Envisioning an Integrated and Open Labor Data Ecosystem

Challenges and Opportunities

WikiRate

ABOUT THIS REPORT

This report is the product of a research project commissioned by Humanity United and conducted by WikiRate to scope out the opportunities and challenges involved in building an open and integrated data ecosystem for labor rights in supply chains.

Turning to organizations working with labor and supply chain data, as well as their funders, the report identifies the technical and data challenges facing organizations in this space and compiles recommendations for more and better data sharing in the future.

The research builds on the work of many organizations and data practitioners working to open up data worldwide.

We are grateful to Humanity United for making this report possible and Ethos Matters for sharing research insights and connecting us with their network. We extend our gratitude to interviewees and surveyed members of labor data organizations who generously gave their time and expertise to the project.

CONTENTS

Introduction

The role of data and data-driven decision making has become central to the supply chains and labor rights space. In tandem, the number of tools and initiatives working on these issues has multiplied.

Despite strides being made on supply chain transparency in certain sectors, the challenge remains that organizations tend to collect labor data in silos with few datasets being shared between initiatives.

Without data sharing, organizations may end up doing double-work: collecting the same datasets rather than combining their efforts to amplify their impact. And as many organizations specialize in certain data sources, they may also be lacking the key pieces of the puzzle that they need to achieve and prove their impact.

This lack of interconnectedness is curbing the ability of organizations to achieve their common goal - that of improving the lives of workers in supply chains around the world.

WikiRate has been working on enabling data sharing between organizations for some time and has brought several labor rights datasets into the public sphere through our open data platform including data on supply chain relationships, wages and working conditions.

But these efforts are small compared to the wealth of data being collected by hundreds of civil society organizations, auditing organizations, traceability tools and worker voice tools.

It was in this vein that Humanity United formulated the question:

> "Could there be an integrated and open data ecosystem of organizations working on labor rights in supply chains?"

We set out to try and help answer this question by creating a better understanding of the technical opportunities and challenges that organizations are faced with.

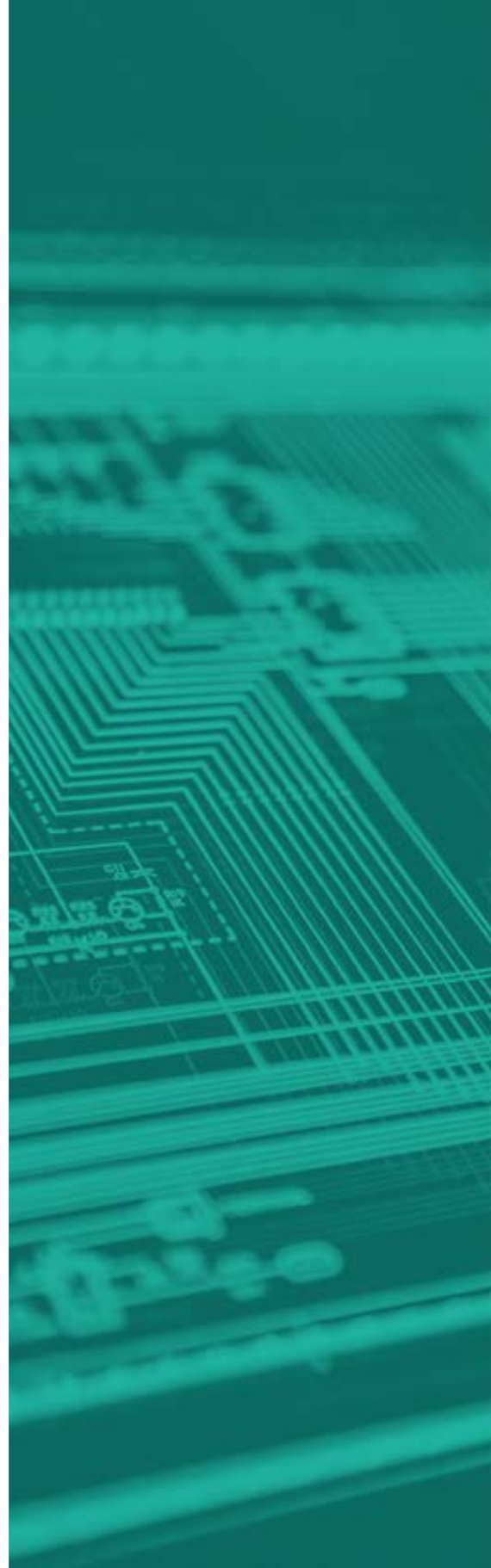
The first steps were to map the landscape of supply chain and labor rights initiatives and build a taxonomy of data types, technologies and organizations. This initial research was based on literature reviews and analysis of publicly available information on thirty-seven organizations.

The information gathered was complemented with insights gathered from a mix of semi-structured expert interviews and survey responses involving twenty-six organizations.

We then defined organizational archetypes based on how organizations use technology and data to meet their various goals.

The findings shed light on the current state of data collection, storage and sharing as well as the different tools and data licenses organizations are using to these ends.

This report presents the most relevant technical and data challenges and opportunities. From this analysis, we have compiled recommendations for organizations in this space, and their funders, to foster more and better data sharing in the future.



What is open, integrated data?

WHAT IS OPEN DATA?

> "Open data and content can be freely used, modified, and shared by anyone for any purpose"
> - Open Data Handbook

Simply put, 'data' refers to bits of information.

It can be both qualitative or quantitative, and although it makes good practice, it does not need to be standardized and structured to be considered 'data'.

It can be rough (also known as raw data) or it can be clean (also known as processed data).

It comes in many shapes and forms, be it statistics, interview records, literature reviews, public records, newsfeeds, company disclosures... the list goes on.

What makes data 'open' is the unrestricted potential for (re)use.

