib. The json from each source is then transformed into json-ld. The RDF predicates and entities are all crm: with literals from xsd: and geo: We are using Jena for the triplestore, and Fuseki as a SPARQL server. Fuseki allows us to expose the data by parsing SPARQL queries to the triplestore, of which MLO and BDM currently contains more than 32 million triples each.

A first prototype interface to navigate MLO was built with a react app using Fuseki as a backend. We are now reworking it using sparql CONSTRUCT queries to generate a relevant graph given a set of resource URLS, visualised with d3 force-directed graphs and leaflet using d3 generated coordinates for entities without geometry. We have also replaced react with jquery and node module(s) packaging for browserify.

The scale and complexity of our existing data and in anticipation of future Bayesian record linkage beyond our two initial datasets means we have also moved our project into two high performance computing environments: VUW's in-house cluster Rāpoi [2020] and NeSI [2020], the New Zealand National eScience Infrastructure. Faster compute times and parallel processing have accelerated our iterative approach.

4.1.2 *Comprehending a Possible*

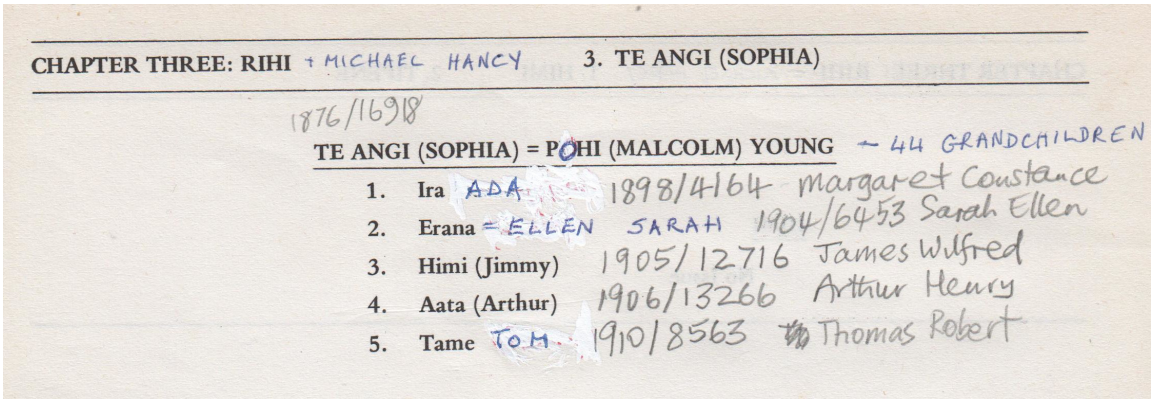
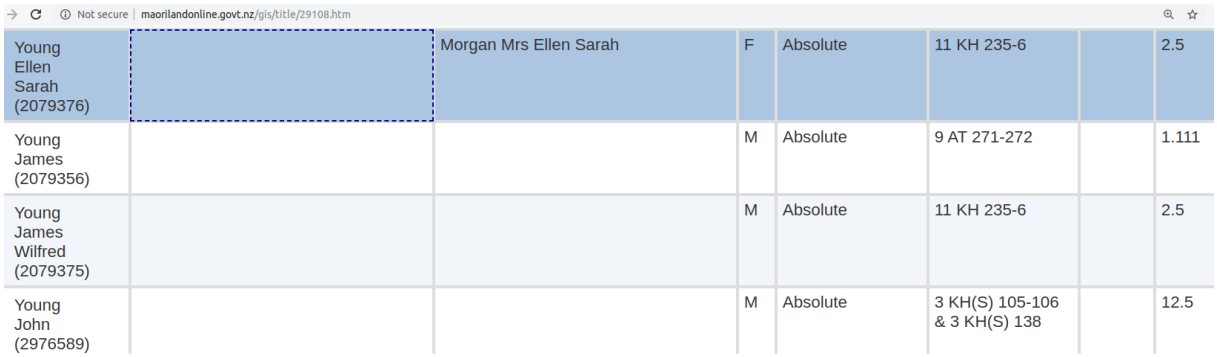


Figure 3: Cassidy-Robson and Harris [1980], p3-6, annotations Alicia Owen and Rhys Owen

In exploring the Harris whānau of one of our Māori researchers through the triangulation of whakapapa, Crown records, and human expert knowledge, a number of issues emerged which exposes the complexity and uncertainty of our research problem (Figure 3). Erana, for example, is also known as Ellen, but appears in the birth records as Sarah, a name not used by the family, and whose correlation in the transliterated te reo Māori corpus is Hera, again unused. To complicate matters, Sarah Ellen appears twice in the official birth records with different registration numbers. Similarly, her brother Himi is also known as Jimmy, but according to the Crown, is legally James. While the linguistic distance from Himi to James can be quantified, Himi is aurally closer to Jimmy, whereas the more common Hēmi is closer to James, thus reflecting the mutability of oral and written exchanges between te reo and English. Data mining from BDM and subsequent json-ld translation produced a whānau sibling group which, when linked to the Māori Land Court data, enabled a tighter cluster to be identified and exposed additional anomalies. The sibling group appears with its minute book reference, land block and number of shares (Figure 4).



Young Ellen Sarah (2079376)		Morgan Mrs Ellen Sarah	F	Absolute	11 KH 235-6		2.5
Young James (2079356)			M	Absolute	9 AT 271-272		1.111
Young James Wilfred (2079375)			M	Absolute	11 KH 235-6		2.5
Young John (2976589)			M	Absolute	3 KH(S) 105-106 & 3 KH(S) 138		12.5

Figure 4: A search result from Māori Land Online [2020].

Here, Sarah Ellen Young is now Sarah Ellen Morgan; her married name is now the primary identifier. Digging deeper in the data throws up two different MLO identifiers. Given the variability of individual names and anomalies in record-keeping systems over time, our social network or whānau approach as expressed in the Harris whānau schema (Figure 9) is proving to be a robust matrix for identifying possible clusters which our data scientists can then transform into probabilities across individual clusters as well as across the project's entire combined dataset.

The first stage of our knowledge engineering was to comprehend each of our two datasets individually and revise our CIDOC-CRM schema prototype over the course of several Taranaki-based consultations with PKW's resident experts.

