Data & Policy (2024), 6: e26
doi:10.1017/dap.2024.18

RESEARCH ARTICLE

Open data, workplace relations law compliance, and digital regulation

Colleen Chen[1], John Howe[2], Timothy Kariotis[1] and Shirley Jackson[1]

[1]Melbourne Regulation & Design Network, University of Melbourne, Melbourne, Australia
[2]Melbourne Regulation & Design Network and Centre for Employment and Labour Relations Law, University of Melbourne, Melbourne, Australia
Corresponding author: John Howe; Email: j.howe@unimelb.edu.au

Received: 25 August 2023; **Revised:** 05 February 2024; **Accepted:** 17 March 2024

Keywords: compliance; detection; open data; regulation; workplace relations

Abstract

This paper discusses the challenges and opportunities in accessing data to improve workplace relations law enforcement, with reference to minimum employment standards such as wages and working hours regulation. Our paper highlights some innovative examples of government and trade union efforts to collect and use data to improve the detection of noncompliance. These examples reveal the potential of data science as a compliance tool but also suggest the importance of realizing a data ecosystem that is capable of being utilized by machine learning applications. The effectiveness of using data and data science tools to improve workplace law enforcement is impacted by the ability of regulatory actors to access useful data they do not collect or hold themselves. Under "open data" principles, government data is increasingly made available to the public so that it can be combined with nongovernment data to generate value. Through mapping and analysis of the Australian workplace relations data ecosystem, we show that data availability relevant to workplace law compliance falls well short of open data principles. However, we argue that with the right protocols in place, improved data collection and sharing will assist regulatory actors in the effective enforcement of workplace laws.

Policy Significance Statement

Our analysis suggests that the detection and prevention of "wage theft," and other forms of noncompliance with workplace laws, could be greatly improved using targeted data science interventions. Our study highlights one of the key challenges in achieving this goal, namely the availability of adequate datasets that will allow predictive models to be developed. Although the currently available datasets in the Australian workplace relations context are inadequate for the tasks described, we believe that there is room for optimism in the recommendations we have made to improve the integration of data science with workplace relations law enforcement in the future.

1. Introduction

Government data has long played a role in informing public policy, spurring both public and private innovation, and supporting greater citizen engagement (Janssen et al., 2012; Chan, 2013). However, recent advancements in data collection, linkage, and analysis have greatly enhanced the opportunities to use data for policy development and implementation (Höchtl et al., 2016). In this paper, we examine the availability of government-held workplace relations data and explore its utility as a means of improving the policy and regulatory outcomes for those who are dependent on their labor for a living.

The use of government data to inform workplace relations policy and regulation remains relatively underexplored in the academic and applied fields (cf. McCann and Cruz-Santiago, 2022), although a number of related issues have received attention, such as the collection and (mis)use of worker data by private firms (Ebert et al., 2021; Bodie, 2022; Rogers, 2023). There is growing evidence that government agencies, including workplace relations tribunals and labor inspectorates, are seeking to make better use of government data in order to improve the efficiency and effectiveness of regulation (Organisation for Economic Co-operation and Development [OECD], 2021; Hannah, 2022; Fair Work Commission [FWC], 2023).

It is important to recognize that policy design, development, and regulation are not solely the domain of government, as open data allows myriad nongovernmental actors to engage in policy debates. This information is critical, as it not only allows external organizations to assess the efficacy of policy and regulation, but helps identify unintended and undesirable consequences (Coglianese et al., 2004). In the context of workplace relations, the task of monitoring and enforcing regulation is frequently carried out by individuals and civil society actors, notably trade unions, augmenting the actions of relevant government agencies (Amengual and Fine, 2017). Currently, although workplace relations regulators are increasingly utilizing government data, there is limited evidence of this data being made available to the public for individual and civil society monitoring and enforcement activities.

This paper articulates some of the challenges and opportunities presented by this unique integration of data science and workplace relations, with particular reference to minimum employment standards such as wages and working hours regulation. The research was informed by a project conducted at the University of Melbourne (the “Fair Day’s Work” project) in which the authors examined how data science could contribute to the improvement of compliance within Australian workplace regulations (Howe and Kariotis, 2021). Based on previous research, which showed that individuals and businesses were more likely to comply with regulation when they perceived a higher risk of noncompliance being detected (Azarias et al., 2014; Hardy, 2021), the project sought to develop a model for predicting individuals and businesses at high risk of noncompliance with labor laws, operating on the assumption that an increased perception of detection by the regulator will alter behavior.

This paper outlines the findings of this project, augmented with some global case studies of the use of machine learning in workplace regulation enforcement and a thorough analysis of relevant workplace and worker data available in the Australian context with reference to the international “open data” principles outlined below. Since accurate, publicly available information relating to compliance with minimum employment standards is crucial to the effective enforcement of those laws, our findings on the issues of the availability, accessibility, and quality of worker and workplace relations data are relevant to policy and practice across a range of regional and domestic economies.

2. Open data policy

While data collected by private and nonprofit institutions has commonly been used to supplement government data (Reggi and Dawes, 2022), government data has only recently become more available to the public. There are also increasing efforts to link data across government with nongovernmental datasets to generate new insights and value (Ubaldi, 2013). These trends culminated in the Open Government Partnership (OGP), established in 2011 to formalize a multilateral commitment towards government data openness, and now covers 75 member states who represent over two billion people (Harrison and Syogo, 2014; Park and Kim, 2022). OGP endeavours to bring government and civil society stakeholders together to develop and implement action plans and policy reforms to protect and enable the principles of open data (GovLab).

Since becoming an OGP member in 2015, the Australian federal government has implemented two National Action Plans to drive the OGP agenda, respectively targeting digital transformation and the use of artificial intelligence (AI) to improve policy outcomes (Attorney General’s Department, 2023). As a

https://doi.org/10.1017/dap.2024.18 Published online by Cambridge University Press

result, Australia is now ranked 6th in the OECD according to the OURdata Index, scoring above the OECD average in data availability, data accessibility, and government support for data reuse (OECD, 2019). However, the achievements of open government data in Australia have not been evenly distributed across government portfolios, and there is a growing critique regarding the quality of government data and its adequacy for regulatory purposes (Safarov et al., 2017).

Some government portfolios oversee sectors with more mature data practices due to a longer history of digitalization and datafication (O'Leary et al., 2021), notably financial services, utilities, and geospatial services (OECD, 2019). Other sectors, such as social services, workplace relations, and legal services, do not share the same maturity in their data practices. They have not undergone the same levels of digital transformation as the more mature areas, as evidenced by data tending to be collected in an unstructured format such as surveys, intake forms, and reports. As a result, data quality, accessibility, and availability lag its counterparts and represent a considerable barrier to utilizing data science in enforcement and regulatory compliance (Frangi et al., 2021).