In practice, openness translates into two dimensions:

- **Legal openness** - licensing that allows re-use, modification and sharing (for free or at most, no more than a reasonable reproduction cost) by everyone
- **Technical openness** - making the data available and discoverable as a whole, in bulk, and in a machine-readable format (see page 7)

INTEROPERABILITY

A key benefit of making data open is the potential to integrate datasets through interoperability. With data stuck in silos, its reach and impact remain limited. Resources are inefficiently spent on duplicative efforts and data life cycles are short.

Open data is the foundation of an integrated and durable data ecosystem that facilitates collaboration and enables us to work at a scale that is needed to achieve systemic change.

Interoperability relies on:

1. The technical ability to connect and integrate different data systems
2. Agreed data standards between different data systems that enable sharing (like metadata, separators, etc)

THE FAIR PRINCIPLES

With its maxim '**as open as possible, as closed as necessary**' the FAIR principles can be a useful tool when considering how to open up datasets about supply chain facilities and the working conditions of the people who work within them.

They set out that data should be:

FINDABLE: The first step in (re)using data is to find them.

ACCESSIBLE: Once the user finds the required data, they need to know how they can be accessed.

INTEROPERABLE: The data usually needs to be integrated with other data. In addition, the data needs to interoperate with applications or workflows for analysis, storage, and processing.

REUSABLE: The ultimate goal of FAIR is to optimise the reuse of data.

MACHINE READABILITY

Publishing data in a PDF can be great for reading by humans, but these files are very hard for a computer to use. Making your data available in a machine readable format ensures the data extraction can be automated and its re-use scaled.

COMMON MACHINE READABLE FORMATS

- Comma-separated values (CSV)
- Excel files (XLS)
- JSON
- XML

WEBPAGES

Standard webpage content (HTML) is not necessarily machine readable. If you choose to report data on your webpage be sure to add structure to the data including HTML data tags and assigning specific identifiers to data tables.

When it comes to updating content, it is important to openly archive copies of the data to ensure previous versions of the dataset are not lost when changes are made.

> "Providing a clear definition of openness ensures that when you get two open datasets from two different sources, you will be able to combine them together, and it ensures that we avoid our own **'tower of babel': lots of datasets but little or no ability to combine them together into the larger systems where the real value lies."**
> - Open Data Handbook

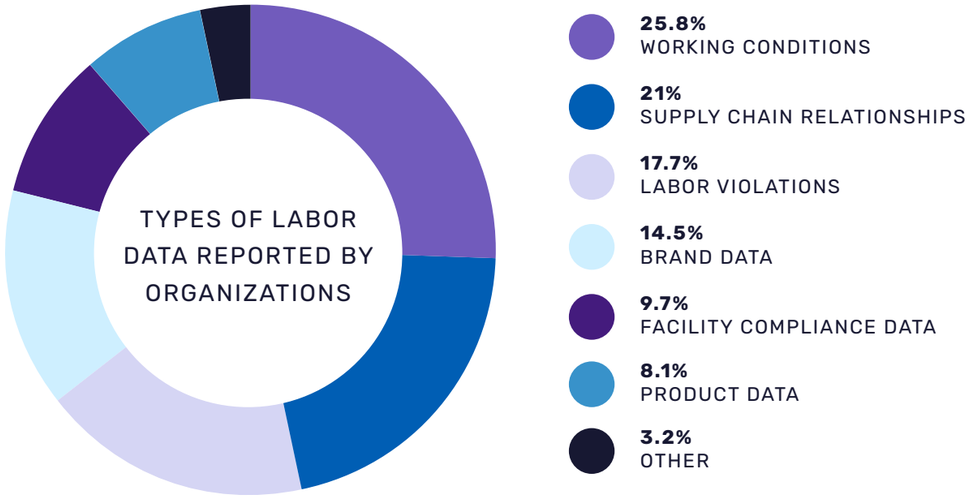
How open can labor data be?

> "As open as possible, as closed as necessary" – FAIR principles

As this study aims to scope out the opportunity for an open, integrated labor data ecosystem that **benefits workers**, labor data is understood as a relatively expansive term. It includes data on working conditions, demographics, trade flows, production capacity, supplier ownership and governance structures, procurement practices, industry initiative memberships, supply chain or product certifications, and so on.

Under this definition, the organizations surveyed and interviewed for this research reported to host and use the following labor data types:

1. Working conditions data (freedom of association, grievance mechanisms, wages, health & safety, worker surveys)
2. Supply chain relationships data (supplier lists, location data)
3. Labor violation data (violation records, forced labor incidents)
4. Brand level data (policies, purchasing practices)
5. Facility compliance data (health & safety, employment records)
6. Product data (certifications, product journeys)
7. Other (environmental data, satellite data)



Although some of this data already is, or may become open, **to what extent can the different types of labor rights data actually be open and shareable?**

While open data offers an unparalleled opportunity to scale the usefulness and impact of data, it should not be forgotten that 'when opening up data, **the focus is on non-personal data**', and so, it should not contain information about specific individuals.[1]

At times, labor data will relate to specific individuals or can be particularly sensitive. For instance, data related to human or labor rights violations, whistleblowing reports, and names or other data collected during worker surveys that could be traced back to an individual or group of individuals.

Other than that, also the privacy and safety of those working in the local context to gather the data should be ensured to avoid retaliation from possibly aggravated stakeholders.

In this vein, it should be remembered that **open data is a means to an end, and not the objective in and of itself**.

Depending on the sensitivity of the data and the risk of adverse impacts, data can be more or less restricted within a data spectrum that goes from closed data to trusted data sharing scenarios to publicly available open data.