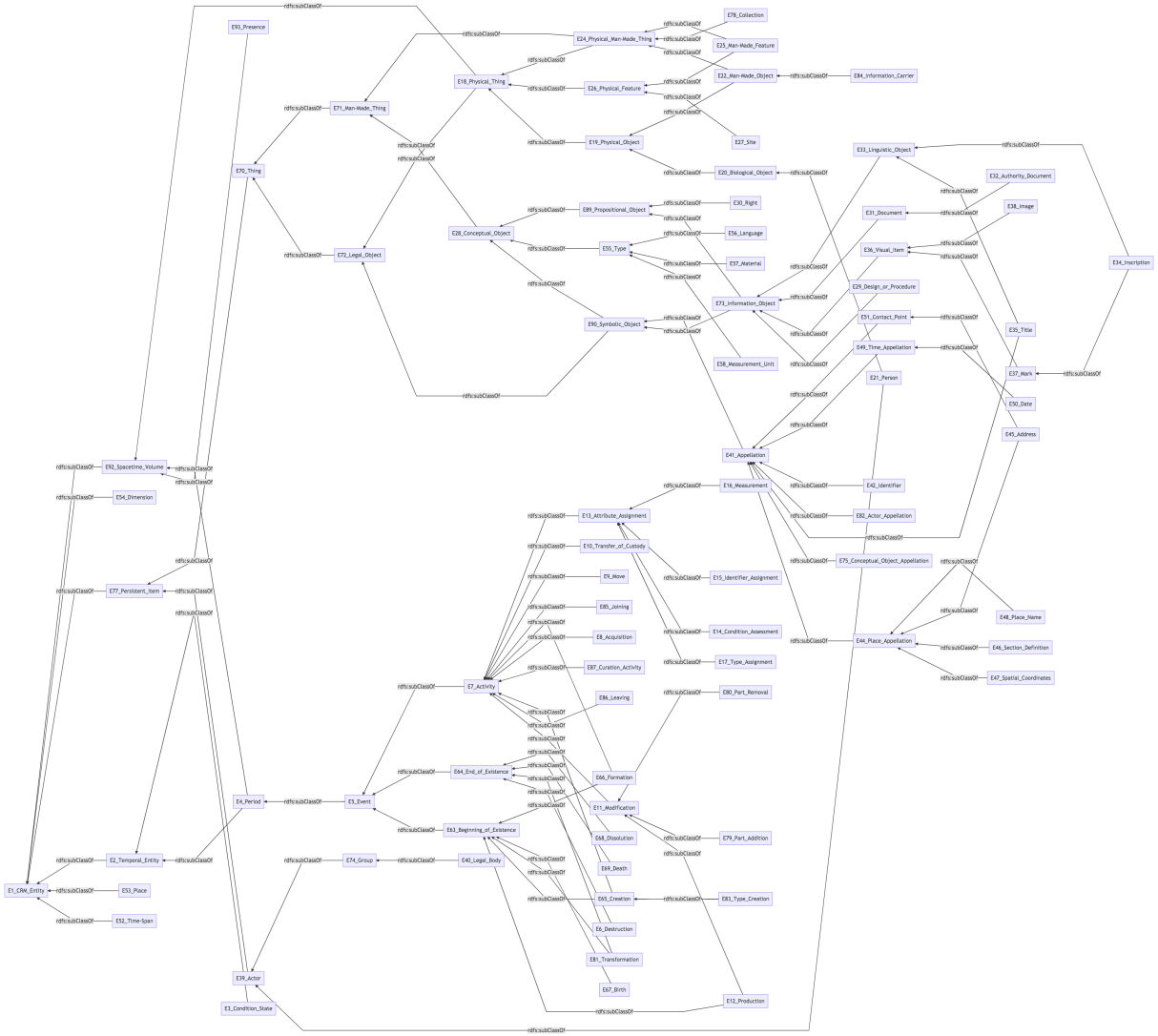


Figure 5: Correctly expressing the semantic meaning of entities and their explicit relationships was achieved by drawing on the hierarchical nature of class inheritance in the CIDOC-CRM model.

In the context of our whānau or network approach, the resonant structural unit with which to join and comprehend the two disparate data sources was the group: in particular, the sibling group. The schema as a critical artifact shows our comprehension of the data in terms of the RDF notions of entity classes, relationship types, resources as first class entities and literals as annotations for entities.

An essentially structural way of finding groups of persons who are possibly siblings is by modelling how shareholders become such through the process of succession and the division of one's shares (Figure 6). The MLO minute book reference (expressed in CIDOC-CRM as an E7 Activity) to minutes of a court hearing represents when a person (or persons) became a shareholder in a particular land block (expressed in CIDOC-CRM as an E27 Site) as a result of a succession.

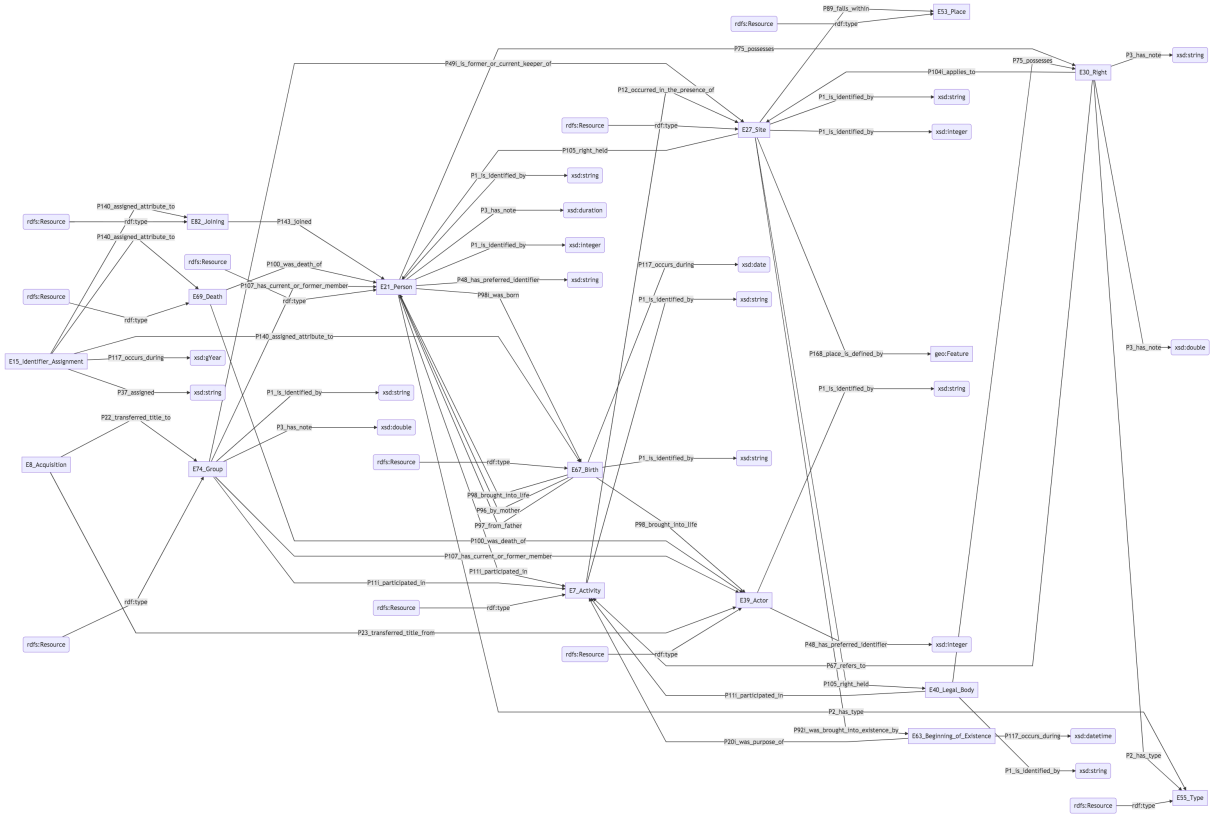


Figure 6: MLO and BDM data comprehended as CIDOC-CRM.

Since the transfer of said shares by one person from another can only occur as a result of whakapapa links, it is almost certainly the case that any two (or more) people who were recorded as participants in the same court hearing, for the same land block, will be related to one another. The previous owner's shares (expressed in CIDOC-CRM as an E30 Right) would often be distributed evenly between their children. Additionally, if applicable, a portion of the total shares amount would be evenly distributed between their surviving siblings and another portion between their nieces and nephews and so on, according to the degree of relation to the shareholder in question. This pattern produces a third characteristic in which we are able to define groups of siblings. Furthermore, siblings who are not just direct descendants to the previous owner of the shares, but are fundamentally siblings, can be captured. As indicated by RDF in the schema (Figure 7) both James and Ellen possess the same value of shares in the same land block Wharau D, which were transferred to them by one of their parents. This trifecta of matching characteristics enables us to knit together records in the MLO dataset. The strength of this approach is in its blindness to names as a way to find or group people. Instead, we rely solely on the structuring principle of the sibling group.