The Australian government has recognized that open data can contribute to deepening governmental transparency and encouraging greater public engagement with government services or policy, and enabling the private sector to create new products or services that generate both private and public value (Reichert, 2017). As noted above, Australia is an OGP member, which has encouraged federal and state governments to provide accessible data collected by all three levels of government. Critically, these platforms reflect the government's intent to "make nonsensitive data open by default…[and] to stimulate innovation and enable economic outcomes" (Department of Prime Minister and Cabinet, 2015), which is ably supported by a range of legislation and policy (Data Availability and Transparency Act 2022 (Cth); Government Information (Public Access) Act 2009 (NSW); DataVic Access Policy 2016 (Vic); Public Sector (Data Sharing) Act 2016 (SA); Queensland Open Data Policy (2022–2024); Western Australia Whole of Government Open Data Policy (2022); Right to Information Act, 2009 (Tas)).

However, even though the quantity of open government datasets in Australia has increased exponentially in the last decade, this has not been matched by a focus on the quality and usability of this data (Sadiq and Indulska, 2017). Australia's progress on open data has been described as "patchy" and "transitional" by its regulator, the Australian Information Commissioner (Heaton, 2016), and there remains a significant gap in understanding between the current state of data management in many public institutions, and the data management practices required to achieve a transformative open data framework (Martin et al., 2015).

Since most government data has been collected for administrative purposes, datasets are often stored in siloed databases within government departments and are not structured with interoperability in mind. Inconsistent storage formats and identification keys limit the utility of these datasets for anything entailing data linkage, reuse, or enhancement. This limitation also impedes the ability of governments to use data for machine-learning approaches when monitoring compliance and anticipating misconduct.

Therefore, while there has been a significant effort by governments worldwide to adopt an open data approach, few have translated this goodwill into high-quality, accessible open data portals, nor have many taken the opportunity to facilitate and promote widespread public use of data (Martin et al., 2015). As a result, policies on open data and real-world projects that could maximize the benefits of this data rarely meet.

3. Navigating data access in the Australian workplace relations context

In Australia, workplace relations policy is mainly the province of the federal government, although regional state governments hold some related policy functions such as workplace health and safety. The Fair Work Act 2009 (Cth) (FW Act) is the key legislation that establishes the Australian workplace relations framework, which includes two statutory bodies to oversee that framework: The FWC and the Fair Work Ombudsman (FWO). The FWC acts as a tribunal ruling on a variety of workplace issues, including the fairness of dismissal for certain employees. It also stewards the "award system" (a series of regulatory instruments that establish minimum employment entitlements across various industries and sectors) and oversees collective bargaining between employer and worker representatives at the enterprise

https://doi.org/10.1017/dap.2024.18 Published online by Cambridge University Press

level. The FWO is the national workplace relations regulator and has a role in promoting compliance with the FW Act, including the power to investigate noncompliance and bring court proceedings to enforce noncompliance.

One of the key challenges in Australia, as in many other jurisdictions around the world, is detecting business noncompliance with minimum employment standards set under the FW Act. Widespread and systematic noncompliance with minimum wages, or "wage theft," has emerged as a major policy problem in recent years (Hardy, 2021). This is notwithstanding the efforts of both trade unions and the FWO to monitor noncompliance and conduct enforcement activities.

While there are many issues that contribute to enforcement and compliance, both public and private regulatory actors currently face a capability gap relating to data collection, usage, and storage to adequately judge the scale of the problem of wage theft (Flanagan and Clibborn 2023, 339). Although the broader experience of industrial relations in Australia is in some ways unique amongst the advanced economies, this capability gap is universal. Proactive detection of noncompliance is critical to the protection of rights in any workplace relations context, especially in an environment where workers are often reluctant to make formal complaints or otherwise act against noncompliance (Weil and Pyles, 2006; Vosko and Closing the Enforcement Gap Research Group 2020; Hardy et al., 2023). Additionally, recent research has found that increasing the likelihood of detection can also act as a deterrent to future noncompliance (Hardy, 2021).

Improved public access to data about noncompliance may also assist with prevention by raising awareness among workers about the risks of exploitation. The potential for data science to act as an efficient mechanism for improving detection is an important consideration, given the limited resources of the FWO relative to its national responsibilities. This relative resource scarcity is a common challenge for unions and other civil society organizations seeking to protect the interests of vulnerable workers.

In the following section of this paper, we set out some case studies of real-world experiments in the use of data and machine learning for workplace relations monitoring and enforcement purposes. We then assess whether workplace relations data is available as open data in the Australian context and to what degree it aligns with accepted open data principles.

4. The use of machine learning in workplace regulation enforcement: global case studies

According to the International Labour Organization (ILO), data analytics and machine learning are increasingly being incorporated into proactive enforcement activities (ILO, 2020, 11). Machine learning (a subset of AI) is the science of teaching computers to learn and generalise using data, and helps to make processes more autonomous, efficient, and effective using training datasets. Today, new data collection, storage, transmission, visualization, and analytic techniques like machine learning have triggered a proliferation of datasets collected by public and private entities, covering diverse areas from health and wellness to consumer purchasing records. Such data is a powerful raw material for problem-solving, and the creation of specific tools capable of furthering the public interest could offer unique insights into how effectively we govern in a diverse array of policy areas. However, increasing advancements in AI, especially generative AI have raised ethical concerns due to their lack of transparency and explainability (Henman, 2020). There is also growing recognition of how biases within established datasets can be exponentially proliferated when fed into machine learning algorithms that allow for low-transparency but high-volume analysis.

By utilizing these novel data analytical techniques, governments and other actors could potentially target scarce enforcement resources more effectively toward compliance. The availability and adequacy of workplace relations data are critical if regulators are to use machine learning for enforcement. This is because data quality impacts the amount of effort required to transform raw information into a format that can be used for training and validating a machine-learning model (Gudivada et al., 2017). Not only does there need to be a sufficiently large volume of data, but the data needs to be collected in a manner that lends itself to enhancement and ease of validation.

https://doi.org/10.1017/dap.2024.18 Published online by Cambridge University Press

We have identified three case studies that demonstrate the range of different approaches that have been taken to the use of government data in the development of machine learning models specific to the context of workplace relations. These case studies provide insights into the opportunities and barriers associated with using open government data for the predictive enforcement of labor regulation. The examples include occupational health and safety (OH&S) monitoring, underpayment detection, and identifying human trafficking. Each of these examples is discussed in turn.

4.1. Case study 1: using data to improve occupation health and safety outcomes by identifying high-risk construction sites

Our first example is a recent study conducted by the OECD (2021), which explored the use of machine learning in the Lombardy region of Italy to identify building sites with a high risk of violating occupational health and safety (OHS) procedures. The researchers had access to historic OHS inspection records data in Lombardy, which they used to predict high-risk construction sites. However, even though the violation records were available, the sample proved insufficient for the development of an effective machine learning model, and the predictions that were produced held accuracy rates that were only marginally better at predicting risk than a coin flip. Although the model could not be used to target individual construction sites, the study contributed to an understanding of how data collection processes could be improved in future iterations of the project. For instance, data collection could be expanded from actual fatal injuries to include close-call incidents and/or complaints made by workers about their OHS conditions. Arguably, this could increase the adequacy of data collection, and not only help reduce the incidence of fatal injuries but improve the overall safety of the targeted industry. A similar study conducted in Norway, which also noted the significant challenges in targeting OHS inspections using big data and machine learning techniques, suggested that a combination of artificial and human intelligence was necessary to optimize the usefulness of such techniques (Dahl and Starren, 2019, 5):

> Rather than allowing the algorithm to pick and choose objects directly, the inspectors are allowed to make risk-informed decisions on the basis of the predictions that the algorithm makes. … When it comes to predictions of complex social events in general, combining the two types of intelligence is probably a necessity.