RESOURCES FOR ORGANIZATIONS

- Open Data Handbook
- FAIR principles
- WEST principles
- Digital principles
- ODI Data Spectrum
- Go to Fostering Trust (page 20) for data management techniques and technical tools for handling sensitive labor data securely

Source: Adapted from the Open Data Initiative Data Spectrum

1 The Open Data Handbook, Open Knowledge Foundation

Taxonomy of organizations

This taxonomy is made up of seven archetypes which were created with the intention of guiding our understanding of how organizations are using technology and data for labor rights work in supply chains.

They were not created to capture every single dimension of the space and some organizations may find they fall between archetypes.

1 TRANSPARENCY & ACCOUNTABILITY ADVOCATE

This archetype gathers and/or uses data to generate evidence-based reports and campaigns to help improve supply chain workers conditions, monitor labor rights compliance, and call for better disclosure practices.

Target audience. Governments, downstream companies, worker rights organizations, unions

Tech. Typically do not use custom tools for data collection and rarely publish data in an open data format. Aggregated data may be made available in text-heavy PDF reports and XLS files

Licensing. Rarely specified

2 COMPLIANCE DATA AGGREGATOR

This archetype uses technology and site visits to gather and share data from suppliers on the working conditions in facilities. It engages with downstream companies who use the paid-for service to gather information about risks along their supply chain.

Target audience. Downstream companies, upstream companies

Tech. Gather and store data using custom tools (e.g. web survey tool) and have API connections

Licensing. Typically this archetype does not make data publicly available

3 DATA INTELLIGENCE HUB

This archetype uses innovative technologies and data science techniques to gather large datasets of disparate data and generates (predictive) trend analyses. Data may be used to track trade flows, detect criminal activity or generate risk profiles

Target audience. Governments, investors, law enforcement, downstream companies, CSOs

Tech. Artificial Intelligence and scraping tools, API connections

Licensing. Typically this archetype does not make data publicly available

4 WORKER VOICE TOOL

This archetype develops digital tools to gather direct and anonymous worker input regarding their working conditions. Workers access knowledge about their rights and downstream companies and workers associations receive information about potential risks.

Target audience. Workers, worker rights organizations, unions, downstream companies

Tech. Mobile phone apps, voice messages, social media and custom messaging applications

Licensing. Typically this archetype does not make data publicly available

5 TRACEABILITY TOOL

This archetype uses technology to help downstream companies trace products through supply chains and, in some cases, monitor facilities. They have high engagement with companies helping them to map their supply chain, and improve their sourcing practices.

Target audience. Downstream companies

Tech. Data is gathered and stored using custom tools (including blockchain) and shared using API connections

Licensing. Typically this archetype does not make data publicly available

6 OPEN DATA PLATFORM

Organizations in this archetype gather, clean, host and link data from different sources and make it publicly available in accessible and open formats. The archetype welcomes and facilitates users' data contributions to the platform.

Target audience. CSOs, academics, journalists, data scientists, the broader public, companies

Tech. Use APIs to gather and share information in different formats (JSON, CSV, XLS)

Licensing. Uses an open data license (CC BY 4.0, ODbL v1.0)

7 KNOWLEDGE SHARING PLATFORM

This archetype gathers and hosts documents and data in different formats to improve access to information and information sharing. Extracts data from reliable sources (reports, lawsuits). Data is discoverable via a search interface

Target audience. Journalists, CSOs, academics, companies, policymakers, the broader public

Tech. Likely to use content management systems and APIs

Licensing. Rarely specified

Map of data challenges

Working with labor rights and supply chain data poses a diverse set of challenges for actors in the space. This section will cover the **data challenges** that were most frequently mentioned during the research process and the opportunities that exist to resolve them.

The challenges are **tagged with the organizational archetypes** that cited them most often.

DATA VERIFICATION 1 2 4 5

Organizations signalled that data verification is problematic for their work, both for the data they have collected themselves and for datasets provided by external organisations.

- **Self-declarations**: data reported directly by facilities or companies is difficult to verify, particularly if organizations have no access to alternative sources
- **Lack of context:** datasets provided by external organizations may arrive without context about the sample, methodology and other key information needed to cross-check the information
- **Unstructured formats**: some organizations left it to the researcher to decide the data collection format or used unstructured spreadsheets or documents to collect data causing data validation problems

RESPONSE QUALITY 1 2 4

> "Sometimes workers or inspectors misunderstand the questions asked or they may not enter the data correctly into the system" - archetype 4

A number of organizations flagged the quality of the responses they receive from surveys of inspectors, facilities and workers as a key data challenge.

- **Shared understandings**: questions may not be properly understood and have conceptual equivalence across linguistic and cultural contexts
- **Reliable data input**: responses may not be entered correctly into the data collection system and online tools may be difficult to access in remote locations
- **Source reliability**: participant responses may be biased, and, in some cases, participants may be coerced into entering false information

STANDARDIZATION 2 3 4 6 7

One blocker to sharing datasets between organizations is the lack of standardized understandings of key concepts which would enable comparability between datasets.

- **Definitions**: each organisation has their own definitions for geographies, key terms (such as supply chain tiers, wages, overtime) and entities (brand, company, legal entity)
- **Entity recognition**: entity names and addresses differ between datasets making the process of entity mapping laborious and error prone
- **Lack of structure**: unstructured, qualitative datasets were problematic for organizations to parse, particularly if they did not have customized data tools

TECHNICAL CAPACITY 1 3 4 7

> "Our small CSO partners do not have dedicated technical capacity and can not efficiently share data with us in a way we can consume" - archetype 7

Technical challenges were cited in the research by organizations frustrated that this placed limitations on the expansiveness of their work and ability to share data.

- **API connections**: several more tech-enabled organizations mentioned datasets with which they were unable to integrate because external organizations had no or limited API infrastructure
- **Machine readability**: there are data sources available that organizations could not process because of a lack of technical infrastructure or challenges with less readable formats
- **Machine learning**: organizations which need to ingest large volumes of data expressed a need for machine learning solutions that would reduce manual work

Map of data opportunities

A number of opportunities exist to help resolve some of the data challenges expressed by the organizations during the research. The opportunities listed in this section are not an exhaustive list, rather they are a starting point for organizations looking to resolve one or more of the challenges we have identified.