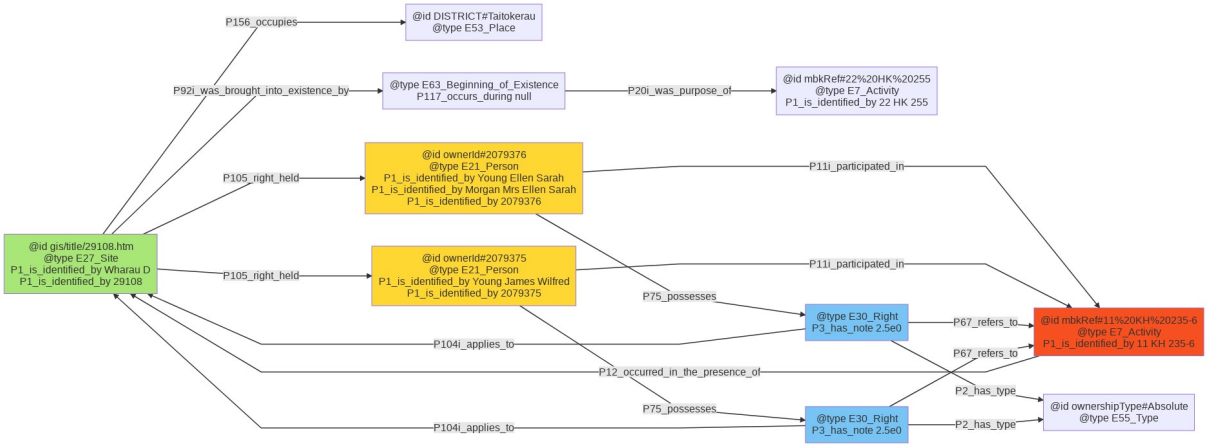


Figure 7: CIDOC-CRM representation of siblings Ellen and James Young in relation to land block Wharau D.

Turning to BDM Historical, the same land block/title owner appears as follows (Figure 8). Within Birth and Death datasets there is a very strong presumption by the Crown that a person has at most one register entry; this example disproves the claim, and is not an isolated instance.

bdmhistoricalrecords.dia.govt.nz/Search/Search?Path=querySubmit.m%3fReportName%3dBirthSearch%26recordsPP%3d30#SearchResults

Registration Number	Family Name	Given Name(s)	Mother's Given Name(s)	Father's Given Name(s)	Still Birth
1895/5704	Young	Sarah Ellen	Sophia	Malcolm	-
1904/6453	Young	Sarah Ellen	Sophia	Malcolm Lake	-

Figure 8: A search result from BDM [2020].

Using our methodology to structurally group siblings, the RDF serves up a multitude of atomic graphs (Figure 9) that provide a relational context and help comprehend anomalies in the data. We used the father's and mother's names together to group persons, so the five persons highlighted in the schema are clustered together as a possible group of siblings. This, again, is an essentially structural rather than content-based way of grouping possible siblings together. Names are used but not the names of the possible siblings, only the names of the mother and father, thereby situating the data in the culturally-meaningful whānau relationship. These groups are then indexed by their one and only surname.

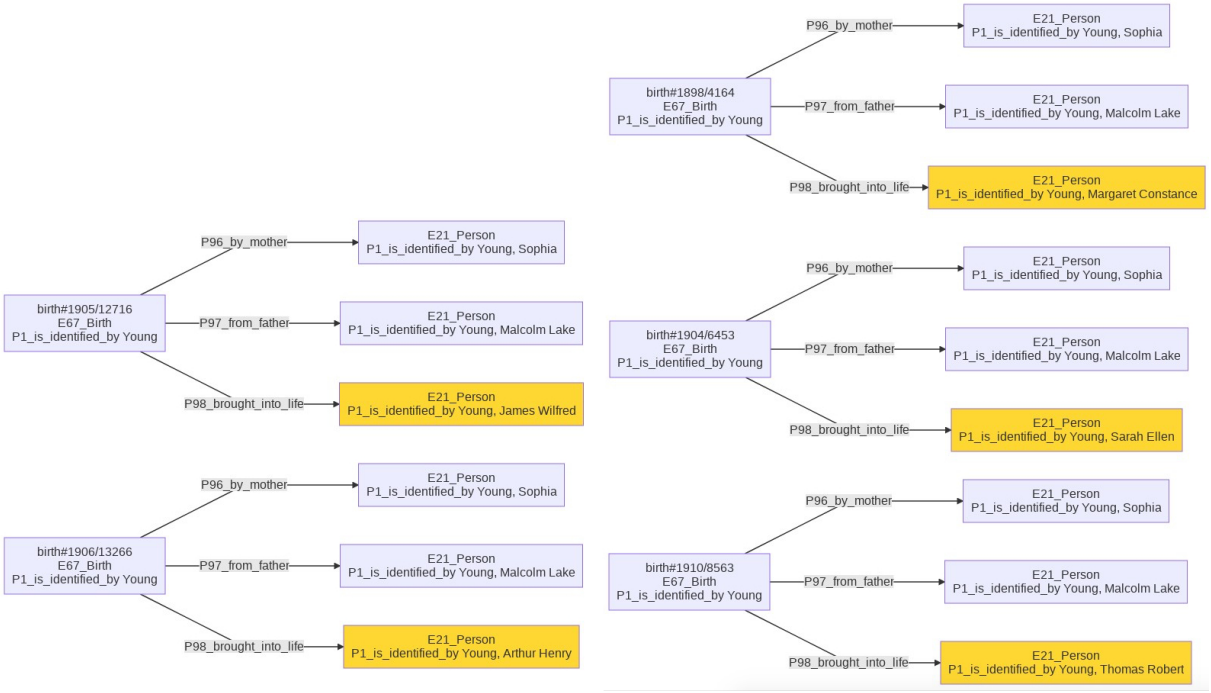


Figure 9: one of our possible sibling groups, (siblings highlighted) assembled by grouping official Birth records, expressed in CIDOC-CRM.

We similarly indexed the groups from Māori Land Online with 0, 1 or many surnames. Is it theoretically and computationally possible to assign a probability of linkage of a BDM possible sibling group from Māori Land Online with a possible sibling group from Historical Births? How likely is it that two or more siblings are included in both groups? The computational effort involved in the project's subsequent probabilistic work is expected to be very large, due to the sheer number of possible sibling groups we have identified.

Datum	Frequency
BDM surnames	22,150
BDM groups	427,084
BDM names	1,654,428
BDM surname : group ratio	19.2
BDM group : name ratio	3.8
MLO unique surnames	8,198
MLO unique groups	356,255
MLO surname-group pairs	414,170
MLO names	1,693,107
MLO unique surname : surname-group pair ratio	50.5
MLO group : name ratio	4.7
Total BDM-MLO group pairings	7,948,780
Total BDM-MLO name pairings	141,488,298

Table 1: The scale of our computational task showing MLO and BDM datum and frequency: 7.9M group combinations; 141M name combinations.

By design, then, our knowledge engineering has been in two major stages. Stage one has been a computational *structural* comprehension of our two datasets. The key idea that surfaced is that of the sibling group. Stage two delivered the *content* results of this knowledge engineering to our data scientists to test their generative modelling on real world data and to form the basis of their Bayesian record linkage. Our CIDOC-CRM ontology has exposed and mapped the key elements and relationships for PKW's shareholder expert to disambiguate hitherto fragmented information sources and enable sound decision-making. It has also opened up the possibility for Bayesian analysis. Additionally, it has provided an interface between two different knowledge systems: Indigenous and western science. In the process, however, we encountered various challenges with CIDOC-CRM for expressing Māori concepts such as rights, interest in Māori land, identity, and stories. Consequently, the next step for our knowledge engineering is to decolonise, indigenise, and localise the ontology, addressing conceptual and cultural gaps as we gain more understanding about how te ture Māori [*Māori law*] intersects with whakapapa. We anticipate this localised version informed, in part, by Ngā Upoko Tukutuku [2020] [*Māori Subject Headings*] will inspire other Indigenous communities to develop new, equally culturally-tuned informatics standards.

4.2 Generative Modelling

When two data sources can identify groups of individuals in their own terms, a bigger picture can be sought via linkages between records across the two sources. Record linkage is often focussed on de-duplication. In our case the goal is not to merge the databases, or to get rid of duplicate entities, but to identify likely connections between groups. An obvious way to do so is to associate individuals with similar names directly, as is done in record linkage. In Bayesian record linkage one seeks to avoid making all-or-nothing assertions about such links - instead focusing on maintaining the associated probabilities (Steorts et al. [2016], Sadinle et al. [2014], Enamorado et al. [2018]), mostly based on a canonical probabilistic model of record linkage (Fellegi and Sunter [1969]) in which links are either matches, non-matches, or in need of manual review.

