Data Enrichment Transparency Template

Purpose of this Document

In order to enable an ecosystem of accountability for stakeholder's actions around data enrichment practices, we believe that there is a need to increase transparency around data enrichment practices across global data supply chains. *Publishing Transparency Reports* is one of the key actions outlined in our proposed Path for Developing Responsible AI Supply Chains. Below represents a template for transparency which is meant to outline what companies should be reporting on, regarding their data enrichment practices. We have included 2 sets of questions:

This document is part of PAI's Responsible Data Enrichment Sourcing Library

1. **Questions that each development team should be reporting on for a given data enrichment project.**

   We recommend that each team reports how they are setting up data enrichment projects and reports this to be tracked by a more central team (for example, a responsible AI team or other internal governance team)

2. **Questions that companies should be tracking and reporting on at a higher level.**

   We recommend that companies adapt their existing reporting infrastructure and workflows to include information about their data enrichment practices. This may look different for different types of companies. For example, companies with existing manufacturing supply chains may be tracking and reporting similar information about labor practices across their supply chains in other types of reports (ESG, Supply Chain Transparency, etc.). These types of companies might adapt their existing reporting infrastructure to include a section to report on their data enrichment practices. For other companies, it may make sense to include this same information in their model card.

We are hoping to get feedback on how this template is organized, the content of the template (whether these questions are comprehensive, phrased clearly, useful for enabling other actors to hold stakeholders accountable, etc.), and the feasibility of monitoring and reporting on these questions.

HOW WAS THIS CREATED?

The first iteration of this document was drafted by PAI based on PAI's white paper, "Data Enrichment Sourcing Guidelines," and input from PAI's community of practice on data enrichment practice.

NEXT STEPS FOR THIS RESOURCE

- Gather more input on these from the broader multistakeholder community during an open comment period and targeted feedback sessions with industry, civil society, human rights experts, supply chain experts, workers and their representatives, and researchers.
- Incorporate feedback/comments into a next version of this document.
- Include a refined version of this resource in PAI resource library for relevant actors across the supply chain to use.

CREDITING

In any future iterations of this document, we will include a section on who contributed to it and the settings in which feedback was collected.

Transparency Template

Team Level Questions to be Answered by Individual Development Teams

These five questions correspond to PAI's five Data Enrichment Sourcing Guidelines.

1 **How did you determine the pay for *data enrichment workers*[1] for this project? Is it at least above a living wage?** PLEASE REFER TO GUIDELINE 1

A. Are you sure workers will be paid in cash?
B. Are workers being paid by the task or by time? How are you calculating the rate?
C. What resource did you use to determine living wage?
D. Are you accounting for time spent training, reading instructions, and any other time needed to find, complete, and review work completed?
E. Are workers paid for all work completed?

2 **Do you plan to run a pilot for this project?** PLEASE REFER TO GUIDELINE 2

A. What are you testing for during the pilot?
B. How are you making adjustments to the project based on what you learn during the pilot?

3 **What skills and background are needed to complete the tasks for this data enrichment project?** PLEASE REFER TO GUIDELINE 3

A. How are you identifying workers to match that?
B. Do these tasks require a high level of domain knowledge? If so, are you thinking about retaining the workers you identify?

4 **Are you providing instructions and training materials for these data enrichment tasks?** PLEASE REFER TO GUIDELINE 4

A. How will you test these instructions / training materials?
B. Will you provide examples of correctly and incorrectly completed tasks?
C. Have you validated your instructions/training against the Good Instructions Checklist? Please highlight any divergences and why.

5 **What communication mechanisms have you put in place between your team and data enrichment workers?** PLEASE REFER TO GUIDELINE 5

A. What is the regular communication cadence with workers?
B. How can workers contact you to clarify instructions, raise questions, contest rejected work, etc.?
C. Please detail how you are collecting feedback from workers at the end of a data enrichment project?

1 **Data enrichment work:** Data curation for the purposes of machine learning model development that requires human judgment and intelligence. This can include data preparation, cleaning, labeling, and human review of algorithmic outputs, sometimes performed in real time.