DATA MANAGEMENT 1 2 4 7

Several challenges around data verification and standardization could be mitigated through better data management practices such as labelling and documentation.

- **Metadata** summarises basic information about data and provides structured context (such as labels) to data which improves findability
- **Documentation** is information provided about a dataset to give it contextual meaning such as an explanation of methodologies, concepts, variables or separators. Better data documentation improves the reusability of datasets by ensuring that dataset users understand the source and meaning of the data
- **Unique identifiers** are identifiers – usually a string of characters – that are not shared with any other record in a dataset. For example, a shared system of unique identifiers for entities facilitates entity recognition between datasets

DATA COLLECTION 1 2 4 5

We found that data collection challenges were well-understood by organizations in the space and several have implemented data solutions to improve the quality of responses which other organizations may find useful.

- **Layer of data verification**: One organization implemented an extra layer of data verification performed by experts working in the same cultural and linguistic context with a direct link with the data provider and who could clarify any discrepancies
- **Structured formats**: Using a structured format for collecting data with room for open-ended answers would improve the standardization of research data even when gathering mostly anecdotal or qualitative data
- **Supplementary evidence**: There are additional sources that can be collected alongside an interview or survey. One organization had a dataset of wage data derived from worker interviews that had some major deviations which they were able to validate because they had also collected copies of worker payslips

DATA TRIANGULATION 1 6 7

In some cases, organizations have a linear model of data collection which involves collecting data from a single source: for instance, facility self assessment or a survey of a group of workers. This leaves gaps in their ability to verify their data against that of other sources and test data accuracy issues related to response quality.

The sharing of datasets could be a mutually beneficial process for organizations, allowing them to triangulate between their datasets and that of other organizations to validate their conclusions and flag deviations.

TECHNICAL CAPACITY 1 7

In some cases, data sharing between organizations is held back by a lack of technical infrastructure. Without funding directed towards building technical capacity, organizations have to rely on time-consuming manual data extraction and analysis work. Civil society organizations in particular may lack technical infrastructure and would benefit from support:

- to build customized databases and data management systems,
- for investments in API infrastructure to enable better data sharing, and
- for machine learning projects.

An important caveat when it comes to tech development is for organizations and funders to **avoid reinventing the wheel** - there are a number of pre-existing open source solutions which can be applied to help solve the challenges above.

RESOURCES FOR ORGANIZATIONS

DATA MANAGEMENT

- Reference Guide for Data Archivists
- Metadata Basics
- The Meta-data Editor
- Using Identifiers

BUILDING TECHNICAL CAPACITY

- Best practices for REST API design
- Digital Insights into Modern Slavery Reporting: Challenges and opportunities of machine readability
- A Primer on Machine Readability for Online Documents and Data

Data sharing practices

The practical reality of an integrated, open labor data ecosystem is that the data is not centrally stored but is left with the data owners or controllers. Therefore, the infrastructure and governance framework for data sharing are paramount.

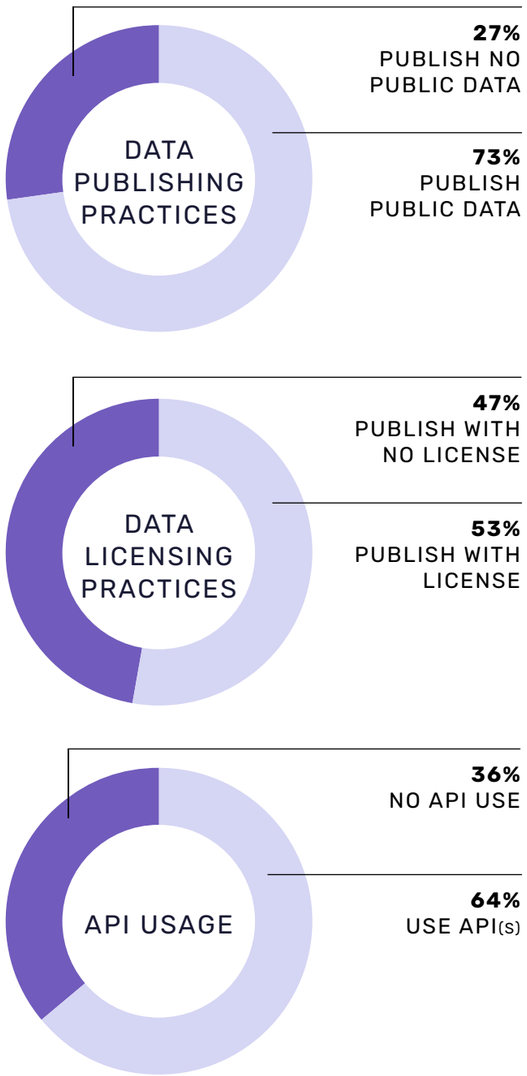
TAKING STOCK

> "73% of the organizations make data 'public' but nearly half of them do not do so under a license"

73% of the organizations make data 'public' but nearly half of them (47%) do not do so under a license, most of which have no licensing specification at all. So, while this means the data is available, **it does not allow re-use, modification and sharing**.

Spreadsheets are by far the most common format for data storage. It is a format that is familiar to many, relatively inexpensive, and easily customizable to fit different data needs.

More than a third of organizations working with labor data do not have or use an API to share data.



OPPORTUNITIES

To advance the ability to share labor data (from a practical standpoint) quick wins include:

- **Select and add clear licensing** to information that is made public (this includes pdf reports, spreadsheets, websites)
- With spreadsheets as the most common format in which data is stored **there is a need to raise awareness on and encourage the use of data practices and standards that enable data sharing**; integrating metadata like identifiers, consistency in data separators and data standardization, to name a few
- **Archive your data in existing open repositories that have API infrastructures** through which others can easily access your datasets.