4.2. Case study 2: using data to predict underpayments among American businesses

The second case study of using data to predict underpayments demonstrates that even where large volumes of labeled data are available, efforts can still be hampered by the lack of related datasets, which are important for generating critical features that improve the accuracy of a model. In a 2018 study conducted by Stanford University and the Santa Clara University School of Law, researchers used 250,000 instances of Wage and Hour Compliance action data from the United States Department of Labor to predict underpayment (Johnson et al., 2018). The researchers had access to a large database of companies proven to have underpaid a worker between 2005 and 2017. However, they still were unable to produce a model that would allow them to deliver a robust prediction of companies most likely to offend.

This limitation was due to the fact that the researchers did not have access to additional descriptive information about the companies, such as their size, business processes, or other risk vectors capable of informing a well-trained model. Without the ability to adequately describe the level of risk identified in the compliance action dataset, it was impossible to train the model to search for indicative features. To offset the lack of descriptive characteristics of firms in the compliance dataset, the researchers used feature engineering techniques, a way of inferring additional information about an entity (Heaton, 2016). For example, one could infer a person's age by referring to their birth date. Feature engineering techniques allowed the researchers to extrapolate additional characteristics of noncompliant firms, such as size, industry context, whether or not the firm was a franchise, and geographic location by using the legal name and business types of the companies. However, the economic and demographic variables identified by the

https://doi.org/10.1017/dap.2024.18 Published online by Cambridge University Press

researchers were not on their own sufficient to produce insights with the requisite level of accuracy that would make them actionable by a regulator.

The inability to link existing compliance datasets to additional datasets is a common challenge when developing machine learning models using government datasets. Without the ability to experiment with generating new features from alternate sources, current government datasets limit the complexity of the models that can be developed.

Furthermore, the sample contained only companies that were caught infringing labor laws and did not include undetected noncompliant firms or compliant firms. Therefore, a sampling bias was inherent in the dataset. Sampling bias occurs when a particular group is so overrepresented in a training dataset that it misleads the model toward associating an unrepresentative trait with what it is trained to identify (Jeong et al., 2018). Severe examples of sampling bias include the overrepresentation of racial groups in current criminal databases used as a data source for training models that identify the likelihood of recidivism or reoffending (Yoon, 2018).

Since the underpayment model in the Johnson et al. (2018) study did not have the requisite information to distinguish the companies that were compliant from those that were noncompliant, this model would incorrectly assume that all are noncompliant. However, the model is unlikely to be as accurate in the real world. This problem is known as overfitting, in which a model is biased towards the existing training dataset, giving the illusion of high performance (Ying, 2019). For instance, if a model is only given pictures of white cats in its training dataset, it would become very good at identifying the white cats as cats, but it would likely reject cats of any other color, as it was trained to assume that all cats must be white. Similarly, given that the underpayment dataset only contained companies that were caught infringing workplace relations laws and did not show compliant companies that shared the same features as the noncompliant companies, if used in a real-world context, the model would start to identify all such companies as underpaying workers. This creates high rates of false positives in which noninfringing companies are wrongly identified by the model as being noncompliant. Additionally, the model is unlikely to be generalizable in the sense that if infringing companies did not share the same characteristics as the current set of infringing companies, then the model would not be able to accurately identify underpayment more broadly (Vosko and Closing the Enforcement Gap Research Group, 2020, 42–43).

4.3. Case study 3: using data to identify forced labor in the fisheries sector

To bypass data management challenges in public institutions, open-source intelligence approaches, where data is sought from a range of open sources, have also been explored by researchers to improve the quality of the training dataset. Our third case study is a seminal study by Global Fishing Watch, a nongovernment organization that sought to use satellite vessel monitoring approaches to identify forced labor on the high seas (McDonald et al., 2020). Data from the satellites enabled the researchers to systematically track over 16,000 fishing vessels at sea. Once this information was gathered, the researchers sought to assign risk scores to vessels based on existing literature about forced labor, such as vessel ownership, crew recruitment practices, and the catch of the species targeted on a vessel level. However, the researchers discovered that this information was not available. Proxy features such as vessel size, maximum distance from the port, the number of voyages per year, and average daily fishing power were instead generated to approximate the risk factors identified.

This approach emphasizes the importance of cross-disciplinary collaboration in developing machine learning models and highlights how existing literature can be leveraged to enrich model features, even when direct collection of requisite information from a single source is unlikely. However, in spite of overcoming data collection difficulties, the researchers acknowledged that there were numerous factors that ultimately rendered their model inadequate. Notably, the lack of a sufficiently large, labeled dataset of vessels known to use forced labor, coupled with an inability to verify the accuracy of the model through a comparison of known noncompliant vessels with the risk profile identified by the model, meant that it was unlikely such a model would be put into production and be used by regulators (Kroodsma et al., 2022).

https://doi.org/10.1017/dap.2024.18 Published online by Cambridge University Press

These three international case studies provide insight into the opportunities available to use government workplace relations data. However, they also highlight the challenges that arise when seeking to utilize this data through machine learning. A common theme to emerge from these examples was that the mere availability and transparency of data is not enough to train a machine learning model. To produce an effective model, the dataset must also be of a sufficiently large sample size, labeled for the condition the model is to predict, and capable of being linked to other government and nongovernment datasets. These properties require an intentional effort to realize a data ecosystem that is capable of being used to develop machine learning models. This explains why few machine learning models have been used for enforcement purposes despite the volume of open government data available in the public domain.

5. Mapping the Australian workplace relations data ecosystem

In this section, we apply the lessons learned from the above case studies to provide an overview and assessment of the current Australian workplace relations data ecosystem. Mapping this data ecosystem enables an assessment of the availability, accessibility, and quality of datasets in Australia. We identified and gathered known datasets in which data collection was mandatory, or those that fell within the purview of Australian regulators due to requirements under legislation. The resulting data sources that we focus on in this paper were either government datasets or government-funded datasets (e.g., academic datasets), which were consistent with an open data framework.

Further consultations were undertaken with academics, government officials, and relevant stakeholders who operated in or were engaged with the workplace relations system, to ensure completeness of coverage. Overall, 22 relevant datasets were identified and obtained through online search engines, relevant publications, and reports, and all requests were made directly to data holders (see Table 1).

The datasets were reviewed based on existing accompanying documentation, which provided relevant information on the dataset. When reviewing the datasets for inclusion in the current study, we considered ease of access, any limitations placed on the use of data, and the type and currency of the data. The documentation collected included information from websites, factsheets, reports, user guides, other academic publications, and news articles. We also contacted the data stewards to seek further information, where none was publicly available.