However individual names are not the only, or even the main, source of confidence in an association. Take for example the two groups (a) `Marcus, Jessica, Nicola, Ben, Rebecca` and (b) `Marcus, Jessica, Nicola, Ben, Roberta`. The probability of linkage between the last two names is minimal if taken in isolation, but (depending on the process provisioning the two lists of course) the surrounding context lends weight to the theory that they are in fact linked - especially if that context is itself unusual (*cf* `John, William...`). Accordingly we seek a group-to-group linkage, as opposed to individual-to-individual.

A generative model gives a probabilistic explanation for the patterns in complex data, in terms of a much simpler but concrete mathematical construction. For example, a simple model of face images is to (i) give a base probability that the face originates from a female (say 0.5), along with (ii) a consistent and plausible probability for any specific face image *given gender*. Working backwards, from these ingredients Bayes theorem provides the way to infer the chance that a particular face is female (say). In this example the effect is merely the classification or clustering of images, but in our case the inference is more complex. In order to infer the likely origin of *groups* from other *groups*, the corresponding ingredients are (i) the probability of any given b-group and (ii) the likelihood of a particular m-group, *given that* b*-group*. The resulting inference

is the degree of belief we should ascribe to the statement "the group of people identified in *this* b-group later gave rise to the names we see in *this* m-group".

As previously noted, we have two very different data sources: Māori Land Court records ($\mathcal{M}$) detailing current groups of owners of land blocks, and Births (Historical) ($\mathcal{B}$) detailing parent-child connections. Each can be used to derive putative groups of siblings. Given one group M derived from the latter, we would like to find which of the groups B from the former are plausible "origins". That is, we want to say which of the groups in one source are identifiable in the other, just from the data sources, without knowing ground truth.

We denote names in MLO as follows. $\mathbf{m} = [m_1, m_2, \ldots m_M]$ is a list of names (M in number, although each may have 2 parts: given and family names), selected by a filtering process utilising the ontology discussed earlier. The process is crafted to generate groups of *possible siblings* in MLO data. Other than that loose assertion, we do not want subsequent processing to depend strongly on the details of the filtering process giving rise to $\mathbf{m}$.

Names in BDM are similar: $\mathbf{b} = [b_1, b_2, \ldots b_B]$ is a group (of size B) of identities in BDM Historical. By harvesting and filtering appropriately, we can be confident that those in a given $\mathbf{b}$ are direct siblings (although not necessarily being *all* the siblings relating to a family). Denote by $\mathcal{B}_\mathbf{m}$ the set of *all* the sibling groups $\mathbf{b}$ that we consider plausible as explanations for some group of names $\mathbf{m}$ (for example this might be every $\mathbf{b}$ containing any of the surnames present in $\mathbf{m}$). As an aside, we have gender for MLO entries, but not for BDM Historical.

Consider the question of the origin of a particular set of names $\mathbf{m}$ derived from MLO. Which $\mathbf{b}$ sets are most likely to contain the true identities of people in $\mathbf{m}$? Stated this way, the question of identity is thereby an inference problem over *groups* as opposed to individuals, foregrounding our culturally-meaningful whānau network approach. Obviously the identities of individuals will eventually play a role. One of the interesting questions is to what extent the precise alignment (ie. the matching up between individuals in a b-group and an m-group) helps in inferring the best $\mathbf{b} \in \mathcal{B}_\mathbf{m}$.

Given a particular $\mathbf{m}$, and a set $\mathcal{B}_\mathbf{m}$, we would like to assess the relative plausibility of each $\mathbf{b} \in \mathcal{B}_\mathbf{m}$ as being the *group of people* behind the names in $\mathbf{m}$. This is the posterior probability $P(\mathbf{b} \mid \mathbf{m})$, given by Bayes theorem, and one could argue that the prior over $\mathbf{b}$'s in absence of any other information is uniform, which leaves $P(\mathbf{b} \mid M, \mathbf{m}) \propto P(\mathbf{m} \mid M, \mathbf{b})$. This makes it clear that to evaluate beliefs about $\mathbf{b}$ given $\mathbf{m}$, we should look to the "forward" probability (likelihood) of $\mathbf{m}$ given $\mathbf{b}$. So what is the probability of some set of names $\mathbf{m}$ corresponding to "unidentified" people, if we were to assume they originate from a specific set of (named, identified) people $\mathbf{b}$? It might help to begin by thinking of the very simplest case, in which each "group" consists of just a single name. To start with, we form single strings that are just the concatenations of all names in $\mathbf{b}$ and $\mathbf{m}$, thus setting aside all questions of the "matching up" of individual elements, for now, and instead treating the entire group as if it were a single name. We still require a form for $P(\mathbf{m}|\mathbf{b})$.

Between $\mathbf{b}$ and $\mathbf{m}$ a lot can happen. First note the contexts were very different (one dominated by compliance with the crown's definition of legal identity, the other with connection to whenua). Then there are shortenings, additions (some predictable, others entirely new), plus

flawed memories, alternative spellings and plain typos, and surname changes through marriage, to name just some of the effects.

Under a simple predictive model of text (the "Ngram" or n-th order Markov model (Murphy [2012]) the overall probability of a string is the product of the predictive probability of each successive character given its n predecessors (pre). In log space,

$$\log P(m \mid b) = \sum_i \log P(c_i \mid c_{\texttt{pre}_i}, b) \tag{1}$$

where c_i are the characters in m. In our case, we want to model the fact that m may differ from b through any of the above processes of intervention and change. We cannot know the details, so adopt an simple approach in which each character either follows either the statistics of the b name, or comes from the M corpus as a whole. This suggests a predictive distribution that is a *mixture* of the two Ngram Markov models:

$$\log P(c_i \mid c_{\texttt{pre}_i}, b) = \log \left[\beta \Pr_{\mathbf{b}}(c_i \mid c_{\texttt{pre}_i}) \; + \; (1-\beta) \Pr_{\mathcal{M}}(c_i \mid c_{\texttt{pre}_i}) \right] \tag{2}$$

Here β is a coefficient determining how much the $\mathbf{b}$ statistics hold sway relative to the background distribution. $\Pr_{\mathbf{b}}$ denotes the predictive distribution over characters based upon the $\mathbf{b}$ string itself, while $\Pr_{\mathcal{M}}$ is the "background" distribution built from the entire corpus of names in $\mathcal{M}$. Alternatively, mixtures of more complex / realistic distributions (such as profile Hidden Markov models) could be used in place of Ngrams. A fully Bayesian treatment would place a prior on β, or we could adopt plausible values and check for robustness. It is important to note that the family name of females is altered by marriage and so that portion of the name should be modelled appropriately. Without an assertion of 1-to-1 matchups between elements of $\mathbf{b}$ and $\mathbf{m}$ though (ie, an alignment), gender is unknown for our $\mathcal{B}$ data, so this is not an option.

We can think of Equation 1 as an automaton that is fed the string $\hat{m}$ as a stream of characters and outputs a float which is the log probability of that string. For each $\mathbf{b}$, we build the associated automaton by computing and storing a dictionary for $\Pr_{\mathbf{b}}$. Then, for each $\mathbf{m}$ in the set of interest, push m through each automaton, giving score S (the log probability of that string under the $\mathbf{b}$-based mixture model). For each $\mathbf{m}$-set we now have the most plausible $\mathbf{b}$-sets and their scores (log probabilities), exponentiating those and normalising yields the posterior probability $P(b \mid m)$.

