Shared Data Visualisations for Private Data: Managing Ownership and Visibility

JUDY BOWEN, University of Waikato, New Zealand
ANNIKA HINZE, University of Waikato, New Zealand

Data ownership and data governance relate to the right to access and share data and the question of who retains this right once data has been collected. In the context of personal and shared health monitoring systems, this question is particularly pertinent. When health monitoring is implemented within the workplace (e.g., as sensor-based Internet of Things applications), the rights of both workers and the employers/company owners may be challenging to manage. This paper explores how control over shared data may be implemented for personal and public data visualisations. We use the case study of a forestry fatigue monitoring system to explore the challenges introduced by internal and external stakeholders, and their interests in data ownership and access to aggregated and longitudinal data. We propose the concept of federated data sharing, and outline how the contributors to shared data visualisations may retain ownership and access to raw data, while giving access to relevant data to the different stakeholders and interested parties.

ACM Reference Format:
Judy Bowen and Annika Hinze. 2024. Shared Data Visualisations for Private Data: Managing Ownership and Visibility. 1, 1 (September 2024), 17 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

In an era where the Internet of Things (IoT) has become an integral part of our daily lives (using technologies such as smart watches, smart clothing, and household technologies etc.), ensuring we have appropriate methods for handling large quantities of streaming data from multiple sources is paramount. This includes considerations of data use, data ownership and data sovereignty.

Data use refers to the different ways data is processed, analysed, visualised etc. after collection. In many commercially-available proprietary systems, data is collected and then streamed to a cloud service for analysis. There it is processed in different ways (depending on the domain of use) with resulting data returned to the individual user, typically as charts visualising trends over time or as insights determined from the data (see Figure 1).

Data ownership concerns the right to access and use the data after it has been collected. Irrespective of the origin of the data (e.g., an individual using wearable technology or smart devices within a smart home), ownership of the data – once collected – typically resides with the company providing the technology and analysis services. It is unusual for users to be able to gain access to the raw data captured; more typically, users are restricted to visualisations or aggregated data provided post-hoc. Users may receive little information on how the data was analysed or aggregated. Furthermore, there may be limited transparency to how the data is used (and continues to be used) once it has been collected. The issues of both ownership and use contribute to a lack of data sovereignty.

Authors' Contact Information: Judy Bowen, University of Waikato, New Zealand, jbowen@waikato.ac.nz; Annika Hinze, University of Waikato, New Zealand, annika.hinze@waikato.ac.nz.

ACM XXXX-XXXX/2024/9-ART
https://doi.org/XXXXXXX.XXXXXXX

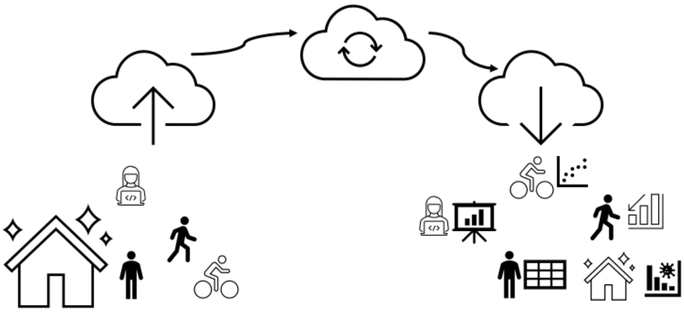


Fig. 1. Typical Data Flow Model for IoT Solutions

Data sovereignty is the concept that data that is collected or stored in a particular geographic location should be subject to the laws of that location. Those laws may then provide rules or guidelines regarding both use and ownership. For example, the European Union's General Data Protection Regulations (GDPR)[1] seeks to give EU citizens the right to full access to their data and control over how it is used. The notion of data sovereignty has been extended in some parts of the world as the basis for Indigenous Data Sovereignty [13]. *Indigenous Data Sovereignty* describes the data rights and interests of indigenous peoples. It is the right of Indigenous peoples to determine the means of collection, ownership, access, use, and dissemination of data pertaining to the Indigenous peoples from whom it has been derived, or to whom it relates [24].

While data sovereignty laws and governance define what is permitted and how data providers should retain rights and control, they do not mandate *how* this should be achieved. This paper addresses this issue by defining a mechanism for aggregating, sharing and visualising data guided by the principles of Indigenous Data Sovereignty. This means that individuals from whom data is collected retain their rights to control use in perpetuity, can control how and where their individual data is stored and used, and retain all of these rights whenever their data is aggregated into larger data sets.

As a case study to elaborate on the practical issues that need to be addressed and also to describe potential solutions, we refer to a long-running project with forestry workers in Aotearoa/New Zealand (described fully in the next section). This involves the collection and aggregation of a variety of different data types which are used in real-time to support safety (through fatigue management) and post-hoc to provide longitudinal data to the workers, work teams, worker families and communities, companies, government etc. The value of the streaming IoT data lies not only in real-time pattern analysis for the individual (e.g. current fatigue markers for individual forestry workers) but also in aggregating streams (e.g., detecting trends over time for groups of workers) and historical analysis (e.g., identifying at-risk work teams or measuring improvements in health and safety). Creating visualisations for the different stakeholders in the forestry project requires understanding which parts of the data are available for which groups of stakeholders (based on personal privacy, commercial sensitivities etc.) and how to provide meaningful results to all stakeholders irrespective of these privacy requirements. In addition, the high proportion of forestry workers that identify as Māori (the indigenous population of Aotearoa/New Zealand) require us to adhere to, specifically, Māori data sovereignty principles [25]. This leads to a number of challenges for the required data visualisations. By investigating ways of addressing these challenges, we now

[1] https://gdpr-info.eu/

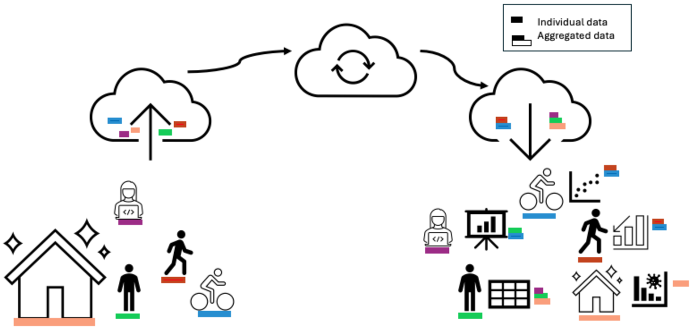


Fig. 2. Visualising Aggregated Data to Multiple Users

propose a solution for data management which can be generalised across all large-scale multi-data situations.