We analyzed the datasets in accordance with the eight Open Data Principles (ODP), and several other relevant principles recognized by the Open Data movement (https://opengovdata.org/; see Table 2). These principles arose out of a 2007 meeting of prominent open government advocates and have been further formalized in the creation of the OGP in 2011. We chose these principles over other possible frameworks (e.g., https://5stardata.info/en/), as they offered flexibility and breadth when considering data sources for analysis. The ODP are also published alongside seven other principles from the Open Data literature which we also included in our analytical framework. In the results, we provide reflections on how we assessed different principles and the challenges each posed when making said assessment. We searched each document, coded relevant information to the components of the framework, and tracked our coding in a spreadsheet that included descriptive details of the relevant dataset.

6. Results of open data analysis

As noted above, we identified 22 data sources as being relevant to Australia's workplace relations framework. Most of these data sources were from governmental sources, namely regulatory agencies associated with workplace relations (e.g., FWO, FWC, SafeWork Australia) and the Australian Bureau of Statistics (ABS), the national statistical agency. During our research, we evaluated each of these data sources based on public information using the open data framework. The findings of this evaluation are outlined in more detail below. We also attempted to access several of these data sources without success. Much like the case studies described previously, the findings of our evaluation were not black and white, and we found that the framework was not as readily adaptable as anticipated. Primarily, we found that the ODP required some interpretation when applying them to the current study, since the evaluation was taken

https://doi.org/10.1017/dap.2024.18 Published online by Cambridge University Press

Table 1. *Australian workplace relations datasets*

Dataset name	Description
FWO Enforceable undertakings data FWO Litigation outcomes FWO Proactive compliance deeds	The Fair Work Ombudsman is Australia’s workplace relations regulator. It provides PDF documents that outline information on its regulatory activities
FWC Modern award pay database	The Fair Work Commission is Australia’s workplace tribunal. It also has oversight of the various pay settings regulations such as modern awards. The MAPD contains the calculated minimum rates of pay, allowances, overtime, and penalty rates in modern awards in CSV format and through an API
ABS–ATO: Weekly payroll and job wages in Australia	The Australian Taxation Office receives information from employer’s payroll and accounting software, which it provides to the ABS to produce statistics on jobs and wages in Australia. This dataset includes data on all payroll jobs in Australia and their associated wages
DEWR: Fair Entitlements Guarantee claims data	The Department of Employment and Workplace Relations operates the Fair Entitlements Guarantee, which is a safety net scheme for employees whose employer has become insolvent
Safe Work Australia data	Safe Work Australia is a Federal statutory agency responsible for developing national WHS policy, including developing and evaluation model WHS frameworks. They are also responsible for collecting, analyzing, and publishing evidence on WHS in Australia
Worksafe Data and Statistics QLD Worksafe Data and Statistics SA Worksafe Data and Statistics WA Worksafe Data and Statistics NT Worksafe Data and Statistics ACT Worksafe Data and Statistics VIC Worksafe Data and Statistics NSW Worksafe Data and Statistics TAS	Each State and Territory has their own WHS regulator, which publishes statistics on WHS incidents in their jurisdiction, as well as incidents These reports are mandated by the respective WHS laws of each State and Territory
Workplace Gender Equality Agency (WGEA)	WGEA is a federal statutory agency responsible for promoting and improving workplace gender equality. WGEA’s compliance reporting program required private sector employers, under the Workplace Gender Equality Act 2012, to report on gender equality in their organization
DEWR: Trends in Federal Enterprise Bargaining	Trends in Federal Enterprise Bargaining is a quarterly report produced by DEWR and contains data about the number of enterprise agreements made in the federal workplace relations system, as well as data about the number of employees covered and the level of wage increases included in collective agreements
DEWR: Workplace Agreements Database	The Workplace Agreements Database, managed by DEWR, contains over 160,000 agreements, with data on developments in coverage, wage increases, and conditions of employment included in collective agreements
	LEED is a government dataset, managed by the ABS, bringing brings together employer information (from BLADE) and

(Continued)

Table 1. *Continued*

Dataset name	Description
Australian Bureau of Statistics (ABS): Linked Employer–Employee Database (LEED)	employee information (from Personal Income Tax data) into a linked dataset
ABS: Multi–Agency Data Integration Project (MADIP)	MADIP is a government dataset, managed by the ABS, containing information on health, education, government payments, income and taxation, employment, and population demographics (including the Census) over time
ABS: Business Longitudinal Analysis Data Environment (BLADE)	BLADE is an economic data tool, managed by the ABS, which combines tax, trade, and intellectual property data with information from ABS surveys to provide a better understanding of the Australian economy and business performance over time
The Household, Income, and Labor Dynamics in Australia (HILDA)	HILDA is a nationally representative longitudinal study of over 17,000 Australian individuals residing in approximately 9,500 households. It is funded by the Australian Government Department of Social Services and managed by the Melbourne Institute at the University of Melbourne

from the perspective of university researchers, rather than from the perspective of the private or government entities for whom the ODP was designed. Further, the documentation available on each dataset did not always provide detailed enough information to make a clear analysis.

The application of the open data framework to the identified data sources revealed several themes related to the openness of workplace relations-related data in Australia. First, many of the available data sources fell short of open data best practices, notably relating to being “complete,” “accessible,” and “timely.” These shortcomings significantly limited the usability of the datasets beyond our initial collection purpose. In some cases, it was not possible to determine if the dataset was complete, due to a lack of information regarding the data collection methods or criteria for exclusion.

Second, there were also significant amounts of qualitative data included in the various data sources we analyzed. These qualitative data sources were compelling in that they provided in-depth information about certain cases. However, they were ultimately inadequate for our purposes as they did not capture the characteristics of this qualitative data through quantitative means. Technically, qualitative data such as administrative reports can be converted into quantitative data through various text processing techniques such as thematic analysis, sentiment analysis, entity recognition, and corpora comparison (Stravrianou et al., 2007; Mironczuk and Protasiewicz, 2018). We determined that readily available quantitative data would be preferable, as the skills required to transform qualitative data to quantitative data are comparatively scarce in regulatory and civil society organizations. Additionally, converting large volumes of qualitative data to quantitative data through text processing approaches can also be computationally costly and requires ongoing maintenance to convert the qualitative data to quantitative data with each new round of data collection. For example, we could access reports on enforcement activities undertaken by the FWO, but we could not access any quantitative data that would have allowed for an analysis of the trends in enforcement activity, the characteristics of the entities being investigated, or the outcomes of these activities.