CHOOSE A LICENSE

- Creative Commons tool: 'find your license'
- Publisher's Guide to Open Data Licensing

CREATIVE COMMONS LICENSING

LICENCE TYPE	SHARE	SHARE WITH ATTRIBUTION	CREATE MODIFIED VERSIONS	REDISTRIBUTE COMMERCIALLY	RELEASE UNDER ANY LICENSE
No Rights Reserved	Yes	No	Yes	Yes	Yes
Attribution	Yes	Yes	Yes	Yes	Yes
Attribution ShareAlike	Yes	Yes	Yes	Yes	No
Attribution-NonCommercial	Yes	Yes	Yes	No	Yes
Attribution-NonCommercial-ShareAlike	Yes	Yes	Yes	No	No
Attribution-NoDerivatives	Yes	Yes	No	Yes	-
Attribution-NonCommercial-NoDerivatives	Yes	Yes	No	No	-
All Rights Reserved	No	-	No	-	-

OPENNESS

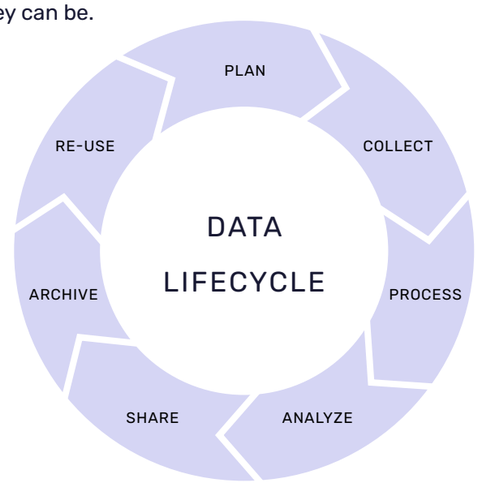
Fostering trust: inclusive approaches

Beside reducing the practical barriers for data sharing, the sensitive nature of labor data underscores the importance of trust between the organizations that share data and the data subjects. There are three approaches that will help build that trust.

INCLUSIVE DATA LIFECYCLES

Firstly it is important to **engage early and often throughout the data lifecycle** with the data subjects, intended beneficiaries (workers) and their 'info-mediaries' (the people and organizations that take the data and transform it to make it useful and actionable).

Involving these stakeholders in the different data lifecycle stages means their needs and concerns can be accounted for and addressed along the way and that the outputs will be as relevant and appropriate as they can be.



Types of questions that these stakeholders should be engaged in:

- To what end should the data be collected?
- What data is needed to achieve that goal?
- Who are the data subjects? How can they be reached? Why would they engage?
- What barriers and risks can we expect at the collection and sharing stages, and how might these be mediated?
- What data should not be collected and/or shared and why?

Moving beyond engaging stakeholders in data lifecycles, another crucial step to foster trust is a transparent and inclusive data governance framework.

DATA GOVERNANCE

Data governance frameworks define the approach taken to managing data availability, usability, accessibility, integrity and security.

They **are closely intertwined with the control of data, which is why it is important to consider the power (im)balance** between the different stakeholders when it comes to responsible data governance.

> "To satisfy all needs, [a data governance framework] is best co-designed by the future participants, with that process facilitated by an independent, neutral body"
> - IceBreaker One

IceBreaker One is an organization that has articulated the need for scalable data sharing of non-financial reporting data. They stress that in order to deliver a functioning data ecosystem there needs to be a governance framework for data access that creates trust.

In their report on 'Shifting Power through Data Governance' Mozilla outlines a number of data governance models that are used to 'steward' data in a way that empowers data subjects. These include:

- **Data cooperative**: Legal construct where individuals/organizations collaboratively pool data for the economic, social or cultural benefit of the group. The cooperative is often co-owned and democratically controlled by its members.
- **Data collaborative**: Data is shared publicly online or strictly between partners to make data that is proprietary or siloed available to inform research or policy. The collaborative acts as data steward to empower their members/the public to solve societal problems.
- **Data trust**: Legal relationship with trustees who steward data rights in the sole interests of a beneficiary and have a fiduciary duty. Data can be pooled from different sources and the trustee can negotiate access by others on behalf of the collective.

A full list of governance models can be found here.

THE CARE PRINCIPLES

In keeping with the democratic and empowering spirit of open data practices, the CARE principles are an example of how to shift the locus of power towards individuals having control and use of the data collected about them.

The CARE principles for Indigenous Data Governance were created by the Global Indigenous Data Alliance to address the imbalance between the dominant open data and open science approaches and the rights and interests of Indigenous Peoples.

They stress that the governance of data should be determined by those who have most knowledge about the risks that this data could pose to individuals if it were shared with other organisations or made public.

In summary the CARE principles center on:

COLLECTIVE BENEFIT

AUTHORITY TO CONTROL

RESPONSIBILITY

ETHICS

It should be noted that the CARE principles were developed to advance the rights and interests of Indigenous Peoples. As the context and use case of workers and labor data cannot be equated with that of Indigenous Peoples, these principles should be regarded as **an inspirational springboard rather than a ready to use guide**.

Fostering trust: data protection

The data and tech infrastructure can also ensure there is trust in the system and address concerns of (unintended) misuse or adverse impacts. In particular the following data management techniques and technical tools can be used to ensure access is only granted to the data that is suitable for sharing (non-personal data).

ANONYMIZATION

When re-identification is no longer needed and a link back to the original data should rather be avoided, data anonymization is the appropriate technique. Recital 26 of the GDPR defines anonymous data as:

"information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable."

Through **data anonymization**, direct and indirect personal identifiers are removed, which often includes data transformation like restructuring and reformatting the data.