The forestry project described above relies on streaming data analysis for real-time alerting as well as aggregation (from both multiple people and places and multiple time periods) for post-hoc reporting to different stakeholders. This requires: robust algorithms to facilitate pattern detection in real time; practical measures to safeguard information from unauthorised use in the future; methods to aggregate data for collective use; the ability for individuals to retain access to their own individual data; and easy ways of controlling how individual data is used in aggregations and by whom. Similarly, data governance and use for communities, such as whānau (Māori family groups), iwi (Māori communities), and work units, have not been considered for IoT data streams [2, 23]. Users of IoT smart devices typically encounter limitations to full data ownership and control, usually only being provided overview graphs and simplified data views. This lack of comprehensive access to all the data generated by their smart devices reduces trust and limits opportunities to use the data. Within the forestry project this required the development of bespoke hardware for the data gathering mechanisms to ensure we could then fully control the data and determine where it was stored, this led to the need to develop solutions to suitably manage such streaming, personal data. The solution we propose is intended to support both individual data ownership as well as collective use and oversight.

Figure 2 enhances the standard model of data transfer shown in Figure 1 by showing how individual data from different sources can be aggregated for different visualisation or analysis purposes, but still be retained in its original form for the data owner.

In the next section, we describe the forestry case study in more detail. Elaborating on the intended use of the different data visualisations proposed for the project, we identify a set of challenges that arose when trying to support the management of data in the manner described above. Following this we describe a new approach for data management, which we call federated data sharing, which addresses these challenges, We describe how this approach can be practically implemented in the forestry scenarios described and more generally within similar scenarios of use. We next discuss related work relevant to our solution and identify the contributions made by our work. Finally we conclude with an explanation of how our work might be generalised and what future work is suggested by this.

2 Background to the Forestry Project

Since 2015 our research team has been undertaking a number of different projects using IoT and wearable technology in New Zealand forestry. The forestry industry in New Zealand has one of the highest number of fatalities and serious injuries across the country.[2] The project investigated how the use of technology in rugged outdoor environments can be used to support health and safety in a variety of different ways. For example, predicting potential health hazards in outdoor work situations by using lightweight, wearable technology[7], relying on known correlations between mental and physical fatigue [11] and hazardous situations. Forestry work involves manual labour in combination with heavy machinery, and other solutions developed within the project use beacons, and worker locations, to define safe-zones which are visualised to machine operators in their cabs or to monitor worker proximity to machinery [14]. Many of the early challenges for this project involved the development of wearable and other IoT solutions that could be used in rugged outdoor workplaces with limited power and internet connectivity. As the project matured, the focus switched to how the data that is gathered can be used in practice and the different uses and visualisations that are required.

Here we focus on the wearable solution developed to support worker safety through fatigue monitoring as this has the most complex set of requirements regarding the data sharing and management. Part of the solution includes a smart shirt worn by forestry workers which records biometric information (such as heart-rate, heart-rate variability, galvanic skin response) which is used in combination with contextual factors (work role, outdoor temperature, terrain types etc.) to identify when the worker is exhibiting signs of fatigue [9]. Fatigue has been shown to be a major contributor of accident risk in hazardous outdoor work environments [10] and the wearable solution is intended to prevent this by alerting workers in real-time so that they can rest and recover. In addition, a buddy monitoring solution was developed to enable co-located workers to support each other (if worker A is fatigued and does not respond to notifications to rest, their buddy, worker B, is also notified so they can encourage worker A to rest). In emergency situations (such as a ‘man down’ incident) supervisors and managers need to be notified so that appropriate action can be taken. The data collected is also used to provide an overview of health and well-being of workers, both individually and collectively. This allows workers to view their fatigue statistics over time (which may support wider lifestyle and health choices) and also allows employers to have an overview of work teams over time. Figure 3 provides an overview of how the data is collected and disseminated.

Development of the data visualisations for post-hoc, aggregated data, initially focused on designing suitable ways of making the data informative for users who may have low digital literacy or little interest in data analytics. The users of the data visualisations consist of both primary users (the workers themselves) and secondary users who may consist of whānau/family and communities of the workers, work teams, worker supervisors, forestry owners, government statisticians etc. (this is the ‘Community feedback’ in Figure 3). In order to support development of requirements for the data, participatory design sessions were run with groups of forestry workers and their whānau/families. The information gathered from these sessions and analysis of this was subsequently used to develop personas and scenarios of use, to guide the design work [3, 9]. The personas were created using a data-informed approach [12] based on both the participatory design sessions as well as meetings and workshops conducted with workers, forestry supervisors, managers and forest owners. We describe these next.

[2] Forestry had a fatality rate of 56.73 per 100,000 workers in 2018. Forestry workers are 6 times more likely to be seriously injured and 22 times more likely to be fatally injured than in other NZ industries [28]

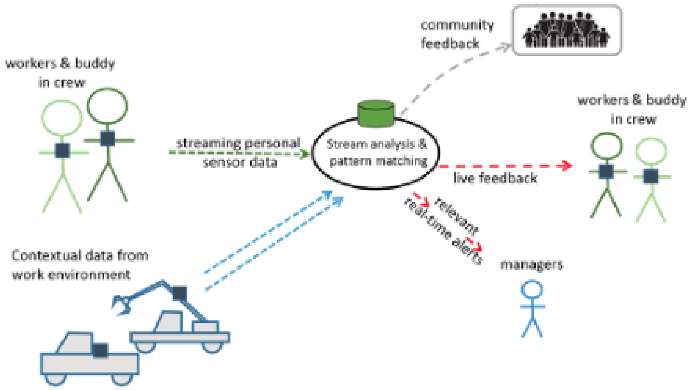

community feedback
workers & buddy in crew
streaming personal sensor data
Stream analysis & pattern matching
workers & buddy in crew
live feedback
relevant real-time alerts
managers
Contextual data from work environment

Fig. 3. Data flow for worker fatigue monitoring solution

2.1 Personas and Scenarios

Personas are a well-established and researched approach in interactive system design, employed by designers as a way of introducing archetypal users and their needs into the design process [8, 22]. They are particularly useful when access to end-users and other stakeholders is limited (as is the case for the forestry project, where distance and work requirements mean we cannot have regular engagement in-person with forestry groups). Different methods of developing personas and validating their effectiveness have been developed. Data-driven approaches often rely on large datasets, for example McGinn and Kotamraju conducted a survey of 1300 potential customers and used factor analysis on the resulting data to create what they argued were statistically valid personas [17]. In the forestry project, the personas were developed based on actual data which included NZ Government statistics on forestry worker profile, for example census data, see Figure 4, as well as data collected from participants in participatory design groups, interview and meetings which took place over several years in the early stages of the project. While this is a relatively small dataset, it has been shown by Faily and Flechais [5] that limited data can still produce effective and accurate personas and their work the uses a grounded theory analysis of empirical data to show the validity of personas developed from such small datasets. The successful design of our personas was validated during their use in a participatory design session (where they served as proxy users to aid concept and requirement elicitation [3]) when several workers were convinced that one of the personas was someone known to them, or perhaps someone's cousin.

In addition to the personas, use cases and scenarios were created for the different users. Table 1 presents a summary of a selection of the personas.