Third, we encountered significant barriers in seeking to access useable data, particularly from government entities. This finding reflects a common theme across many data sources, especially from workplace relations regulators and agencies, that the data is maintained in a qualitative form, rather than in a quantitative, machine-processable format. It may be possible to apply certain techniques to the

https://doi.org/10.1017/dap.2024.18 Published online by Cambridge University Press

Table 2. *Open data principles*

Open data principles	Definition
Complete	All public data is made available
Primary	Data is provided as it was at the source, without being aggregated or modified
Timely	Data is available in a timely manner to maintain its value to end–users
Accessible	Data is made available, via the internet, to the widest range of end users. Accepted standards and protocols are adopted to ensure the easy reuse of the data
Machine processable	Data is structured in such a way to allow for automated processes, including documentation on the form of the data
Nondiscriminatory	Data is made available to everyone anonymously, thus not requiring registration, payment, or application
Nonproprietary	Data is made available in a format that does not add restrictions to the use of data and which is free to access
Online and free	Data is made available for free online, in a way that is easily findable. If a fee is charged, it should only be to cover the cost of reproducing the data online
Permanent	Data is made available in a stable format indefinitely
Trusted	Information should be provided as to the authenticity and integrity of data, such as digital signatures, to ensure the public can trust the data has not been modified since publication
A presumption of openness	Data is made available as a default in a proactive way—this may be supported by legislation or regulation that provides for open data
Documented	Documentation should be provided alongside data that provides information as to the accuracy and current of the data
Safe to open	Data should be safe to open and utilize without risk to the end–user's digital security
Designed with public input	The public has an opportunity to input into the design and dissemination of public data

Adapted from https://opengovdata.org/.

qualitative data, such as natural language processes, to analyze them and pull out quantitative data. However, it is unlikely that this data would be validated by the agency that collected the data. This finding may also reflect the fact that most of this data was not designed with public input. The majority of this data was drawn from operational data, which is likely to be affected over time by changes in regulator behavior and evolving approaches to compliance and enforcement. There was little evidence to suggest that regulators considered the secondary use of their data by outside sources.

Fourth, concerning the "presumption of openness," we noticed how the legislative requirement to publish data was translated differently by distinct jurisdictions and regulators. For example, most Workplace Health and Safety (WHS) regulators are required to publish statistics related to WHS under their jurisdiction's WHS Act. However, there is a significant difference in how some jurisdictions go about providing this data on their websites. Some States in Australia, like South Australia, provide downloadable CSV files of their data on their website, while other states provide infographic-style images capturing high-level statistics. Similarly, these agencies are required to publish data on incidents, however, there was no way to validate if this data was complete or up to date, as much of it was captured qualitatively. This appears to reflect several data sources where there is a legislated requirement to publish data; the method of publication is left up to the regulator. Ultimately, this encourages variable approaches, and many do not meet the principles of open data.

https://doi.org/10.1017/dap.2024.18 Published online by Cambridge University Press

Fifth, there was a lack of primary data availability across the data sets identified. Most data available, especially those from government sources, had been aggregated. There was also limited documentation describing the method of aggregation or noting if there had been a process of modification allowed for end-users. In the few data sets that did provide primary data, such as the university-run Household, Income, and Labor Dynamics in Australia (HILDA) survey and the government-operated Multi-Agency Data Integration Project (MADIP), we noted that documentation was in-depth and training was provided to encourage utilization of the datasets.

Sixth, while much of the data we assessed was "trusted" due to it being provided on a government website, digital signatures and other ways to confirm authenticity were not utilized effectively across the analyzed datasets. This could conceivably impact the trustworthiness of the data, as a lack of oversight could allow unscrupulous actors to forge and share copies of the data without safe management protocols. There also appeared to be a low uptake of application program interfaces (APIs), which may allow agencies to provide direct access to their data without sharing the data files themselves. One exception to this is the FWC's Modern Award Pay Database API. The API provides registered users with access to information in the Modern Award Pay Database, which includes the minimum rates of pay, allowances, overtime, and penalty rates set by the industry and sectoral level instruments (Modern Awards), which are the regulations governing the pay and conditions of many Australian workers (Fair Work Commission, 2023).

Seventh, the most significant data available from the ABS is limited through the use of proprietary platforms, coupled with considerable restrictions regarding analysis and sharing. Users are required to access the data through an ABS-controlled online platform, either TableBuilder or DataLab, depending on the data source being accessed. While some publicly available employment data can be exported at will from TableBuilder, more sensitive data from the MADIP and the ABS Business Longitudinal Analysis Data Environment (BLADE) is accessed through DataLab, which requires ABS sign-off before it can be shared with those outside the authorized research team. In addition, MADIP and BLADE data cannot be joined with other datasets without the permission of the ABS. It is likely that these limits are in place to protect the privacy of the data, however, these controls also place additional burdens on its use beyond descriptive analytics. For the purpose of this study, this hinders the ability of machine learning models to generate new features by joining and manipulating diverse datasets, as well as curtailing their ability to experiment with several combinations across different modeling approaches to select the most suitable model.

Finally, these highly valuable data sources from the ABS did not meet the criteria of being "nondiscriminatory" as significant levels of understanding and expertise are required to successfully gain access. Potential users must complete a detailed application process to access the data and agree that the data only be used for the listed project/s in the application, which must be approved by the ABS and the relevant agencies that provide data to the agency. Project applications must include a justifcation of the value they will provide, which is assessed by the ABS and relevant agencies. There is also a significant cost to accessing the data if the applicant is not from an approved institution that has a partnership with the ABS, such as a university.

7. Our experience in accessing and using these data sources

Our research began with government datasets that offered the most extensive coverage of workers and businesses. These were the MADIP and BLADE datasets, respectively, which as we have noted are administered by the ABS. MADIP and BLADE are linked datasets about individuals and businesses, aggregated from several government agencies that collect data on these entities through social services, taxation, immigration, or other government channels.

We sought to access this data to assess its suitability for use in training a machine learning model that could identify businesses at high risk of underpaying staff. However, it soon became clear that a key challenge would be the restrictions placed on the BLADE dataset for workplace compliance purposes conducted by any entity other than the FWO (Department of Industry, Science, and Resources, 2017). According to the existing taxation laws under which the majority of administrative data on businesses is

https://doi.org/10.1017/dap.2024.18 Published online by Cambridge University Press

collected, only the FWO is authorized to receive this information for the purpose of compliance with the Fair Work Act 2009 (Cth) (Item 5, Table 7, Schedule 1, Taxation Administration Act, 1953 (Cth)). It is understandable that restrictions are put in place on these datasets, as businesses provide this data to the government with the guarantee it will be used for express purposes outlined in the legislation, and a breach of this guarantee may discourage accurate business reporting.

When this barrier was discovered, we attempted to requisition aggregated data on noncompliance in relation to young workers from the FWO. Although representatives of the agency indicated they were willing to consider this request, and we complied with the FWO's protocols for requesting release of data, no data was provided to us.