If the information for each person that is part of the dataset cannot be distinguished from at least k-1 individuals, it is called **k-anonymization**. Simply put, this technique sets a minimum threshold for anonymity (e.g. number of workers) at which an individual can no longer be identified.

Please note that the thresholds will depend on the characteristics and sensitivity of the dataset (very detailed datasets about individuals will be harder to anonymize).

The first frontier to protect data subjects is **data pseudonymization**, defined in article 4(5) of the GDPR as:

" … the processing of personal data in such a way that the data can no longer be attributed to a specific data subject without the use of additional information, as long as such additional information is kept separately and subject to technical and organizational measures to ensure non-attribution to an identified or identifiable individual."

This approach is useful when data subjects should not be identifiable by those who have access to the data, but should be re-identifiable by, for instance, a specific organization or case worker.

This might be the case when after months of advocacy with factory owners, remedy will be provided to specific workers for sexual harassment.

In combination with using thresholds, **data aggregation** can be a helpful technique to achieve anonymization as well. For example, disclosing average worker wages at a facility level, only if a certain number of worker payslips have been used to determine that average.

DATA PROTECTION TOOLS

Apart from data techniques, there are also tech tools that can be used for data protection.

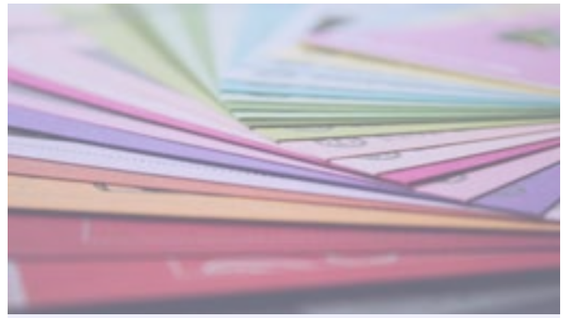
Data and server encryptions convert data into code to conceal the original information and avoid unauthorized access. Both digital information on computers and that which is sent between computers can be encrypted. Having such infrastructure in place will make sure you can collect, store and share worker data safely and responsibly.

Moreover, there are innovative tools that are in early stages of development, which aim to move from a model of data sharing that is intertwined with sharing data ownership to a **model that enables data sharing without sharing ownership**.

- OPAL: Open Algorithms[1] focuses on making data accessible without violating personal data privacy. By performing algorithm-execution on data at the location of the data repository, raw data never leaves its repository and access to it is controlled by the repository owner. Only aggregate answers or "Safe Answers" are returned.
- OpenPDS[2] provides a secure computation space for 3rd party applications to interact with the data and allows users to audit when and how applications have used their data.

1 A Trust::Data Consortium project

2 Developed by the Human Dynamics Group at the MIT Media Lab, led by Pentland et al



RESOURCES

- What Does it Mean? | Shifting Power Through Data Governance
- IceBreaker One: Data Infrastructure: Enabling secure and scalable non-financial reporting and data flows
- Shifting Power Through Data Governance

OPPORTUNITIES

Funders can support the development of an open and integrated labor data ecosystem by:

- convening stakeholders and facilitating the definition of a multi-tiered inclusive data governance framework that account for the different openness "layers" of data
- providing funding to data holding organizations to acquire data science skills that can help with wrangling large datasets and getting them ready for responsible sharing
- providing funding and open access to technologies such as encryption tools and tech infrastructure that allow data sharing without sharing ownership

What to make of all of this?

The labor data landscape includes many different data types and stakeholders, all of which have different needs, strategies, models and approaches to advance labor rights. It is important to identify opportunities for opening up data that do not have adverse impacts on the intended beneficiaries (workers) and do not undermine other effective strategies leading to improved working conditions. Rather it is about identifying pathways for how open data could help scale the advancement of labor rights.

A holistic overview of the different data sharing opportunities and needs in the data ecosystem can encourage improved data practice. That way, **even when your own strategy and data type does not enable you to open things up, you can still bake in the possibility for others to access that data, following a responsible framework of protections**. This table sketches out an example of how labor data (in this case wage data) can move across the data spectrum, and can be thoughtfully and meaningfully opened up, using the different tools and techniques outlined in this report.

	CLOSED	CLOSED	SHARED	SHARED	OPEN
Data example	Case file data on individual workers' wages (eg. interview records and pictures of pay slips)	Worker reports of wages and pictures of pay slips; anonymized but still records on the level of individuals	Aggregate wage indications per production facility, combined with data on facility ownership	Aggregate wage indications per production facility, combined with supply relations data	Aggregate wage indications per production facility, combined with supply relations data
Use case example	Front-line organization working on individual remediation	Organization building a legal case to litigate exploitation and unpaid wages	Organizations working with financial sector to freeze assets of facility owners who are exploiting workers	Organizations providing aggregate wage data on supply chain facilities as a service to investors for supply chain risk assessments, whilst also making it publicly available to worker representatives	Organizations campaigning with brands to improve purchasing practices and leverage their supplier relations to increase supply chain wages
Access	Internal	Named (explicitly assigned by contracts)	Group-based (via authentication)	Public (license that limits use)	Anyone
Licensing options	-	-	-	Attribution NonCommercial Attribution NoDerivatives Attribution NonCommercial NoDerivatives	No Rights Reserved Attribution Attribution ShareAlike
Governance options	Data trust	Data collaborative Data trust	Data collaborative Data trust	Data collaborative Data commons	Data collaborative Data commons
Technology options	Encryption (files and server), Pseudonymization	OpenPDS, OPAL: Open Algorithms, Encryption (files and server), K-anonymization	OpenPDS, OPAL: Open Algorithms, RESTful API, Aggregation and record thresholds	RESTful API	RESTful API