Each of the personas described in Table 1 should have a different view of the data. Jordan can see his own personal data as well as aggregated data across his work team. Within the aggregated data he cannot identify the other team members but can see which data is his. Kerry can see her husband Ari's data but not any of the aggregated data. Tamar can see aggregated data for all work teams that come under the umbrella of his management company. He cannot identify any individuals from the data but he can identify the different teams. Joe can see aggregated data for his workers (over time) and also individual trends for each worker. He cannot see any of the details of the personal data of each worker only the trends over time.

Table 1. Personas developed using forestry data

Persona Image	Persona Info	Example Scenario
	Jordan Henare: Jordan's father and three uncles all work in forestry and he was keen to join them as soon as he could leave school. His cousin Kahurangi works with the same crew and has been helping him to get used to the job. Jordan has asthma but he doesn't always remember to use his inhaler.	Jordan's boss has told him about the data dashboard that reports on his fatigue and safety status at work. When he arrives home in the evening he logs onto the website to take a look at the graphs. For a few days he compares how he is tracking each day, but once the novelty of this wears off he stops looking at it regularly.
	Kerry Donaldson: Kerry is 52 years old and works part-time as a teaching assistant. Her husband Ari works as a logger for a small family-run forestry business. She worries about Ari's safety at work and does not want her sons to work in the industry but recognises there are not many other opportunities for them locally.	Ari has been put on new medication by his doctor to help with sleep problems. Kerry is worried about the effect this is having on his fatigue levels. Each evening she logs into the dashboard to view his data and downloads weekly graphs showing trends over time. She plans to give these to Ari to take to his doctor the next time he visits.
	Tamar Fauolini: Tamar is an ex-forestry worker who now heads up the health and safety team for a large North Island forestry management company. He is responsible for reporting accident data to the Government and developing and implementing new safety initiatives for forestry crews.	Tamar is preparing the monthly newsletter and wants to include some graphs showing health and safety data from different regions. He logs into the dashboard and selects a monthly view of data by region, he can then download the graphs for each region and include these in the newsletter.
	Joe Gattis: Joe is an experienced bushman who runs his own contracting company which employs ten tree fellers. He has 3 children aged 10, 12 and 15 and currently lives alone. He spends most of his free time fishing and co-owns a small boat with his brother. Joe is very focused on health and safety and keen to try out anything that might help with that.	One of Joe's workers has been involved in a near-miss incident on a work-site. The management company have warned Joe that if this happens again his contract will be terminated. He logs into the dashboard and views the data for his team on the day of the accident. He then views the data for the previous week. He takes some screenshots and makes notes about the data to discuss with his team at the next morning's toolbox meeting.

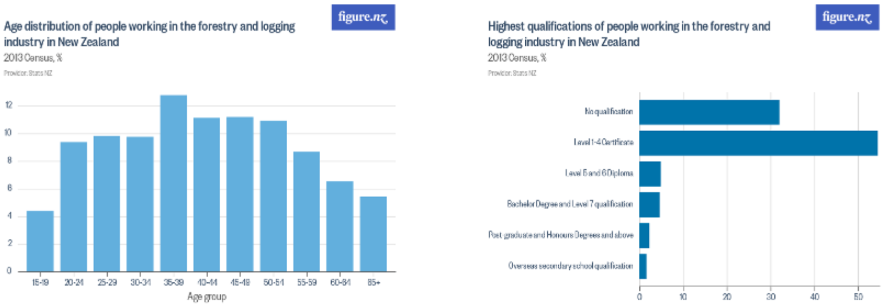


Fig. 4. Foresty worker age and education statistics, 2013 NZ census. (Figure.NZ using data from Stats NZ)

From the fully-developed scenarios, we developed requirements for a *data visualisation dashboard tool*. Some of the requirements focused specifically on aspects of data privacy and sharing, such as how a user controls sharing and which things they can share or hide. This also required definitions for the different use groups to control data views. We present an overview of the dashboard tool next.

2.2 Data Dashboard

The initial set of functional requirements for the data visualisation dashboard focused on a single user (the worker) and the access they enabled for their whānau/family. These initial requirements were:

- login
- view account information
- edit account information
- add whānau account
- edit whānau permissions
- view full data
- filter data
- select data items
- view user guide
- logout

To design appropriate visualisation methods, we explored research projects and commercial products that provided data dashboards and visualisations for different types of health and personal performance-based data. Here we briefly discuss on example of each type.

The EU CARRE project [21] provided data visualisations that aimed to empower patients with medical conditions, particularly chronic heart and kidney disease. A data dashboard was used to display visualisations that allow patients and clinicians to explore the risk associations and the possible development of disease – see Figure 5 (left). It also visualises patient data collected through sensors and enables comparisons between the risk association data and the patient's data which allows a better understanding of their disease progression. The types of risk associations and data over time is similar to the fatigue warnings and health metrics displayed for the forestry project. The time-dependent data is mapped across multiple metrics. Our interface adopted the dynamic adding and zooming of metrics similar to that of the CARRE visualisation.

Among the commercial dashboards we explored were the Hexoskin Connected Health Platform, Garmin Connect mobile application and Apple Health mobile application. Figure 5 (right) shows the interface of the Hexoskin platform. The analysis of commercial dashboards provided insights relating to use of colour, expandable information, time and user filtering methods and multiple views of a single dataset.

We identified common properties of such dashboards, such as: panning and zooming; use of icons; hovering over data points; insights and highlights; use of colour to mark zones etc. Each of these contributed to the final look and feel of the data dashboard prototype, shown in Figure 6, which consists of the following components:

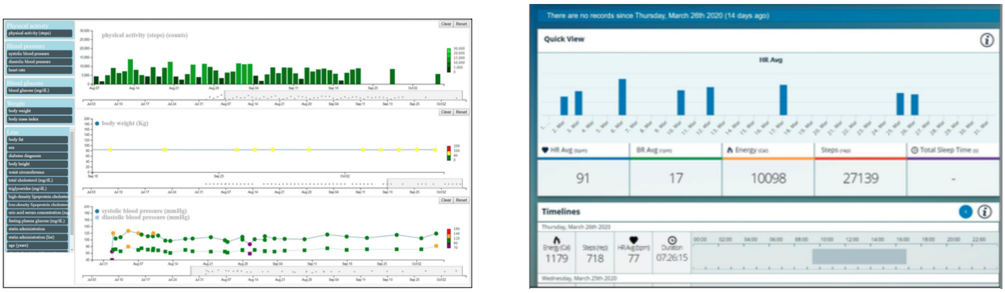
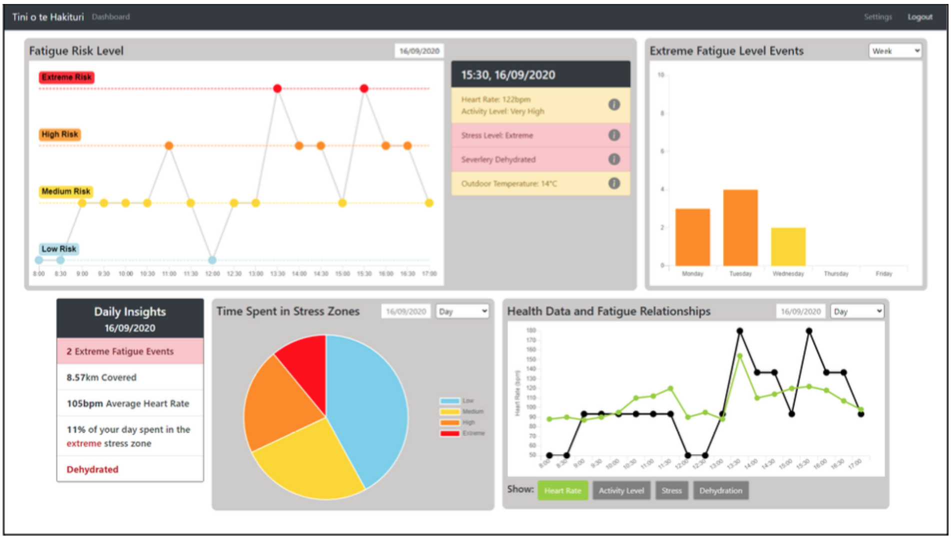