We also sought to access data from organizations that frequently represent or support employees, including community legal centers (CLCs) and unions. A key challenge in engaging with these organizations related to their digital capabilities and the resultant quality of the data they collected. In many cases, we were told that the extent of data used in these organizations was for case management and membership management purposes. CLCs have access to financial accounting and management software called CLASS, which is provided by their peak body so that they can meet their government funding reporting requirements. However, many of the CLCs we spoke to when trying to access their data outlined how this software did not meet their everyday practice needs. These CLCs had set up other data collection systems using spreadsheets and different office management software. They then exported data from these systems into the CLASS system for reporting purposes. Many of the CLCs we engaged with spoke of the challenges they faced in managing their data, and how their data collection was primarily driven by the reporting requirements imposed upon them by government funding arrangements. We did not end up accessing or utilizing any CLC data due to the data quality issues posed.

Due to these barriers, an alternative was sought in the HILDA Survey, a nationally representative longitudinal study of over 17,000 Australian individuals residing in approximately 9,500 households. The household-based panel study collects information about economic and personal well-being, labor market dynamics, and family life and is funded by the Australian Government Department of Social Services through the Melbourne Institute, an economic and social policy research center at the University of Melbourne. The data collected by the HILDA Survey is of a general nature and does not identify the individuals that it tracks. Significantly, it also does not contain any labels on whether the individuals had encountered underpayment. Therefore, efforts were made to build a labeled dataset by identifying individuals who reported being paid below the minimum wage based on their calculated hourly wage. Once the below-minimum-wage-earning individuals were identified and tagged, several modeling approaches were explored to show whether leading indicators such as age, industry sector, level of education, and having children had high predictive value for identifying underpayment.

To explore the full range of opportunities offered by data science approaches, various statistical, geometric, tree-based and clustering methods to compare performance. Five common models were fitted to the data, in order to identify the relative performance of each approach on the dataset available. They are as follows: baseline model, logistic regression, support vector machine, decision tree, and K-means clustering. These models go beyond the existing methods used in the case studies in which multiple (multivariate) linear regression was primarily employed. In our study, the most performant model was the support vector machine, which proved to be 77.8% accurate. However, the more meaningful comparison was not between the performance of the different models but the comparison between the different datasets used to train the models. Regardless of the type of models used, the models trained on a dataset that was further enhanced with feature engineering performed on average 4% more accurately. Feature engineering, as discussed in Case Study 2, is a way of inferring additional information about an entity from existing data points (Heaton, 2016). We used data on weekly wages and weekly hours to calculate whether a worker was paid below the minimum wage. We then used their job role and industry code to infer which industry award the worker was likely to be subject to, in order to identify whether this low payment was illegal or justified under their Industry Award. This new label, a new feature generated through the inferences, would then become a key

https://doi.org/10.1017/dap.2024.18 Published online by Cambridge University Press

indicator in the model to determine whether a worker was at high risk of being underpaid in contravention of the FW Act.

However, in an ideal scenario, rather than making such inferences, this task could have been simplified if it were possible to link the data effectively. This would combine the worker's real-time income, social security support status, and demographic information (such as age or migration status) to make the inference of illegal underpayment more accurate. The inability to link data due to the previously mentioned complications (e.g., differing data structures, terminology, collection cadences, barriers to access, and limits on use) hampers this option. The Fair Day's Work project hopes to use a minimum viable product to demonstrate the potential to detect underpayment at scale using data science and machine learning. Additionally, the findings could be used to strengthen the use of government and civil society data in protecting workers from underpayment and other contraventions of their minimum entitlements. In particular, the project seeks to improve the quality of available data, reduce information silos, and improve organizational maturity to sustain predictive models that provide adequate confidence amongst regulators and other stakeholders interested in enforcing compliance.

8. Recommendations to overcome institutional challenges to using open data for workplace compliance

The technical and institutional challenges that impede the accessibility of quality, open workplace relations data in Australia also hamper proactive regulatory activity. The FWO has been striving to move beyond a reactive approach when addressing underpayments, but the agency has been historically reliant on public complaints, investigate journalism, and business self-reporting to recover underpayment (Hannah, 2022; Hardy et al., 2023). As noted earlier, trade unions are hampered by capability gaps and limited resources. However, as our study has shown, the lack of adequate data containing key indicators of underpayment in existing literature curtailed the predictive ability of our model from the outset. Nonetheless, the study exposed opportunities for revising existing data collection approaches, which could potentially increase our model's ability to systematically identify workplace noncompliance.

To overcome the existing institutional barriers, we suggest the following three "next steps." Although directed to the Australian context, these steps may be adaptable to different workplace relations ecosystems:

1. Develop a centrally-maintained data management framework to support workplace compliance;
2. Shift perceptions towards data science tooling among organizers and worker representatives; and,
3. Lift data literacy among organizers and worker representatives and encourage collaboration with data specialists.

8.1. Recommendation 1: developing a centrally-maintained data management framework to support workplace compliance

Data access and data quality are key limitations in the predictive model development process. Administrative tax records collected by the ATO could not be used for enforcement purposes even though general noncompliance matters identified by the ATO are likely to indicate noncompliance in other areas of business regulation, such as workplace relations (Bernhardt et al., 2010). Other available data sources, such as the HILDA dataset and other government agencies, lacked the relevant granularity and completeness of income information to generate a reliable model.

Further, since occupational change is under-reported in tax data, there are additional challenges in attributing underpayment due to occupational mobility, even if tax records could be used for underpayment detection purposes (Hathorne and Breunig, 2022). As a result, future attempts to assemble existing data sources for the purpose of workplace compliance appear futile.

As a review of the *Privacy Act 1998* (Cth) is currently underway in Australia (Attorney-General's Department, 2023), the most direct solution to overcome data access challenges would be for any legal

https://doi.org/10.1017/dap.2024.18 Published online by Cambridge University Press

reforms following the review to include removal of the existing "employee records exemption" under the Act. This exemption currently excludes employee data held by employers from the protection of the Act. The removal of the exemption would require accurate and timely recording by employers of wage payments and other employee-related data at workplaces. The data gathered could be made accessible by default to both the employees (who are the data subjects) and their accredited representatives in order to deliver timely actions when responding to any detected underpayments (Chen and Howe, 2022).

8.2. Recommendation 2: shifting attitudes toward data science in enabling compliance with workplace laws

During consultation with government and civil society stakeholders, there was a perception that data science tooling or "AI" was expensive to use and opaque for its users. Stakeholders were anticipating a costly procurement of external expertise and raised concerns that use of these tools would incur high levels of risk. However, this perception is increasingly inaccurate, given the maturing nature of the data science discipline, and the increasing number of products available to drive down the cost of adopting data science techniques across the economy.

For instance, our project used only open source and well-established data science resources, such as the scikit-learn library and Jupyter Notebook, to assemble the data and build the model (Sci-kit learn, 2023). Both resources were free to use and easy to access, meaning that the key constraint associated with adoption comes from user data literacy and awareness of data science techniques, rather than the costs anticipated by stakeholders.

8.3. Recommendation 3: lifting data literacy and capability

Many worker representatives and their industrial organizers come from the industry sectors they represent, or are graduates in law, industrial relations or related fields. In carrying out their responsibilities, they are supported with advice from legal professionals (both internal and external) during industrial action. Despite the technical nature of the legal issues canvassed by Australia's complex industrial relations system, worker representatives are able to navigate the enterprise bargaining process and help workers uphold their contractual rights under the law with the support of legal specialists.