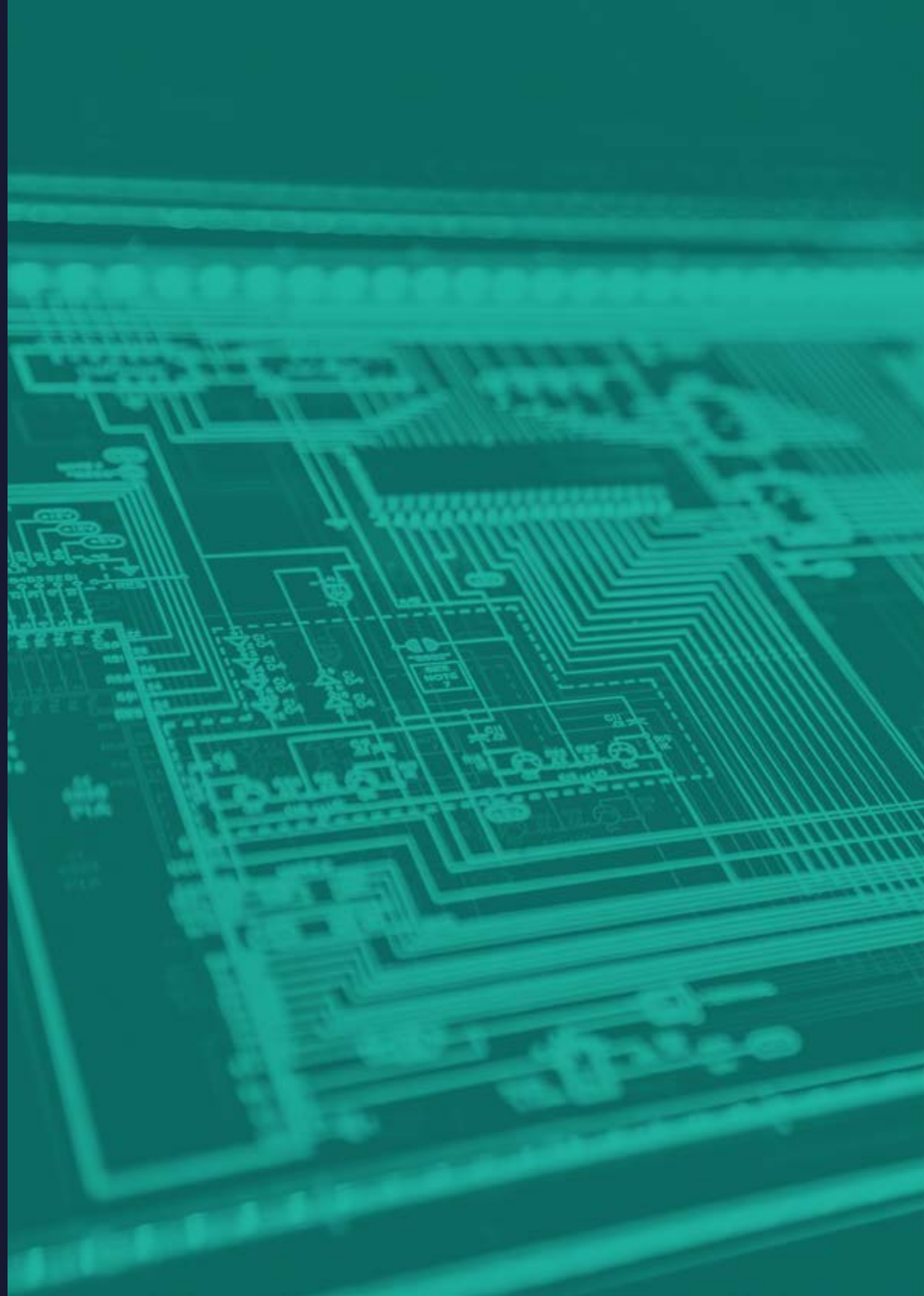
Recommendations

GENERAL

1. Creation of a working group of labor rights organizations and worker representatives to explore the practicalities and governance of sharing data and opening up certain aggregated datasets to the public
2. Creation of a pilot which tests opening up a specific type of labor rights data with a small group of organizations
3. Creation of a registry of labor rights and supply chain organizations listing their technical infrastructure, which types of data they hold and what data they make publicly available
4. Strengthening the technical infrastructure and data science capacity of organizations to improve data collection, protection, storage, archiving and sharing
5. Targeted interventions with organizations to set-up strategies for adding explicit licensing to publicly available data
6. Even when working with low tech tools or small budgets, use good data practices and standards that enable (automated) data sharing, like using metadata tags, publishing your data with an open license in an excel format and/or uploading it to an open registry that already has an API infrastructure

FOR FUNDERS AND INVESTORS

1. Include feasibility assessments for data sharing in funding ventures
2. Ringfence resources for building technical infrastructure for sharing (such as APIs) and robust and responsible data management practices



Getting started with opening up your data

For those organizations who are ready to open up a dataset, a step-by-step guide is detailed below.

The content in this section has been adapted from the Open Data Handbook published by the Open Knowledge Foundation.

Before you get started on opening up a dataset, keep in mind these **three guiding principles**:

- **Keep it simple**. Start out small, simple and fast. Deciding to open up datasets can be daunting, but focusing on a smaller dataset first is a good way to start getting data out there while testing your process and gaining feedback.
- **Engage early and often.** In the realm of labor rights and supply chain data it is essential that you engage with the actual and potential users and re-users of your data throughout the process. This helps to ensure that what you're putting out there is needed, impactful, accessible and safe to be shared.
- **Address common fears and misunderstandings.** This is especially important when sharing aggregated labor rights data. Acknowledge the fears and misunderstandings people may have around data sharing and ensure that you (a) identify the most important of these and (b) address them at as early a stage as possible.