physical activity (steps) (counts)
body weight (Kg)
systolic blood pressure (mmHg)
diastolic blood pressure (mmHg)
There are no records since Thursday, March 26th 2020 (14 days ago)
Quick View
HR Avg
HR Avg
BR Avg
Energy
Steps
Total Sleep Time
91
17
10098
27139
-
Timelines
Thursday, March 26th 2020
1179
718
77
07:26:15
Wednesday, March 25th 2020

Fig. 5. CARRE Patient Data Visualisation [21] (left) and Hexoskin Data Visualisation (right)

.



15:30, 16/09/2020

Heart Rate: 122bpm
Activity Level: Very High

Stress Level: Extreme

Severlery Dehydrated

Outdoor Temperature: 14°C

Daily Insights
16/09/2020

2 Extreme Fatigue Events

8.57km Covered

105bpm Average Heart Rate

11% of your day spent in the extreme stress zone

Dehydrated

Fig. 6. Single User Data Dashboard Prototype for Forestry Data Visualisation

- The **Fatigue Risk Level Graph** shows a daily snapshot of the overall fatigue risk levels throughout a day (top left corner of Figure 6). The day is split into 30-minute intervals, with the overall level of fatigue calculated using a physiological, environmental, behavioural and demographic data and ranked as low, medium, high or extreme risk. Users can select a specific day to view using a date picker or by entering a specific date. This data is based on the real-time data analysis and alerts that are calculated and reported to the worker in the workplace, so the dashboard provides a summary of these. One of the benefits of this is, the worker may not have received any alerts all day, but on reviewing the graph they may see that they were borderline for high risk at several times and can therefore modify or change behaviours in the following days to reduce their risk. There are multiple ways of interacting with this graph. There is a tooltip that is displayed when a user hovers over, or near a point on the graph. This displays the time and the overall fatigue risk level. which makes it easier for a user to identify times of the points that fall under the extreme risk level. All of the points on the graph are also clickable to show more information about the worker's physiological and behavioural data, in addition to environmental data in an information panel (top centre of Figure 6.

- **Extreme Fatigue Risk Events Graph** which is a bar graph displaying the number of times a worker has been at the extreme fatigue risk level (top right corner of Figure 6). This can be shown as a weekly graph (one bar per day), monthly or yearly. This allows workers to identify trends in their data over time. Hovering over one of the bars displays the day/date and number of fatigue events for that period.
- **Daily Insights Panel** is an information panel which gives an overview of a user's day (bottom left of Figure 6). This provides a summary of insights about a day at a glance, without the need to refer to the graphs.
- **Stress Zones Graph** is a pie chart showing percentage of a selected day (or week, month, etc.) that a worker has spent in each of the low, medium, high and extreme stress zones (bottom centre of Figure 6). This provides the worker with a clear visual representation how much of their day (or selected time period) is spent in the different stress zones. Hovering over one of the sections will display the specific percentage of the day spent in that stress zone.
- **Health Data and Fatigue Relationships Graph** allow users to identify how different data is related to their overall level of fatigue (bottom right of Figure 6). A line showing the overall fatigue risk level throughout the day is permanently displayed on the graph, while other metrics can be changed by selecting one of the buttons beneath the graph. By viewing the different data in addition to the overall fatigue risk level, users will be able to recognise patterns which may lead to behaviour change. Similar to the other graphs the user can select a day, week, month or yearly period to visualise. Hovering over the points of the graph displays the data for both datasets of the graph.

Following the development of the initial dashboard, an iterative process of user studies and refinement were conducted. These followed a task-based approach where six participants (non forestry workers) found answers to a series of questions by interacting with different parts of the dashboard. For example, "What stress zone were you in at 3.30pm today? Is this reading good, or bad? Why do you think this?". To answer this, participants needed to correctly navigate to the stress zone graph and select the right time of day to find the answers. These initial user studies were intended to identify any major issues before development on the rest of the system continued.

2.3 Challenges of data ownership and governance

The next stage of development focused on users sharing data with whānau and family members. This initially involved looking at privacy settings to support the workers in controlling the level of detail that could be shared. Again, we looked to several existing health applications to understand best practice for controlling shared data in the health space, for example Apple Fitness, Garmin Health and similar. We also considered the importance of making explicit the results of sharing settings so that the workers could easily see the effect of their choices. A good example of providing this type of feedback (through a feed-forward mechanism) is described by Coppers et al. [4]. Although their work is looking at UI controls more generally, the methods they describe, which allow users to see the effects of widgets before selecting them, is also a good option for the types