In the emerging paradigm of data-driven insights for regulatory compliance, worker representatives should be empowered to collaborate effectively with data specialists in the same way that they currently work alongside legal specialists. This does not mean that all worker representatives need to be able to code, but we argue that achieving a level of data literacy that allows them to adequately assess their data needs could be beneficial. By learning which data could help inform the identification of noncompliance, worker representatives could collaborate with data specialists (either internally or externally, as with legal specialists) to direct scarce resources towards the greatest need.

9. Conclusion

As we have shown, the relatively underexplored nature of data science in the context of compliance with workplace relations laws offers potential for further scholarship and future application. Our analysis suggests that the detection and prevention of "wage theft," and other forms of noncompliance with workplace laws, could be greatly improved using targeted data science interventions. However, while there are some innovative real world examples where data science has been used for enforcement, there is a way to go in realizing the potential of data science within a workplace relations context. Our study has highlighted one of the key challenges, namely the availability of adequate datasets that will allow predictive models to be developed. This study has provided an in-depth exploration of the data available in the Australian workplace relations context and evaluated it using an open government data framework. We found that the currently available datasets are inadequate for the tasks described. Nevertheless, we believe that there is room for optimism in the recommendations we have made to improve the integration of data science with workplace relations law enforcement in the future, not just in Australia, but in many jurisdictions around the world.

https://doi.org/10.1017/dap.2024.18 Published online by Cambridge University Press

Data availability statement. The data that support the findings of this study is publicly available data, as listed in Table 2 of the article. The authors are not the custodians of any data referred to in this article.

Author contribution. C.C. led the data curation and formal data investigation and analysis for this article, with support from T.K. C.C., J.H., and T.K. contributed equally to the conceptualization, methodology, and writing of the article. S.J. assisted with the review and editing of the article.

Funding statement. The research conducted for this article was supported by funding from the Paul Ramsay Foundation (the Fair Day's Work project). The Foundation played a role in the study design of the overall project, but had no role in the study design of this article, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interest. The authors declare no competing interests in relation to this article.

References

Amengual M and Fine J (2017) Co-enforcing labor standards: The unique contributions of state and worker organizations in Argentina and the United States. *Regulation and Governance 11*, 129.

Attorney General's Department (2023) Australia's Open Government Partnership. Australian Government. https://www.ag.gov.au/rights-and-protections/australias-open-government-partnership (accessed 23 August 2023).

Azarias J, **Lambert J**, **McDonald P and Malyon K** (2014). Robust New Foundations: A Streamlined, Transparent and Responsive System for the 457 Programme. An Independent Review into Integrity in the Subclass 457 Programme, September 2014. https://www.homeaffairs.gov.au/reports-and-pubs/files/streamlinedresponsive-457-programme.pdf (accessed 23 August 2023).

Bernhardt A, **Polson D and DeFilippis, J** (2010) Working without laws: a survey of employment and labor law violations in New York City, National Employment Law Project. https://www.researchgate.net/publication/242757880_A_Survey_of_Employment_and_Labor_Law_Violations_in_New_York_City (accessed 23 August 2023).

Bodie M (2022) The law of employee data: Privacy, property, governance. *Indiana Law Journal 97*, 707.

Chan CM (2013). From open data to open innovation strategies: Creating e-services using open government data. In *46th Hawaii International Conference on System Sciences*. Piscataway, NJ: IEEE, pp. 1890–1899).

Chen C and Howe J (2022) Worker Data Right: The Digital Right of Entry. Centre for Employment and Labour Relations Law Policy Brief No. 5, Melbourne: University of Melbourne.

Coglianese C, **Zeckhauser R and Parson E** (2004) Seeking truth for power: Informational strategy and regulatory policy-making. *89 Minnesota Law Review 87*, 277.

Dahl O and Starren A (2019) *The Future Role of Big Data and Machine Learning in Health and Safety Inspection Efficiency.* EU-OSHA Discussion Paper, 15 May.

Data Availability and Transparency Act 2022 (Cth). https://www.legislation.gov.au/C2022A00011/latest/text.

DataVic Access Policy 2016 (Vic). https://www.data.vic.gov.au/datavic-access-policy.

Department of Industry Science and Resources (2017) Business Longitudinal Analysis Data Environment (BLADE). https://www.industry.gov.au/publications/business-longitudinal-analysis-data-environment-blade (accessed 23 August 2023).

Department of Prime Minister and Cabinet (2015) Australian Government Public Data Policy Statement. Australian Government. https://www.finance.gov.au/government/public-data/public-data-resources/public-data-policy-resources (accessed 23 August 2023).

Ebert I, **Wildhaber I and Adams-Prassl J** (2021) Big data in the workplace: Privacy due diligence as a human rights-based approach to employee privacy protection. *Big Data & Society 8*(1). https://doi.org/10.1177/20539517211013051.

Fair Work Commission (2023) Modern Awards Pay Database https://www.fwc.gov.au/agreements-awards/awards/modern-awards-pay-database (accessed 23 August 2023).

Flanagan F and Clibborn S (2023) Non-Enforcement of Minimum Wage Laws and the Shifting Protective Subject of Labour Law in Australia: A New Province for Law and Order? *Sydney Law Review 45*(3), 337–370.

Frangi L, **Masi AC**, **Faust R**, & **Rohraff, N** (2021). *Digital Technologies and Workplace Relations: managers, colleagues, trade unions.* Draft report for the SSHRC Knowledge Synthesis Grant: Skills and Work in the Digital Economy. Ecole des Sciences de la gestion, Universite du Quebec a Montreal and Desautel Faculty of Management, McGill University, Montreal, Canada.

Government Information (Public Access) Act 2009 (NSW). https://legislation.nsw.gov.au/view/html/inforce/current/act-2009-052.

GovLab. The State of Open Data Policy Repository. https://repository.opendatapolicylab.org/ (accessed 23 August 2023).

Gudivada VN, **Apon A and Ding J** (2017) Data quality considerations for big data and machine learning: Going beyond data cleaning and transformations. *International Journal on Advances in Software 10*(2), 1–20.

Hannah K (2022) Speech to the Australian Industry Group PIR Conference on behalf of the Fair Work Ombudsman, 8 August 2022. https://www.fairwork.gov.au/newsroom/speeches#2022 (accessed 23 August 2023).

Hardy T (2021) Digging into deterrence: An examination of deterrence-based theories and evidence in employment standards enforcement. *International Journal of Comparative Labour Law and Industrial Relations 37*(2/3), 133–160

Hardy T, **Cooney S and Howe J** (2023) A balancing act: The difficulties of detecting labour violations and the implications for employer compliance and deterrence. *Australian Journal of Labour Law 36*(1) 1–29.

https://doi.org/10.1017/dap.2024.18 Published online by Cambridge University Press

Harrison TM and Syogo DS (2014) Transparency, participation and accountability practices in open government: A comparative study. *Government Information Quarterly 31*(4), 513–525.