Examples of data enrichment work: Data preparation, annotation, cleaning, validation, intent recognition, sentiment tagging, image labeling, human review of algorithmic outputs such as content moderation, validation of low confidence algorithmic predictions, speech to text error correction, red teaming.

Organization-Level Questions for Transparency Reporting

For the following questions, please highlight any relevant variances in policies and practices for how workers might be treated for different types of models (e.g. direct hires vs. contractors, or via managed service vs. via a platform)

1 **What policies do you have for how data enrichment workers are treated for your organization's data enrichment projects? Please describe your policies on the following:**

A. What are your policies on how workers are paid for data enrichment projects and how do you ensure that these are met?
   - Does this differ across different contract terms (e.g. direct hires vs. contractors vs. via third party)?

B. Do your teams consistently run pilots to create appropriate baselines and test the experience for workers?
   - Please describe how often pilots are not run and describe the criteria for projects not needing pilots.

C. What policies are in place to ensure that teams are recruiting appropriate workers for a task (i.e. matching the skills of workers to the tasks)

D. What policies are in place to ensure that instructions and training materials are tested and provided for data enrichment tasks?
   - Please describe how often instructions/training materials are not tested or provided and the criteria for projects where this is the case

E. What are the communication mechanisms have you put in place between your team and data enrichment workers?
   - Do you track response time to worker queries and collect worker feedback over time? How do you act on these results?

2 **Describe how you ensure that workers are being treated in accordance with those policies:**

A. How do you ensure that teams setting up data enrichment projects are following the above policies?
   - Who is involved with establishing that?

B. Please describe what steps you take to work with any vendors involved with managing data enrichment workers to ensure that impact on workers is monitored in accordance with the above policies.

3 **Based on the total number of workers involved with enriching data:**

A. What percentage are directly hired by your company?
   - What are their employment terms?

B. What percentage are contracted?
   - What are their employment terms?
   - Who takes on the cost of acquiring any tools needed to complete the work? (e.g. computers)

C. What percentage come through a *managed service provider*?[2]

D. What percentage come through a *platform*?[3]

4 **Do you have different policies for navigating how workers are treated across different contract terms? (e.g. how do terms differ across direct hire, contractors, workers through managed service, workers through platforms?) If so, please describe.**

A. On platforms if a worker is flagged for suspension, is there an appeal process?

5 **What geographic locations do your workers comes from?**

A. Are there any different policies in place for workers in different locations?
B. How do you ensure you are complying with local labor laws for workers in various locales?

6 **What resources are available for data enrichment workers interacting with traumatizing content?**

A. What policies/process does your company follow to support the mental well-being of DE workers intentionally/unintentionally exposed to traumatizing content?
B. How does your company define traumatizing content that would qualify workers to receive additional support services?
C. What purpose does data enrichment of traumatizing content serve to your company/ product?

7 **What information is collected about your workers and what precautions are taken to protect this data?**

8 **When working with your vendors/other actors in the supply chain, did you identify any additional issues with how workers were being treated?**

A. Did you identify additional harms that workers were subject to over the course of data enrichment projects?
B. What steps did you take to remedy those harms?

2 **Managed service provider:** Managed service providers can work in a variety of configurations including employing an in-house team, working with a set of subcontractors, or even setting up tasks on crowdworking platforms on behalf of clients. Depending on the configuration of the service, workers can be full-time employees, consultants, or independent contractors. Managed service providers typically support their clients in developing and refining instructions and task design, monitoring quality, and determining the price for the work.

3 **Platform:** Crowdsourcing platforms act as an intermediary for task-based work, connecting clients and workers. Crowdsourcing platforms can have curated workforces or may be open for anyone to join. Some platforms provide clients with the ability to work with a "private crowd" specifically assembled for the duration of the project. Others provide application programming interfaces (APIs) which allow clients to customize the platform's core functionality to meet their unique needs. This model can be considered as a sub-segment of what is often referred to as the "gig economy," or "platform economy." While there are platforms that are fully dedicated to providing data enrichment work, tasks such as data labeling are also frequently done on platforms that offer other kinds of task-based work.