Figure 10 *(left)* shows an illustrative example (using real but anonymised data) of $P(\mathbf{b}|\mathbf{m})$ using a surname for which there are about 20 different possible "families" derived from BDM data and 60 possible "family" groupings in the MLO data . Each row corresponds to one group of names $\mathbf{b} \in \mathcal{B}_\mathbf{m}$, while the columns correspond to groups of names from the other data source, $\mathcal{M}$. Each grid site is displayed with a colour indicating its posterior probability, which is the probability of that row as origin (out of those on offer) for the m of that column. Since those probabilities are normalised, the total colour of each column is the same (total probability is 1). A column with a single dark blue square thus means that the row in question is deemed to be likely to be the source of the names in that column's m, compared to the other options on offer. Note the intermediate colours (mid blues) however, indicating that some m-groups (columns) have more than one plausible b-group (rows).

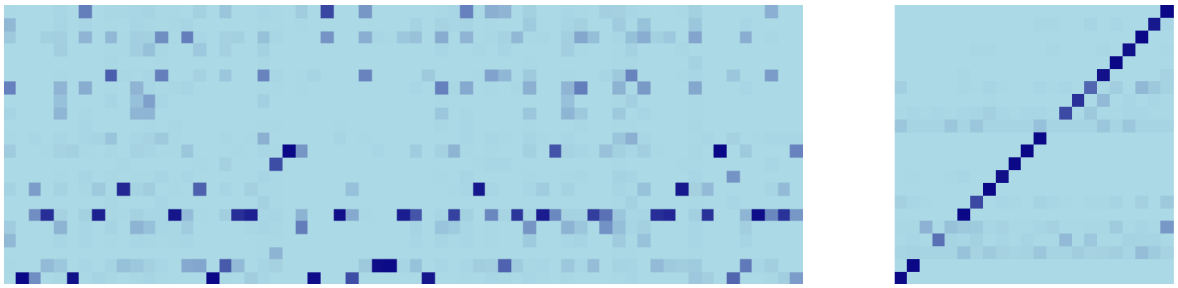


Figure 10: Associating groups of names, under the most basic form of our algorithm. Each row corresponds to a b-group (essentially, a family) identified in the BDM data. Columns correspond to m-groups, ie. lists of *possible* siblings identified via a completely different processing pipeline based on MLO data. The colour indicates preference for that b (row) as the originating family, given that m (column) as the evidence. Darker colours represent greater confidence. On the right, we test $\mathbf{b}$ data against itself, with a high error rate in transcribing letters. A strong diagonal band in this figure indicates the method is usually able to recover the true source despite the substantial change in the names.

Since we do not have absolute ground truth, expert opinion will be critical in testing hypotheses about the model and seeking improvements. However we can do a partial test simply by taking elements from the $\mathbf{b}$ set itself, adding "noise" (changes to the names) and asking whether the model can recover the correct source, since in this case we effectively possess "ground truth". Figure 10 *(right)* shows this for the same $\mathbf{b}$ data (rows) and and a noise probability of 0.4, meaning that around 40% of the letters are randomly corrupted in generating the so-called $\mathbf{m}$ data (columns) as a test.

This suggests that good guesses are possible provided the correct group is actually among the available b options, but we also need to detect (and reject) the possibly numerous cases where no such good option is present ('false positives'). One option is to make use of the Shannon entropy of the posterior distribution $P(b|m)$, which reflects how much the distribution is spread out over the available families. For example in Figure 10 the first column is more equivocal than the second, and would have a higher entropy. Low entropy points to a strong winner or winners, and we might reasonably reject all that exceed some threshold. To evaluate this idea we need some notion of ground truth, which in general we don't have. Instead, and as earlier, we can take actual b-groups from some set $\mathcal{B}_{\mathbf{m}}$, add 'noise' to them in the form of new / omitted names and mis-spellings, and use them as proxy m-groups in a test. From these 'pretend' m-groups we can then try to distinguish between the (known) b-groups that lead to them ($\mathbf{b} \in \mathcal{B}_{\mathbf{m}}$), or a *different* set of b-groups, corresponding to a (randomly chosen) other surname ($\mathbf{b} \in \mathcal{B}_{\mathbf{m}'}$). Without conditioning on the surname, can entropy distinguish between these two cases?

Figure 11 shows the results on these two groups. In order to simulate the poor data quality of MLO, the m set is a substantially corrupted and augmented version of the original b, as follows. Each letter of each name has 0.2 chance of getting changed to another letter, and each word has a 0.2 chance of being removed or joined with another word. There is also a 0.2 chance of having an additional word added. A family *[ralph, morton, harold, oscar, arnold, james, myra, ellen, colleen, alice]*, will become much harder to decipher when noise of this kind is introduced, becoming for example: *[ralphomortdn, harxudmoscar, arnold, jameo, ellen, collren, acize]* . This "noise" is applied in all the tests reported below. The figure suggests that entropy may nonetheless be used to distinguish between the two cases.

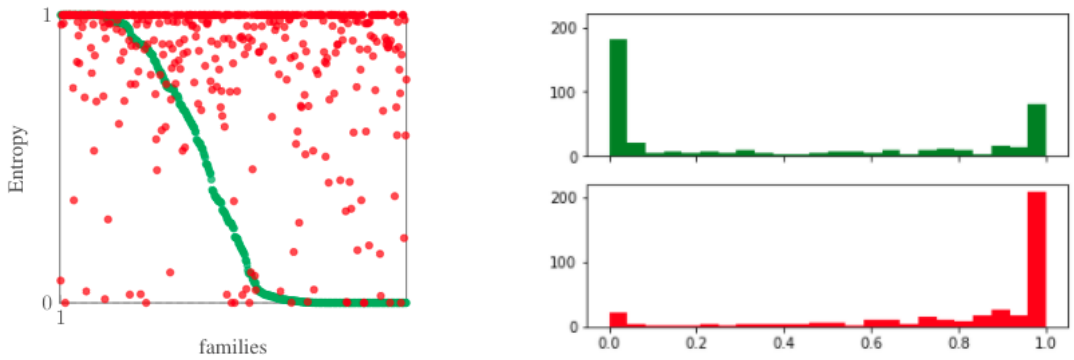


Figure 11: Distributions of the entropy of the posterior over b groups, for artificially altered data in which noise is set at 0.2 (see main text) - this generates substantial change in most of the names. *Left*: Each green point represents a family that does originate in $\mathbf{b} \in \mathcal{B}_{\mathbf{m}}$ (ie. a case we would like to pass) while each red is the same for a different (and wrong) set of families, which we would like to exclude. The families appear on the abscissa sorted by the former value, hence the green values decrease from left to right. Most cases that have high entropy should be excluded (red) while most that are low should ideally pass (green). *Right*: Histograms of entropies from the two cases, suggesting it can be used to distinguish them apart in the majority of cases. Note however some examples that we would prefer to pass do acquire high values under our noise process.

Table 2 shows larger scale results over the same test environment as the figures in 11, and at different threshold levels. These are aggregate results over 446 families and 60 surnames. A true positive is considered to occur when the linkage is correct *and* it is below the entropy threshold. An example of this is seen in Table 2 where one would assume that the true positive accuracy at 1 would also be 1. The "missing" 0.171 of true positives is caused by the incorrect linkage rather than entropy rejection where the $p(\mathbf{b}|\mathbf{m})$ was higher for the wrong family than the real original B. The entropy threshold appears robust, in light of the highly corrupted data used in these tests. Table 2 displays high precision scores, and reasonable accuracy scores. The F1 score ideally peaks at 1, while the best F1 score we received is 0.7, probably because the measure does not include the true negatives. The F1-score is typically used when the False Negatives and False Positives are crucial, whereas in our case True Positives and True Negatives (captured by the Accuracy) are more important. Tests show similar robustness to the setting of our other main model parameter β (here it is 0.5).