Hathorne C and Breunig R (2022) Occupational Mobility in the ALife Data: How Reliable are Occupational Patterns from Administrative Australian Tax Records? *Economic Papers 41*(4), 297–324.

Heaton J (2016) An empirical analysis of feature engineering for predictive modeling. *IEEE SoutheastCon 2016*, 1–6.

Henman P (2020) Improving public services using artificial intelligence: Possibilities, pitfalls, governance. *Asia Pacific Journal of Public Administration 43*, 209–221.

Höchtl J, **Peter P and Schöllhammer R** (2016) Big data in the policy cycle: Policy decision making in the digital era. *Journal of Organizational Computing and Electronic Commerce 26*(1-2), 147–169.

Howe J and Kariotis T (2021) *A Fair Day's Work: Detecting Wage Theft with Data*. Melbourne: Melbourne School of Government. https://government.unimelb.edu.au/research/regulation-and-design/Home/detecting-wage-theft-with-data (accessed 23 August 2023).

International Labor Organisation (2020) Innovations in Labour Law Enforcement and Inspection—A Field Scan by the Governance Lab. https://www.ilo.org/global/topics/labour-administration-inspection/areasofwork/policies-and-methods/lang–en/index.htm (accessed: 19 October 2022).

Janssen M, **Charalabidis Y and Zuiderwijk A** (2012) Benefits, adoption barriers and myths of open data and open government. *Information Systems Management 29*(4), 258–268.

Jeong W, **Lee K**, **Yoo D**, **Lee D and Han S** (2018) Toward reliable and transferable machine learning potentials: Uniform training by overcoming sampling bias. *Journal of Physical Chemistry 122*(39), 22790–22795.

Johnson T, **Peterson F**, **Myers M**, **Taube RS and Fischer M** (2018) Predicting, Analysing and Educating on Wage Theft with Machine Learning Tools, CIFE Technical Report #TR229, Stanford University. https://stacks.stanford.edu/file/druid:mx396wr3611/TR229.pdf (accessed 23 August 2023).

Kroodsma DA, **Hochberg T**, **David PB**, **Paolo FS**, **Joo R and Wong BA** (2022) Revealing the global longline fleet with satellite radar. *Scientific Reports 12*(21004). https://doi.org/10.1038/s41598-022-23688-7.

Martin AS, **De Rosario AHD and Perez MCC** (2015) An international analysis of the quality of open government data portals. *Social Science Computer Review 34*(3). https://journals.sagepub.com/doi/full/10.1177/0894439315585734.

McCann D and Cruz-Santiago A (2022) Labour data justice: A new framework for labour/regulatory datafication. *Journal of Law and Society 49*, 658–690.

McDonald GG, **Costello C**, **Bone J**, **Cabral RB**, **Farabee V**, **Hochberg T**, **Kroodsma D**, **Mangin T**, **Meng KC and Zahn O** (2020) Satellites can reveal global extent of forced labor in the world's fishing fleet, *Proceedings of the National Academy of Sciences* 202016238; https://www.pnas.org/doi/10.1073/pnas.2016238117.

Mironczuk MM and Protasiewicz J (2018) A recent overview of the state-of-the-art elements of text classification. *Expert Systems with Applications 106*, 36–54.

OECD (2019) OECD OURdata Index: 2019. https://www.oecd.org/gov/digital-government/ourdata-index-australia.pdf (accessed 23 August 2023).

OECD (2021) *Data-Driven, Information-Enabled Regulatory Delivery.* Paris: OECD Publishing. https://doi.org/10.1787/8f99ec8c-en.

O'Leary K, **O'Reilly P**, **Nagle T**, **Filelis-Papadopoulos C and Dehghani M** (2021). The sustainable value of open banking: Insights from an open data lens. In *54th Hawaii International Conference on System Sciences, Kauai, Hawaii, USA, 4–8 January 2021*. University of Hawai'i at Manoa, pp. 5891–5901.

Park CH and Kim K (2022) Exploring the effects of the adoption of the open government partnership: A cross-country panel data analysis. *Public Performance & Management Review 45*(2), 229–253. http://doi.org/10.1080/15309576.2022.2042703.

Public Sector (Data Sharing) Act 2016 (SA). https://www.legislation.sa.gov.au/lz?path=%2FC%2FA%2FPUBLIC%20SECTOR%20(DATA%20SHARING)%20ACT%202016.

Queensland Open Data Policy (2022–2024). https://www.data.qld.gov.au/_resources/documents/qld-data-policy-statement.pdf

Reggi L and Dawes SS (2022) Creating open government data ecosystems: Network relations among governments, user communities, NGOs and the media. *Government Information Quarterly 39*(2), 101675.

Reichert C (2017) Australian government unveils open data framework for cities, ZDNet, 7 December 2017. https://www.zdnet.com/article/australian-government-unveils-open-data-framework-for-cities/ (accessed 23 August 2023).

Right to Information Act 2009 (Tas). https://www.legislation.tas.gov.au/view/whole/html/inforce/current/act-2009-070.

Rogers B (2023) *Data and Democracy at Work: Advanced Information Technologies, Labor Law, and the New Working Class*. Boston: MIT Press.

Sadiq S and Indulska M (2017) Open data: Quality over quantity. *International Journal of Information Management 37*(3), 150–154.

Safarov I, **Meijer A and Grimmelikhuijsen S** (2017) Utilization of open government data: A systematic literature review of types, conditions, effects and users. *Information Polity 22*(1), 1–24.

Stravrianou A, **Andritsos P and Nicoloyannis N** (2007) Overview and semantic issues of text mining, *ACM SIGMOD Record*, *36*(3): 23–34. https://dl.acm.org/doi/abs/10.1145/1324185.1324190.

Taxation Administration Act 1953 (Cth). https://www.legislation.gov.au/C1953A00001/latest/versions.

Ubaldi B (2013), *Open Government Data: Towards Empirical Analysis of Open Government Data Initiatives*. OECD Working Papers on Public Governance, No. 22. Paris: OECD Publishing.

https://doi.org/10.1017/dap.2024.18 Published online by Cambridge University Press

Vosko L, and the Closing the Enforcement Gap Research Group . (2020) *Closing the Enforcement Gap: Improving Employment Standards Protection for People in Precarious Jobs* Toronto: University of Toronto Press.

Western Australia Whole of Government Open Data Policy (2022). https://data.wa.gov.au/sites/default/files/Open%20Data%20Policy%20v2%202022.pdf

Weil D and Pyles A (2006) Why complain? Complaints, compliance, and the problem of enforcement in the U.S. workplace. *Comparative Labor Law and Policy Journal 27*, 59.

Ying X (2019) An overview of overfitting and its solutions. *Journal of Physics 1168*(2).

Yoon H (2018) A machine learning evaluation of the COMPAS dataset. In *IEEE International Conference on Big Data*, IEEE.

Cite this article: Chen C, Howe J, Kariotis T and Jackson S (2024). Open data, workplace relations law compliance, and digital regulation. *Data & Policy*, 6: e26. doi:10.1017/dap.2024.18