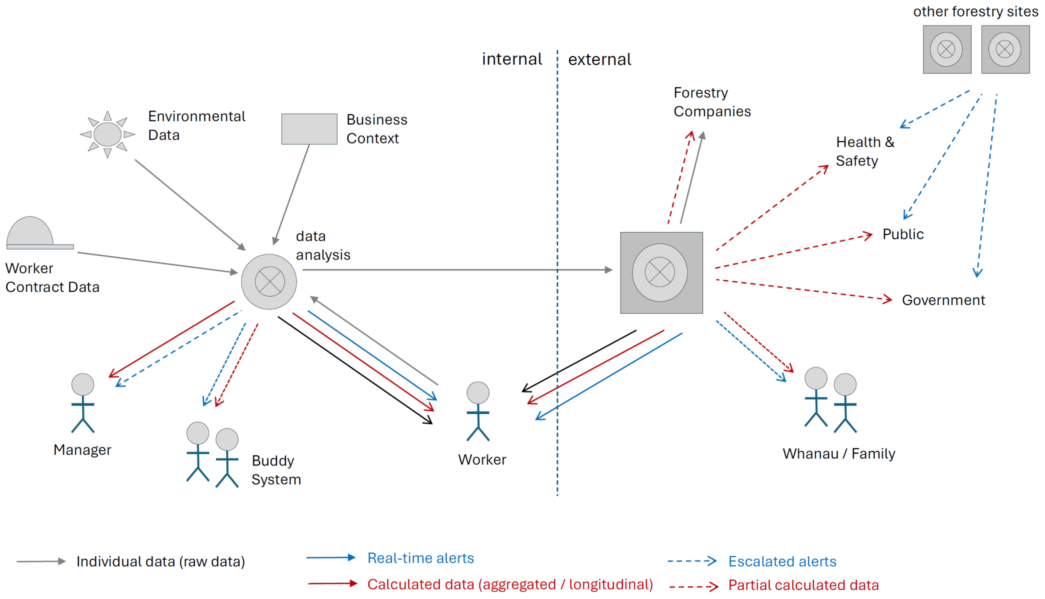


Fig. 7. Dataflow within forestry project: internal and external stakeholders

of privacy and sharing controls we need. Further considerations of how best to manage this data sharing whilst also considering the wider context of sharing parts of the data with external parties, led to further considerations of the data ownership and governance of all of the data. This led to the identification of the following challenges:

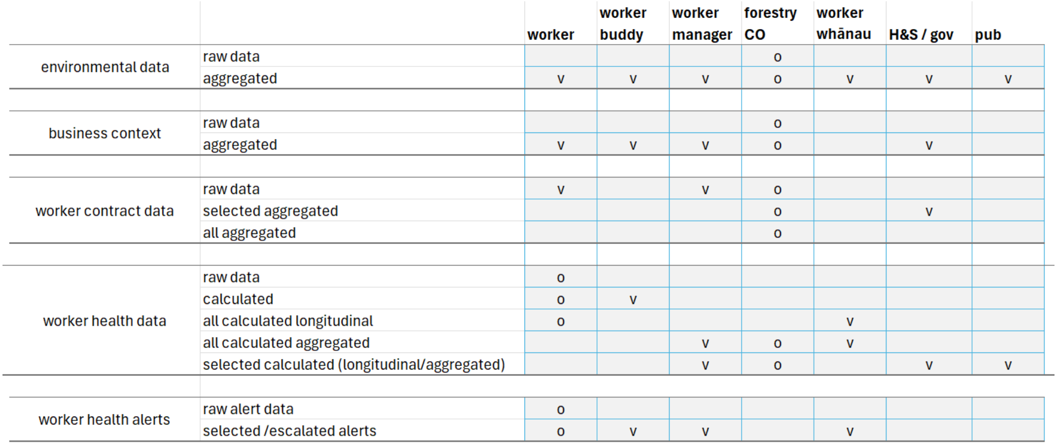
(1) stakeholders have different needs
(2) personal data requires privacy and sharing support to be built in
(3) combined data comes from multiple sources, each with different ownership
(4) users with limited technical skills or interest can't reply on complex permission settings
(5) privacy requirements and use cases change over time
(6) workers move between sites / employers but need to retain view of combined historical data
(7) levels of anonymity require careful handling and differ between groups

Our primary focus in this work addresses Challenges 2, 3, 6 and 7 as these relate to how data is aggregated, anonymised and shared. In the next section we examine the different stakeholders and their relationships to the data in more detail.

3 Stakeholders and Data Ownership

The data in the forestry project consists of both collected data (i.e., data that is measured from workers and their environment) and data generated by the analysis on the collected data (e.g. a fatigue identification for a worker). For data sovereignty, we consider both types of data in the same way, i.e., rights on both collected and calculated data need to be protected. Calculated data may further be aggregated (e.g., across teams) or longitudinal (e.g., considered over a time span).

Data and internal stakeholders. Figure 7 shows the detailed dataflow within the forestry project. The left side shows the dataflow at the worksite: collection of data from the worker, worker contract data (such as information about their role), environmental data and the business context. This real-time data is analysed to identify fatigue events, and send warnings to the worker, their buddy



		worker	worker buddy	worker manager	forestry CO	worker whānau	H&S / gov	pub
environmental data	raw data				o			
	aggregated	v	v	v	o	v	v	v
business context	raw data				o			
	aggregated	v	v	v	o		v	
worker contract data	raw data	v		v	o			
	selected aggregated				o		v	
	all aggregated				o			
worker health data	raw data	o						
	calculated	o	v					
	all calculated longitudinal	o				v		
	all calculated aggregated			v	o	v		
	selected calculated (longitudinal/aggregated)			v	o		v	v
worker health alerts	raw alert data	o						
	selected /escalated alerts	o	v	v		v		

Fig. 8. Matrix of data ownership and control (o = ownership, v = visible data): raw data refers to the sensor readings, calculated data refers to the outcome of data analysis (e.g. fatigue status), aggregated data is across teams, and longitudinal data across time

(via the Buddy System), and the manager. While the worker has access to all data, their buddy will receive relevant alerts and only selected aggregated data (i.e., not including detailed health information). The manager in turn will receive alerts escalated from the buddy system (fewer than the buddy system, as indicated by wider dashes), and aggregated data at the end of each day showing fatigue data across all of the work team.

Data and external stakeholders. Data from the workers in each team is aggregated further, and made available to external stakeholders at the end of the work day or throughout the following days and weeks (see right side of Figure 7). Immediate relations (e.g., family, whānau) may receive aggregated data and relevant alerts, while the community may be provided with selected aggregated data via a dashboard to support the workers' health. The work should retain guardianship and access rights over their personal health data both in detailed and aggregated form, as well as any alerts received over time.

The forestry companies covering this worksite may receive aggregated data relating to their sides. Furthermore, aggregated data from across forestry sites may be provided to the Health & Safety team, the government, and the general public.

Data ownership and control. Each of these shared data contains potentially elements of a worker's health data that are under their personal governance and guardianship, as well as information about the business context (e.g. worksite locations and conditions) and managerial decisions that are under the company's governance and guardianship. *Data governance* refers to the right to determine the use of one's own data (i.e., relating to ownership), while *data guardianship* refers to the care of data that originates from others (also referred to as stewardship). The matrix shown in Figure 8 gives an overview of the data ownership, visibility, and use as outlined above. The data list on the left refers to the internal data sources shown in Figure 7 (left), while the data recipients across the top refer to the main players shown Figure 7.

We identified seven challenges that need to be addressed for data ownership and governance (see Section 2.3. Four of the challenges relate to sharing of data that is calculated and considered in aggregated/longitudinal contexts. The matrix shown in Figure 8 reflects these challenges: Challenge 2 (personal data privacy and sharing) is shown here for the worker health data that is to be owned by each of the workers (both raw sensor data and calculated fatigue data) while also sharing necessary details with the buddy system. The related fatigue alert data similarly belongs to the respective worker, while alerts are to be shared via the buddy system, and escalated to the manager where needed. Challenge 3 related to combining data that may have different ownerships, such as environmental data combined with worker health data to allow for analysis of the impact of environmental context on the health of the working teams. Challenge 6 refers to the retention of worker data ownership and access while moving between different teams or companies. Challenge 7 highlights the management of data governance and guardianship while providing necessary access to meaningful calculated/aggregated data. For example, selected information about worker health issues in forestry teams is to be shared with the general public, while ensuring protection of the individual workers.