1. CHOOSE YOUR DATASET

The first step is to choose the dataset you plan to make open. Remember that this process is iterative so you can return to this step if you encounter problems later on. If you already know which dataset you want to make open, you can move to the next step. If you still need to decide which dataset to focus on, you can:

Ask the community: The people who will be accessing and using the data will have a good understanding of what data is valuable to them. If the dataset includes aggregated labor rights data about workers, it is important to gain feedback from them and worker representatives on what they would find most useful

- Prepare a list of datasets you want feedback on
- Create a request for comment and publicise your request via a webpage
- Provide easy ways to submit responses
- Circulate the request to relevant mailing lists, forums and individuals
- Run a consultation event with the key stakeholders

Consider ease of release: Consider the time and resources you have to open up a dataset. What data do you have that would be easiest to make open? Small, easy releases can be a catalyst for developing a strategic approach towards opening up data. Note: do consider the value of small releases to your audience,

if the value is too low then it can undermine faith in the open data approach.

Observe peers: Learn from your peers in the labor rights and supply chain data space who are opening up data. They will have learnings they can pass on to you and can help amplify your datasets once they are published.

2. APPLY AN OPEN LICENSE (LEGAL OPENNESS)

It is essential that you **provide clarity on which open license your data is published under**. An open license will let your data users and re-users know under which terms the data can be transformed and shared. There are open licenses available from Creative Commons (table on page 17) and Open Data Commons.

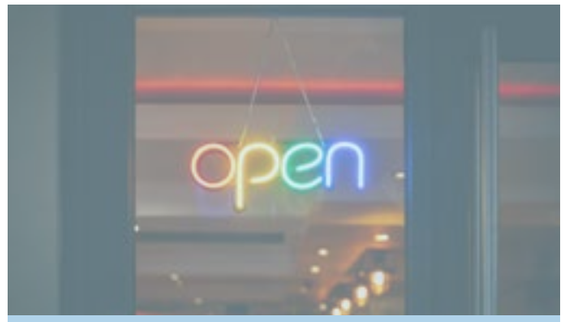
3. MAKE DATA AVAILABLE (TECHNICAL OPENNESS)

The data you publish needs to be available **in bulk in a machine-readable format**. See page 7 for more information about machine-readability and machine readable formats. You may wish to consider making data available via an API.

4. MAKE DATA DISCOVERABLE

Once your data has been made open, you need to **make it possible for your data users and re-users to find it**. As well as making them available on your website, there are a number of places you can publish datasets (see text box to right)

Data discoverability is also facilitated through metadata (page 14). Adding accurate descriptive information about your data will help potential users to find the data most valuable to them.



OPEN DATA REPOSITORIES

GENERAL

- GitHub
- Zenodo

SPECIALIST

- Open Apparel Registry: for global apparel facilities, their affiliations and IDs
- Open Development Initiative: for datasets about the Mekong region
- WikiRate: for data that can be tied to corporate entitites

Bibliography

- Au-Yeung, J. & Donovan, R. (2020, March 20). Best Practices for REST API Design. The Overflow. https://stackoverflow.blog/2020/03/02/best-practices-for-rest-api-design/
- Creative Commons. (2021). Choose a License. https://creativecommons.org/choose/
- Data.gov. (2012, September 24). A Primer on Machine Readability for Online Documents and Data. https://www.data.gov/developers/blog/primer-machine-readability-online-documents-and-data
- Dodds, L., & Young, L. (2020, December 23). Using Identifiers.. Open Data Institute. https://theodi.org/article/using-identifiers/
- Dublin Core Metadata Initiative. (n.d.). Metadata Basics. Retrieved July, 2021 from https://www.dublincore.org/resources/metadata-basics/
- Dupriez, O., Sanchez Castro, D. & Welch, M. (2019). Quick Reference Guide for Data Archivists. Household Survey Network. https://guide-for-data-archivists.readthedocs.io/en/latest/index.html
- Go FAIR (n.d.). FAIR Principles. Retrieved July, 2021 from https://www.go-fair.org/fair-principles/
- Mozilla Insights, van Geuns, J. & Brandusescu A. (2020, September). Shifting Power Through Data Governance. Mozilla Foundation. https://foundation.mozilla.org/en/data-futures-lab/data-for-empowerment/shifting-power-through-data-governance/
- Open Data Institute (n.d.). Open Data Spectrum. Retrieved July, 2021 from https://theodi.org/about-the-odi/the-data-spectrum/
- Open Data Institute. (2013, December 15). Publisher's Guide to Open Data Licensing. https://theodi.org/article/publishers-guide-to-open-data-licensing/
- Open Knowledge Foundation. (n.d.). The Open Data Handbook. Retrieved July, 2021 from http://opendatahandbook.org/guide/en/
- Principles for Digital Development (n.d.). Digital Principles. Retrieved July, 2021 from https://digitalprinciples.org/principles/
- Regulation (EU) 2016/679. (2016). General Data Protection Regulation. GDPR.eu. https://gdpr.eu/tag/gdpr
- Research Data Alliance International Indigenous Data Sovereignty Interest Group. (September 2019). CARE Principles for Indigenous Data Governance. The Global Indigenous Data Alliance. https://www.gida-global.org/care
- Starks, G., Cheetham, M., & Patchay, J. (2021). Data Infrastructure. Enabling Secure and Scalable Non-financial Reporting and Data Flows. Ice Breaker One. https://icebreakerone.org/report-nfdf/
- Walk Free, The Future Society, WikiRate, Business and Human Rights Resource Centre. (2020). Digital Insights into Modern Slavery Reporting: Challenges and Opportunities of Machine Readability. https://www.walkfree.org/reports/digital-insights-into-modern-slavery-reporting/
- Worker Engagement Support by Technology (n.d.). WEST Principles. Retrieved July, 2021 from https://westprinciples.org/about/
- World Bank.(n.d.).The Meta-data Editor Guidance. Read the Docs. Retrieved July, 2021 from https://metadata-editor.readthedocs.io/en