Accuracy of Entropy rejection method with noise							
Threshold	True +	False -	True -	False +	Precision	Accuracy	F1
0.100	0.315	0.685	0.994	0.006	0.981	0.655	0.477
0.300	0.428	0.572	0.954	0.046	0.903	0.691	0.581
0.500	0.538	0.462	0.914	0.086	0.862	0.726	0.663
0.700	0.613	0.387	0.834	0.166	0.787	0.723	0.689
0.900	0.735	0.265	0.655	0.345	0.680	0.695	0.707

Table 2: Accuracy for the test dataset. We use accuracy, precision and the f1 score to assess the performance at various threshold levels. Accuracy is a measure of truth, precision is a measure of variability and F1 is a the harmonic mean of the precision and recall (recall being the fraction of relevant instances retrieved), which is a commonly used measure for natural language processing applications. Precision, Accuracy, and F1 are high and robust over a wide range of the threshold.

To summarise, we adopt a combination of *(i)* an entropy-based filter performing rejection of negatives, and *(ii)* ranking via the posterior, for identification of the group-of-origin of a set of names. Even with the very substantial corrupting processes used in our test, we are able to reject a substantial majority of the negative cases and, within the positives, to successfully identify the correct "source" families. While the end application does not provide ground truth, we expect the same method to provide useful information "in the wild", at least to the extent that the character of actual name change is comparable with our test's augmentation process.

4.2.1 Inference with Alignment

This section motivates and outlines a way forward in addressing the full alignment problem, which is work in progress. An immediate "win" of taking alignment seriously would be the ability to use a sex-dependent model for name change, something that is not possible because our BDM Historical data does not include gender. In practice it is also, of course, very natural to think of a specific alignment when considering the plausibility of one b-group versus another: "What if James W in b is Jimmy in m?" and so on. But in trying to carry out formally consistent inference, a much more significant reason to incorporate alignment (indeed multiple possible alignments) into the picture is that it gives a more correct answer.

The fact that the core linkage of interest is between whānau (rather than the specific identities of individual people making up those whānau) results in a potentially high computational burden since, in a generative model, quantifying the former correctly must involve integrating out the latter, as explained below.

By the sum and product rules of probability, the overall likelihood $P(\mathbf{m} \mid M, \mathbf{b})$ can be expanded into a sum over possible alignments $\mathbf{z}$:

$$\underbrace{P(\mathbf{m} \mid M, \mathbf{b})}_{F} = \sum_{\mathbf{z} \in \mathcal{Z}} P(\mathbf{m} \mid \mathbf{z}, M, \mathbf{b})\ P(\mathbf{z} \mid M, \mathbf{b}) \tag{3}$$

We can drop M in the first term, and the $\mathbf{b}$ in the second is effectively just B, leaving

$$\sum_{\mathbf{z} \in \mathcal{Z}} \underbrace{P(\mathbf{m} \mid \mathbf{z}, \mathbf{b})}_{f(z)}\ \underbrace{P(\mathbf{z} \mid M, B)}_{p(z)} \tag{4}$$

Thus the quantity we need to calculate has the form of a large sum, $F = \sum_z f(z)\, p(z)$. The z-specific likelihood $f(z)$ can be found by a mixture of Ngrams as described above, while $p(z)$ is a prior on the number of genuine linkages. If we know F up to a proportionality constant for each $\mathbf{b} \in \mathcal{B}$ we can find the quantity we are really interested in: $P(\mathbf{b}|\mathbf{m})$, by normalising over all the F values arrived at for $\mathbf{b} \in \mathcal{B}$.

The "brute force" approach is to find F exactly, by calculating the whole sum, but this means working out all of the different possible alignments/linkages and then calculating the $p(x)$ for each, of which there are a potentially huge number (factorial in the group sizes). Computing the entire sum will only be feasible for combinations in which one of the groups is very small. The computational intractability of this sum is a significant theoretical obstacle to drawing inferences about identity in a principled way.

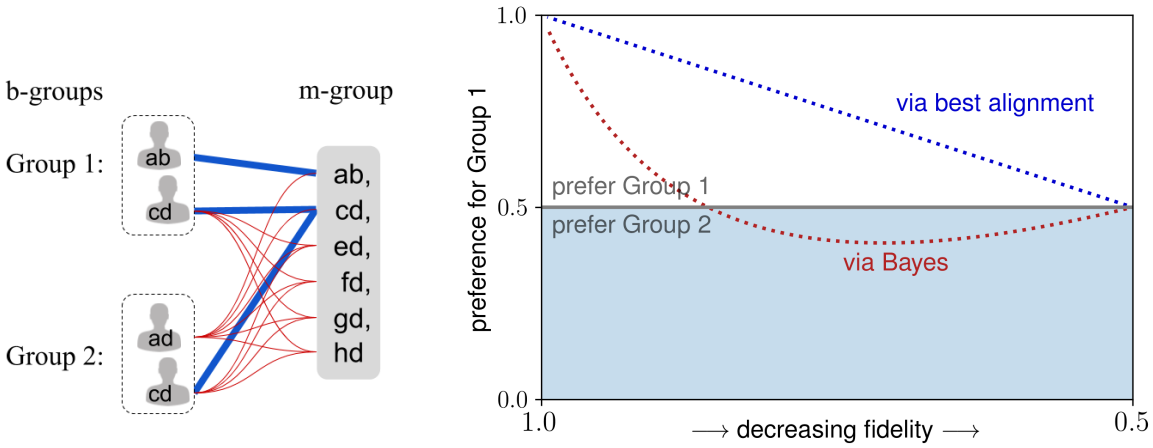


Figure 12: Why the awkward computation could matter. Consider a borderline case (*left*) with just two b-groups, each with 2 members, and an m-group with 6 members. Constructed names are used for illustration: blue lines indicate perfect matches, and red lines partial matches. The best overall alignment in terms of good matches is obviously Group 1. The plot (*right*) shows the preference for one $\mathbf{b}$ group over the other, as the model parameter controlling fidelity is decreased (fidelity is the probability that a character is left the same by the transition $b \rightarrow m$, in a zeroth-order Markov model for simplicity). The optimal single alignment (blue) always favours Group 1, but fully Bayesian inference (red) would switch preferences, depending on the level of discontinuity we think exists between $b \rightarrow m$. We are, of course, unsure of the *true* fidelity. This uncertainty leads, in this case, to an appropriate ambivalence as to which underlying Group is the correct one. Using the optimal alignment alone leads to an unjustified confidence.

Figure 12 illustrates why the distinction between optimising (picking the best alignment) *vs* integrating (doing the whole sum) could matter for our application. Depending on what we believe about the causal relationship between BDM data and MLO data, the two approaches might make different conclusions about how they rate competing hypotheses ($\mathbf{b}$ groups) for the identity of a given $\mathbf{m}$ group. An explicit "self-doubt" about exact identities adds to the credibility of results, and is built into a fully Bayesian solution.

In ongoing work we are exploring two potential estimates for F that remain tractable to compute, even for large groups. One is to approximate it by the f value of the *best* alignment, which can be found in polynomial time by an optimisation algorithm based on dynamic programming (Bellman [1961]). While fast, this could be a poor approximation in some cases because it commits to one specific alignment (ie. one set of identities) even when the evidence remains equivocal. A second approach is to use a form of Importance Sampling (Press et al. [2007], Lee [2012]) to generate a Monte Carlo estimate of F. There are two potential advantages of this over the optimisation approach. Firstly, it directly addresses the correct F (meaning it takes appropriate account of ambiguous identities, rather than just "taking the best match"). And secondly, the computational load involved is readily tuned up and down as needed: more computational resource can generate better approximations when uncertainty is large, or be cut back when the evidence is clear. We speculate that an approach in which dynamic programming is used to initialise an importance sampler might give the best of both worlds, but this is work in progress.