4 Related Work

We here first discuss research that addresses related challenges to those illustrated above (Section 4.1). Section 5 then introduces our approach to federated data sharing, which draws on both the concepts of federated data management (described in Section 4.2) and federated learning (Section 4.3).

4.1 Privacy Settings

Within the context of the forestry project there are two ways in which data can be shared. The first is determined by the architecture and requirements of the system (described in Section 3) while the second relates to a more typical use of privacy settings where a worker can determine the level of detail that a family member has access to, for example. This relates to Challenge 4 from Section 2.1.

Bokove et al. [1] investigated user-centric approaches for protecting privacy of users in applications which collect sensor data for health and well-being, where the type of personal data they consider is similar to that of the forestry workers. They identified that users should be able to control which parts of their data should be considered sensitive and who they might share it with and under what conditions. They note that ensuring privacy settings and controls are intuitive and easy to understand is crucial to ensuring users remain in control and that users should be able to easily change their mind and revoke consent for others to use their data at any time. This dynamic nature of consent and privacy is reflected in the needs of the forestry solution. Workers may change employer, which might mean moving to a different contracting company or moving to a different forestry site. Privacy settings and controls they have in place over their data use should move with them and not be dependent on the new work context.

According to Nasah et al. [18] there is a strong correlation between a user's age, gender and socioeconomic status (relating to educational levels) and their tendency to use technology, with socioeconomic status, rather than age, being the dominant factor. This impacts not only their digital literacy, but also understanding of, and familiarity with, privacy settings. Similarly, Becker found that even though younger users were typically considered to be more 'tech savvy', older users tend to have higher cognitive skills to evaluate new technologies compared to younger users. Given the demographics of the forestry workers, who typically come from lower socio-economic backgrounds and have lower educational outcomes, this is of particular relevance. According to Torre, users often do not "really read" permissions statements, nor do they understand what the permissions mean even when they do [26]. Permissions that are vague, confusing and poorly grouped are some of the common reasons for users' lack of understanding in privacy settings. In addition, Watson et al. identified that users do not change or modify default settings, nor do they use granular settings to avoid what might possibly be tedious, difficult and time consuming work [27].

Lin et al. conducted a study that identifed that a user's willingness to provide personal information for a specific application relies strongly on how much the user trusts that application [16]. Their work suggests that willingness to provide information is also likely to increase if users are provided controls over how much is shared and who can see their data. However, allowing users to set their privacy preferences for every single piece of data is likely to get tedious for the user who will eventually lose interest in setting and updating their privacy preferences.

While the use of privacy settings provides users with some autonomy over their data and how it is shared, the difficulties in ensuring fine-grained control and updating settings regularly suggests that we can't just rely on users to manage data privacy themselves. Rather, we need solutions which integrate ownership and control at the point the data is collected.

4.2 Federated Data Management

Federated data management is an approach that allows the integration and management of data from multiple, distributed sources without the use of a centralised data repository. This enables different organisations to combine data from various sources or systems while maintaining the autonomy and security of each individual data source [6]. Each part of the data retains its own governance, access controls, and update mechanisms. The data is combined into a virtual dataset which can be queried as if it were a single dataset.

4.3 Federated Learning

Federated learning is a machine learning technique which is related to the concept of federated data management. Data from multiple clients is used to collaboratively train a model without data being shared between clients or individuals losing access to their own data. Depending on the approach used, the model is then provided to each client for local use, or is centralised to provide individual data results to each client. In both cases the raw data from each client is not shared with any other party. The concept of privacy for federated learning is twofold. The ability to collaborate on training machine learning models, without exposing raw data to other parties, provides data privacy and ownership to each client. However, what is often also considered is security of data transfer to ensure it is kept private from external parties, this is typically managed using encryption mechanisms.

Patros et al. [20] demonstrate how federated learning can be used in situations where indigenous data sovereignty is a requirement of the proposed solution. Their work focusses on rural primary industries in Aotearoa/New Zealand (for example in agriculture) where IoT is used to gather real-time data and AI provides insights and analysis to support decision making. With Māori land-owners identified as one of the stakeholders for their work, they propose federated learning as a method which not only meets technical requirements for edge computing on low-resourced devices, but also supports data ownership and indigenous data sovereignty (IDS) for individual land-owners.

While federated learning is now well accepted as a privacy-preserving machine learning approach, there are still many technical requirements that must be met in order for the approach to be effective, efficient and genuinely privacy-preserving. In 2023, Li et al. [15] conducted a survey on federated learning to investigate the different architectures, privacy mechanisms and effectiveness. Among their findings were the importance of heterogeneity, which is required for ease of integration of the various clients into a single learning system, as well as autonomy. Ooi et al. [19] similarly discuss the opportunities and challenges of a hierarchical approach to federated learning in sensor applications, and proposed the use of edge computing to enable data management at the local (rather than cloud) level.

5 Federated Data Sharing

Building on the concepts of federated data management and federated learning, we developed a new data sharing concept which we call federated data sharing. This allows data to be aggregated into a single data set (as in federated data management) and combined to produce aggregated results (as in federated learning) whilst maintaining individual ownership and privacy. Unlike federated data management, the collective data is not visible and accessible to all entities, but rather the approach enables control over what is shared, and to whom, based on data policies for each data stream. Similar to federated learning, the data can be combined to produce aggregated data sets for visualisation, statistical analysis etc. with individuals retaining the right to ownership and management of their personal data. Based on the problems with supporting users in controlling privacy themselves, as outlined above, and building on the example of federated learning as a privacy-preserving method we now describe our proposed approach of federated data sharing in more detail.

Figure 9 shows a comparison of federated data sharing with federated learning. The top image shows an example of federated learning in rural AI, taken from [20]. Sensors on each farm site send data to a local gateway where they are aggregated and then shared to an edge server for regional model aggregation. Finally the regional models are aggregated and trained at the cloud layer. The bottom image shows how our federated data sharing example follows this model. Each individual worker and environmental sensor can be seen as analogous to the local layer sensors, but each retains autonomy to their individual data which is first aggregated across teams (for buddy sharing and real-time alerting) before being sent to the edge computing. In this example the edge computing occurs at a site level and aggregates multiple work teams from that site. Finally the data is aggregated at the cloud level, which may be regional forestry sites, or Nationally across all sites etc.

To consider how this works for the forestry data visualisation example, we have annotated Figure 3 to show how each part is represented by the federated layers, see Figure 10 (left). In order to now generalise this approach across different IoT solutions we can similarly annotate the shared data visualisation of Figure 2 with the federated layer, see Figure 10 (right).

Within the local layer, each data owner (denoted by the coloured bar in right hand image) always retains full ownership and access to their raw data. At the edge layer this may be aggregated or analysed to produce new data. Visibility of aggregated data is set according to the system requirements, while data derived from analysis is restricted to the source of the analysed data. If we return again to the forestry example, a worker's data that is collected during a particular work period on a given site will always be available to them irrespective of future changes in employment. Similarly, derived data for that worker (fatigue events etc.) will also remain available to them.

6 Discussion and Conclusion

Our federated data sharing approach was developed in response to some of the challenges elicited from the data dashboard development for the forestry project. Specifically:

2: stakeholders have different needs
3: personal data requires privacy and sharing support to be built in
6: privacy requirements and use cases change over time
7: workers move between sites / employers but need to retain view of combined historical data

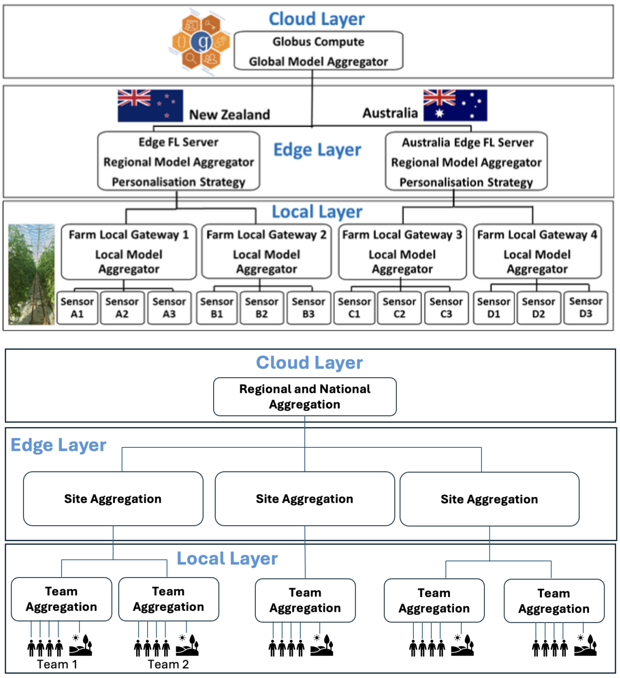


Fig. 9. Comparison of Federated Learning with Federated Data Sharing. Top image from [20]

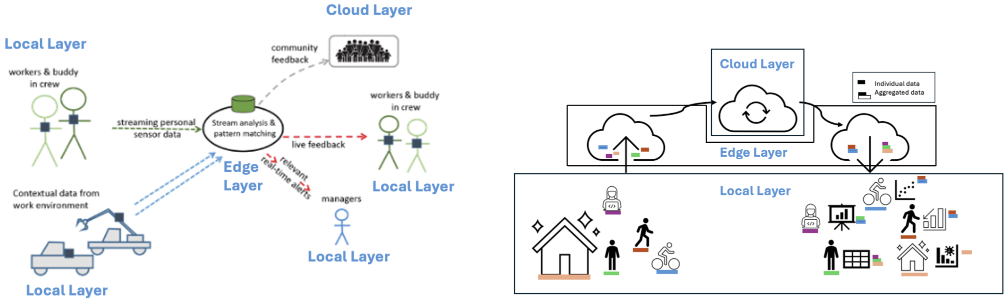


Fig. 10. Data flow for fatigue monitoring (left) and Generalised Data Aggregation Visualisations (right), both annotated with federated layers

Different needs of stakeholders are addressed by ensuring they can always access their own data and any aggregations which their data contribute to. This also means that privacy and sharing are, by default, preserved. Requirements on users to manage sharing permissions are minimised to cases where they want to share their personal or aggregated data to parties who do not otherwise have access (e.g. the whānau/family groups). If privacy requirements change, for example if a new Government agency requires access to forestry data nationwide, this can be addressed at the aggregated data level which does not affect individual data ownership. Similarly, if workers change employer, or even leave the forestry industry their rights to access their personal data is ensured.

While these principles satisfy the challenges identified, the way in which they are implemented will differ between systems and use cases. Within the forestry project the data storage and access is designed to support the ownership requirements identified at the edge and local layers. This includes addressing one part of indigenous data sovereignty be ensuring all data is kept onshore in Aotearoa/New Zealand. One of the implications of this is that all hardware and wearable technology had to be built from scratch as off-the-shelf products did not adhere to these requirements. This enabled us to also control the 'cloud', meaning that all data was contained within our systems giving us full control. We are not yet at the stage of integrating the data produced with external systems, for example government databases. This will require a more general approach which allows tracking and control of data in a wider context of use. This is also true for any solutions which do not have full control over where data is stored and used and has implications for implementing federated data sharing more generally. Further work is required to address this. We have been investigating an approach using meta-data to tag data akin to the use of traditional knowledge labels (TK Labels) for indigenous data[3] but this work is in its early stages at present.

We introduced the concept of federated data sharing as a mechanism for preserving ownership and governance of personal data collected within IoT solutions. Using a real-world example of wearable technology developed for the forestry industry in Aotearoa/New Zealand we have outlined how and where data aggregation and data visualisation requirements arise and how these can be addressed by our federated data sharing principles. The success of this approach relies on how it is implemented, particularly when IoT systems contain off-the-shelf products which come with their own built-in data controls. As discussed above, the first step in generalising the solution is to find ways of retaining information about different data streams which are used within off-the-shelf systems. A harder problem is addressing how (and if) commercial vendors would support such an approach and provide the appropriate mechanisms within their own software solutions.

Another consideration is that of autonomy, which is a central principle of federated data learning. Autonomy allows for clients to disconnect and disengage, which in turn has implications for the collaborative model training. Similarly, if we envisage a federated approach for data learning which extends beyond a single system, this autonomy may have major implications. In the simplest case, aggregated data may become less meaningful if the number of contributors reduce - either due to choice, or technical connectivity issues. In the worst case, anonymity of participants may be compromised if the number of contributors drops below a certain level. In future work, we need to consider how autonomy can be managed to ensure these issues do not occur.

References

[1] Wouter Bokhove, Bob Hulsebosch, Bas Van Schoonhoven, Maya Sappelli, and Kees Wouters. 2012. User Privacy in Applications for Well-being and Well-working. In *Proceedings of AMBIENT 2012, The Second International Conference on Ambient Computing, Applications, Services and Technologies* (Barcelona, Spain). 53–59.

[2] Judy Bowen and Annika Hinze. 2022. Participatory Data Design: Managing Data Sovereignty in IoT Solutions. *Interacting with Computers* 34, 2 (2022), 60–71. https://doi.org/10.1093/iwc/iwac031

[3] Judy Bowen, Helen Petrie, Annika Hinze, and Sanjit Samaddar. 2020. Personas revisited: Extending the Use of Personas to Enhance Participatory Design *(NordiCHI '20)*. Association for Computing Machinery, New York, NY, USA, Article 62, 12 pages. https://doi.org/10.1145/3419249.3420135

[4] Sven Coppers, Kris Luyten, Davy Vanacken, David Navarre, Philippe A. Palanque, and Christine Gris. 2019. Fortunettes: Feedforward about the Future State of GUI Widgets. *Proc. ACM Hum. Comput. Interact.* 3, EICS (2019), 20:1–20:20.