V CONCLUSION: DECOLONISING THE FIELD

Figure 13: A New Zealander by Parkinson and Ajax by Ezekias play draughts. Lithograph from *Southern Myths The Odyssey of Captain Cook*, Marian Maguire [2005]

Kimihia te Matangaro - finding missing Māori shareholders is a local challenge with real world impact and one also of universal import. Working within Indigenous paradigms has required a profound flaxroots rethinking of community collaboration, research design and computational approaches. Our ongoing research journey has highlighted several key questions: are the computational tools and techniques developed for predominantly western/European digital humanities suitable for Indigenous worldviews, languages and practices; are the philosophies and methodologies underpinning digital humanities culturally aware; does the field's emphasis on open access and open data perpetuate a neocolonial agenda? In discussing the historian's place in indigenisation and decolonisation, Hogan and McCracken [2016] remark that "indigenization cannot be attempted without first making space to decolonize what types of knowledge the academy sees as legitimate, otherwise projects have the potential to become tokens used to absolve settler guilt." Similarly, Roopika Risam [2018] explains that digital humanities' diversity agenda has occluded the need for a greater self-awareness of the field's own colonising theories and practices. The dominant narratives of digital humanities driven by the Global North relegate Indigenous perspectives, positionality, and practices to subaltern status and deny agency. They have also derailed deep engagement with decolonising the production of knowledge. She advocates for "the creation of new methods, tools, projects, and platforms to undo the epistemic violence of colonialism" and celebrates the "hybridity, plurality, contradiction, and tension that are necessary strategies of decolonization" (Risam [2018]). One approach deployed in the context of HGIS has been 'indigitalization' described by Palmer [2012] as "the amalgamation of

indigenous, scientific and digital technological knowledge systems; characterised as fragmentary, contradictory, and full of uncertainties." In their enactment of decoloniality, Mignolo and Walsh [2018] would agree: "decoloniality is not a new paradigm, or mode of critical thought. It is a way, option, standpoint, analytic, project, practice, and praxis." It is, moreover, a creative force of resistance that reimagines and celebrates re-existence. By focusing on the local and situated nature of knowledge, decolonial computing (Ali [2016]) amongst other decolonising dh strategies can rewrite the relationship between space and time (De Landa [1997], De Landa [2016]) embrace the complexity of Indigenous cultures, and resist the decoupling of decolonial projects like ours from the rematriation of Indigenous land and lives. But, as contemporary New Zealand artist Marian Maguire [2005] suggests, we all need to sit around the same table, under the same maunga (Figure 13). Decolonising, localising, and indigenising practice is more than an awareness of data provenance, algorithmic bias, uncritical tool use. Tuck [2018] eloquently argues that "decolonization is not the endgame, not the final outcome of a long process, but the next now, the now that is chasing at our heels." Enacting data sovereignty and stewardship, and working with/in communities for their collective benefit shape and sustain our dialogic interactions, our enduring kōrero.

Acknowledgements

Our researchers' iwi and hapū affiliations: Rhys Owen, Te Rarawa; Rere-No-A-Rangi Pope, Ngāruahine; Pikihuia Reihana, Ngapuhi (Ngati Hine), Ngati Kahungunu (Ngati Kere), Rangitane ki Wairau, Ngai Tahu. He tangata, he tangata, he tangata [*It is the people, it is the people, it is the people*].

To our research colleagues at Parininihi ki Waitotara (PKW): E Mitchell, he mihi nunui tēnei ki tō tautoko i a mātou. Whakatere koe te waka nei i a tātou katoa e hoe ana. He tangata māhaki, he tangata awhina. Koia ngā tohu o te taniwha hikuroa. Waimarie mātou ki te mahi i tō ake taha. Tēnā koe. E Adrian, e te puna o te mātauranga, tēnā rā koe. He tangata koi, he tangata matatau i tēnei ao o te whakapapa me te ture māori. He taonga tō mōhiotanga ki ngā tāngata e hiahia ana ki te whakahokia ki tō rātou whenua. Tēnei te mihi kau ana ki a kōrua tahi.

The authors acknowledge support and funding from the Science for Technological Innovation [2020] National Science Challenge Spearhead Project "Analytics to identify and connect successors to whenua"; the Digital Learning and Research group at VUW, and the use of the Rāpoi [2020] cluster; the use of New Zealand eScience Infrastructure (NeSI [2020]) high performance computing facilities as part of this research. New Zealand's national facilities are provided by NeSI and funded jointly by NeSI's collaborator institutions and through the Ministry of Business, Innovation & Employment's Research Infrastructure programme.

References

Syed Mustafa Ali. A brief introduction to decolonial computing. *XRDS: Crossroads, The ACM Magazine for Students*, 22(4):16–21, June 2016. ISSN 15284972. doi: 10.1145/2930886.

M. M. Bakhtin and Caryl Emerson. *Problems of Dostoevsky's poetics*. Number v. 8 in Theory and history of literature. University of Minnesota Press, Minneapolis, 1984. ISBN 9780816612277 9780816612284.

BDM. Birth, death and marriage historical records, 2020. URL https://www.bdmhistoricalrecords.dia.govt.nz/.

Richard Bellman. *Adaptive control processes: a guided tour*. Princeton University Press, Princeton, N.J, 1961.

Richard Boast. Te Ara - The Encyclopedia of New Zealand, July 2015. URL http://www.TeAra.govt.nz/en/te-tango-whenua-maori-land-alienation/page-4.

Debbie Broughton and Kim McBreen. Mātauranga Māori, tino rangatiratanga and the future of New Zealand science. *Journal of the Royal Society of New Zealand*, 45(2), 2015. doi: 10.1080/03036758.2015.1011171.

Andrea Byrom. Why should we include Vision Mātauranga and Mātauranga Maori in our research? In *Not Your Usual National Meeting*, Te Papa,Wellington, May 2017.

Reitu Cassidy-Robson and Paul Harris. *Te Whanau Harris*. Reitu Robson and Paul White, 1980.

CIDOC-CRM. Cidoc conceptual reference model (crm), 2015. URL http://www.cidoc-crm.org/sites/default/files/cidoc_crm_version_6.2.pdf.

Manuel De Landa. *A thousand years of nonlinear history*. Swerve editions. Zone Books, New York, 1997. ISBN 9780942299311 9780942299328.

Manuel De Landa. *Assemblage theory*. Speculative realism. Edinburgh University Press, Edinburgh, 2016. ISBN 9781474413626 9781474413633. OCLC: ocn953197501.

Mason Durie. Exploring the interface between science and indigenous knowledge. In *Capturing Value From Science*, Christchurch, March 2004.

Ted Enamorado, Benjamin Fifield, and Kosuke Imai. Using a probabilistic model to assist merging of large-scale administrative records. *Available at SSRN 3214172*, 2018.

Ivan P. Fellegi and Alan B. Sunter. A theory for record linkage. *Journal of the American Statistical Association*, 64(328):1183–1210, 1969. ISSN 01621459. URL http://www.jstor.org/stable/2286061.

Andy Fyers and John Hartevelt. Nā Niu Tīreni / New Zealand Made, July 2018. URL https://interactives.stuff.co.nz/2018/07/na-niu-tireni-new-zealand-made/.

David Gaertner. Why we need to talk about indigenous literature and the digital humanities, 2017. URL https://novelalliances.com/2017/01/26/indigenous-literature-and-the-digital-humanities/.

Heather Gifford and Kirikowhai Mikaere. Te Kete Tū Ātea: Towards claiming Rangitīkei iwi data sovereignty – Te Mauri – Pimatisiwin. *Journal of indigenous wellbeing*, July 2019.

Shawn Graham, Ian Milligan, and Scott Weingart. *Exploring big historical data: the historian's macroscope*. Imperial College Press, London, 2016. ISBN 9781783266081 9781783266371.

Jennifer Guiliano and Carolyn Heitman. Indigenizing the digital humanities: Challenges, questions, and research opportunities. Digital Humanities 2017, 2017. URL https://dh2017.adho.org/abstracts/372/372.pdf.

Garth R. Harmsworth and Shaun Awatere. *Indigenous Māori knowledge and perspectives of ecosystems*. Manaaki Whenua Press, Lincoln, New Zealand, January 2013. URL http://api.digitalnz.org/records/35867430/source.