[5] Shamal Faily and Ivan Flechais. 2011. Persona cases: a technique for grounding personas. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (, Vancouver, BC, Canada,) *(CHI '11)*. Association for Computing Machinery, New York, NY, USA, 2267–2270. https://doi.org/10.1145/1978942.1979274

[6] Olaf Görlitz and Steffen Staab. 2011. *Federated Data Management and Query Optimization for Linked Open Data.* Springer Berlin Heidelberg, Berlin, Heidelberg, 109–137.

[3] https://localcontexts.org/labels/traditional-knowledge-labels/

[7] Christopher Griffiths, Judy Bowen, and Annika Hinze. 2017. Investigating Wearable Technology for Fatigue Identification in the Workplace. In *Human-Computer Interaction - INTERACT 2017 - 16th IFIP TC 13 International Conference, Mumbai, India, September 25-29, 2017, Proceedings, Part II (Lecture Notes in Computer Science, Vol. 10514)*, Regina Bernhaupt, Girish Dalvi, Anirudha Joshi, Devanuj K. Balkrishan, Jacki O'Neill, and Marco Winckler (Eds.). Springer, 370–380.

[8] Jonathan T. Grudin and John S. Pruitt. 2002. Personas, Participatory Design and Product Development: An Infrastructure for Engagement. https://api.semanticscholar.org/CorpusID:8238498

[9] A. Hinze, J. Bowen, and J.L. König. 2022. Wearable technology for hazardous remote environments: Smart shirt and Rugged IoT network for forestry worker health. *Smart Health* 23 (2022), 100225. https://doi.org/10.1016/j.smhl.2021.100225

[10] Annika Hinze, Jemma L König, and Judy Bowen. 2021. Worker-fatigue contributing to workplace incidents in New Zealand Forestry. *Journal of Safety Research* 79 (2021), 304–320. https://doi.org/10.1016/j.jsr.2021.09.012

[11] Robert Hockey. 2013. *The Psychology of Fatigue: Work, Effort, and Control.* Cambridge University Press.

[12] Soon-Gyo Jung Joni Salminen, Kathleen Guan and Bernard J. Jansen. 2021. A Survey of 15 Years of Data-Driven Persona Development. *International Journal of Human–Computer Interaction* 37, 18 (2021), 1685–1708. https://doi.org/10.1080/10447318.2021.1908670

[13] Tahu Kukutai and John Taylor. 2016. *Data Sovereignty for Indigenous People: Current Practice and Future Needs.* Australian National University Press, 1–24.

[14] Jemma L. König, Judy Bowen, Annika Hinze, and Dylan Exton. 2024. IoT in forestry: Human-focused assistive safety technology. *Safety Science* 176 (2024), 106525. https://doi.org/10.1016/j.ssci.2024.106525

[15] Qinbin Li, Zeyi Wen, Zhaomin Wu, Sixu Hu, Naibo Wang, Yuan Li, Xu Liu, and Bingsheng He. 2023. A Survey on Federated Learning Systems: Vision, Hype and Reality for Data Privacy and Protection. *IEEE Transactions on Knowledge and Data Engineering* 35, 4 (2023), 3347–3366.

[16] Jialiu Lin, Bin Liu, Norman Sadeh, and Jason I. Hong. 2014. Modeling users' mobile app privacy preferences: restoring usability in a sea of permission settings. In *Proceedings of the Tenth USENIX Conference on Usable Privacy and Security* (Menlo Park, CA) *(SOUPS '14)*. USENIX Association, USA, 199–212.

[17] Jennifer (Jen) McGinn and Nalini Kotamraju. 2008. Data-driven persona development. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Florence, Italy) *(CHI '08)*. Association for Computing Machinery, New York, NY, USA, 1521–1524. https://doi.org/10.1145/1357054.1357292

[18] A. Nasah, C. DaCosta, B.and Kinsell, and S. Seok. 2010. The digital literacy debate: an investigation of digital propensity and information and communication technology. *Education Tech Research Dev* 58 (2010), 531–555.

[19] Melanie Po-Leen Ooi, Shaleeza Sohail, Victoria Guiying Huang, Nathaniel Hudson, Matt Baughman, Omer Rana, Annika Hinze, Kyle Chard, Ryan Chard, Ian Foster, et al. 2023. Measurement and applications: Exploring the challenges and opportunities of hierarchical federated learning in sensor applications. *IEEE Instrumentation & Measurement Magazine* 26, 9 (2023), 21–31.

[20] Panos Patros, Melanie Ooi, Victoria Huang, Michael Mayo, Chris Anderson, Stephen Burroughs, Matt Baughman, Osama Almurshed, Omer F. Rana, Ryan Chard, Kyle Chard, and Ian T. Foster. 2023. Rural AI: Serverless-Powered Federated Learning for Remote Applications. *IEEE Internet Comput.* 27, 2 (2023), 28–34.

[21] CARRE project. [n. d.]. Visualizing HealthLines in CARRE. available at: https://www.carre-project.eu/visualizing-healthlines-in-carre/.

[22] John Pruitt and Jonathan Grudin. 2003. Personas: practice and theory. In *Proceedings of the 2003 Conference on Designing for User Experiences* (San Francisco, California) *(DUX '03)*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/997078.997089

[23] Anna K Rolleston, Judy Bowen, Annika Hinze, Erina Korohina, and Rangi Matamua. 2021. Collaboration in research: weaving Kaupapa Māori and computer science. *AlterNative: An International Journal of Indigenous Peoples* 17, 4 (2021), 469–479. https://doi.org/10.1177/11771801211043164

[24] C Matthew Snipp. 2016. *What does data sovereignty imply: what does it look like?* Australian National University Press, 39–76.

[25] Te Mana Raraunga. [n. d.]. Te Mana Raraunga: Māori Data Sovereignty Network Charter. online at https://www.temanararaunga.maori.nz/s/Te-Mana-Raraunga-Charter-Final-Approved.pdf.

[26] I. Torre, O. Sanchez, F. Koceva, and G. Adorni. 2018. Supporting users to take informed decisions on privacy settings of personal devices. *Pers Ubiquit Comput* 22 (2018), 345–364.

[27] Jason Watson, Heathe Richter Lipford, and Andrew Besmer. 2015. Mapping User Preference to Privacy Default Settings. *ACM Trans. Comput.-Hum. Interact.* 22, 6, Article 32 (nov 2015), 20 pages.

[28] Forestry New Zealand. 2019. Forestry health and safety. New Zealand Ministry for Primary Industries online information on Forestry NZ. available from https://www.mpi.govt.nz/growing-and-harvesting/forestry/taking-care-of-your-forest/forestry-health-and-safety.