D. Hikuroa. Mātauranga Māori—the ūkaipō of knowledge in New Zealand. *Journal of the Royal Society of New Zealand*, 47(1):5–10, January 2017. doi: 10.1080/03036758.2016.1252407.

Skylee-Storm Hogan and Krista McCracken. Doing the work: The historian's place in indigenization and decolonization, 2016. URL http://activehistory.ca/2016/12/doing-the-work-the-historians-place-in-indigenization-and-decolonization/.

Maui Hudson, Tiriana Anderson, Te Kuru Dewes, Pou Temara, Hēmi Whaanga, and Tom Roa. "He Matapihi ki te Mana Raraunga" - Conceptualising Big Data through a Māori lens. In *He whare hangarau Māori language, culture & technology*. Te Pua Wānanga ki te Ao, Te Whare Wānanga o Waikato, 2017. ISBN 9780473426927.

Vivienne Kennedy. Social network analysis and research with māori collectives. *MAI review*, 2010(3), 2010. ISSN 1177-5904. URL http://www.review.mai.ac.nz/mrindex/MR/article/view/372/567.html.

Tanira Kingi. Maori landownership and land management in New Zealand. In *Making land work*, volume 2, pages 129–151. AusAID, Canberra, A.C.T., 2008. URL https://www.dfat.gov.au/sites/default/files/MLW_VolumeTwo_CaseStudy_7.pdf.

Tahu Kukutai and Donna Cormack. Census 2018 and implications for Māori. *New Zealand population review*, 44: 131–151, 2018. ISSN 0111-199X. URL https://population.org.nz/app/uploads/2019/02/NZPR-Vol-44_Kukutai-and-Cormack.pdf.

Tahu Kukutai and John Taylor. *Indigenous Data Sovereignty*. ANU Press, November 2016. ISBN 9781760460310. doi: 10.22459/CAEPR38.11.2016. URL https://press.anu.edu.au/publications/series/caepr/indigenous-data-sovereignty.

Tahu Kukutai and Melinda Webber. Ka Pū Te Ruha, Ka Hao Te Rangatahi: Maori identities in the twenty-first century. In *A land of milk and honey? Making sense of Aotearoa New Zealand*. Auckland University Press, 2017.

Peter Lee. *Bayesian Statistics: An Introduction, 4th Edition*. Wiley, 1st edition edition, 2012.

Raymond Lovett, Vanessa Lee, Tahu Kukutai, Donna Cormack, Stephanie Rainie, and Jennifer Walker. *Good data practices for indigenous data sovereignty and governance*. Institute of Network Cultures, 2019. ISBN

9789492302274. URL https://hdl.handle.net/10289/12919.
Marian Maguire. A New Zealander by Parkinson and Ajax by Exekias play draughts, 2005.
Ocean Mercier. Mātauranga and science. *New Zealand science review*, 74(4):83–90, 2018. ISSN 0028-8667.
Walter Mignolo and Catherine E. Walsh. *On decoloniality: concepts, analytics, praxis*. On decoloniality. Duke University Press, 2018. ISBN 9780822370949 9780822371090 9780822371779.
Kevin P. Murphy. *Machine Learning: A Probabilistic Perspective*. MIT Press, 2012.
Māori Land Online. Māori land online, 2020. URL https://www.maorilandonline.govt.nz.
NeSI. New Zealand eScience Infrastructure, 2020. URL https://www.nesi.org.nz/.
Ngā Upoko Tukutuku. Ngā upoko tukutuku / māori subject headings, 2020. URL https://natlib.govt.nz/librarians/nga-upoko-tukutuku.
NZ Institute of Economic Research. *Māori Economic Development: te Ōhanga Whanaketanga Māori*. NZ Institute of Economic Research, Wellington, N.Z, 2003.
T. W. Palin. Sketch map of the north island of new zealand shewing native tribal boundaries, topographical features, confiscated lands, military and police stations, etc. 1869., 1869. URL https://kura.aucklandlibraries.govt.nz/digital/collection/maps/id/710.
Mark Palmer. Theorizing indigital geographic information networks. *Cartographica: The International Journal for Geographic Information and Geovisualization*, 47(2):80–91, 2012. doi: 10.3138/carto.47.2.80.
Parininihi ki Waitotara. Parininihi ki Waitotara, 2020. URL https://pkw.co.nz/.
William H Press, Saul A Teukolsky, William T Vetterling, and Brian P Flannery. *Numerical recipes 3rd edition: The art of scientific computing*. Cambridge university press, 2007.
Roopika Risam. Decolonizing The Digital Humanities In Theory And Practice. *The Routledge Companion to Media Studies and Digital Humanities*, May 2018. URL https://digitalcommons.salemstate.edu/english_facpub/7.
Mike Ross. The throat of Parata. In *Imagining Decolonisation*, BWB texts. Bridget Williams Books, March 2020. ISBN 9781988545783. doi: 10.7810/9781988545783.
Rāpoi. Rāpoi VUW's high performance compute cluster, 2020. URL https://vuw-research-computing.github.io/raapoi-docs/.
Mauricio Sadinle et al. Detecting duplicates in a homicide registry using a bayesian partitioning approach. *The Annals of Applied Statistics*, 8(4):2404–2434, 2014.
Science for Technological Innovation. Analytics to identify and connect successors to whenua, 2020. URL https://www.sftichallenge.govt.nz/our-research/projects/spearhead/analytics-to-identify-and-connect-successors-to-whenua/.
Science for Technological Innovation NSC, Data ILG, and Victoria University of Wellington. Māori Data Futures - Hui Report. In *Māori Data Futures*, Wellington, NZ, May 2018. URL https://www.sftichallenge.govt.nz/assets/Uploads/Download-PDFs/Maori_Data_Futures_Report-2018.pdf.
Science for Technological Innovation NSC, Data ILG, and Te Hiku Media. Māori data futures – intellectual property. In *Māori Data Futures*, 2019. URL https://www.sftichallenge.govt.nz/assets/Uploads/Download-PDFs/Maori-Data-Futures-Report-2019.pdf.
Kevin Shedlock and Marta Vos. A Conceptual Model of Indigenous Knowledge Applied to the Construction of the IT Artefact. In *31st Annual CITRENZ conference*, Wellington, NZ, July 2018. URL https://www.citrenz.ac.nz/conferences/2018/pdf/2018CITRENZ_1_Shedlock_Indigenous.pdf.
Andrea Siodmok. Lab Long Read: Human-centred policy? Blending 'big data' and 'thick data' in national policy, January 2020. URL https://openpolicy.blog.gov.uk/2020/01/17/lab-long-read-human-centred-policy-blending-big-data-and-thick-data-in-national-polic
Linda Tuhiwai Smith. *Decolonizing methodologies: research and indigenous peoples*. Zed Books, London & New, second edition edition, 2012. ISBN 9781877578281 9781848139503 9781848139510. OCLC: 809167003.
Stats NZ. Integrated Data Infrastructure, July 2018. URL https://www.stats.govt.nz/integrated-data/integrated-data-infrastructure/.
Rebecca C Steorts, Rob Hall, and Stephen E Fienberg. A bayesian approach to graphical record linkage and deduplication. *Journal of the American Statistical Association*, 111(516):1660–1672, 2016.
Te Mana Raraunga. Te Mana Raraunga - Māori Data Sovereignty Network charter, 2016. URL https://static1.squarespace.com/static/58e9b10f9de4bb8d1fb5ebbc/t/5913020d15cf7dde1df34482/1494417935052/Te+Mana+Raraunga+Charter+.
Eve Tuck. Losing patience for the task of convincing settlers to pay attention to indigenous ideas. In *Indigenous and Decolonizing Studies in Education : Mapping the Long View*. Routledge, 2018. doi: 10.4324/9780429505010.
Pakake Winiata. Guiding kaupapa of Te Wānanga-o-Raukawa. Technical report, Te Wānanga-o-Raukawa, 2001. URL https://www.wananga.com/user/inline/2/Guiding_Kaupapa.pdf.