# ACF Tech Data Surge Team Project Scope Document

## Project Title

CCDF Data Aggregation and Normalization

## Project Overview

Goal: Gather, aggregate, and normalize data about entities that receive Child Care and Development Fund (CCDF) funding through state block grants, with a focus on collecting business names and addresses.

The Administration for Children and Families (ACF) provides substantial funding to states through CCDF block grants. Due to the nature of block grants, states have flexibility in how they distribute these funds to childcare providers, resulting in limited visibility for ACF regarding the end recipients of funding. This project aims to address this gap by collecting data on CCDF-funded childcare providers across all states and territories.

This information will serve dual purposes: 1. Enhance accountability by improving visibility into where CCDF funds are ultimately distributed 2. Support disaster response planning by identifying childcare resources in potential disaster areas, helping to prevent the creation of “childcare deserts” during emergency situations

## Objectives

1. Assess data availability across state CCDF systems
   1. Review each state’s CCDF plan website and related resources
   2. Identify states with accessible data and those requiring alternative approaches
   3. Document data formats, accessibility methods, and completeness for each state
2. Develop and implement data collection strategies
   1. Create web scraping solutions for states with online data
   2. Establish alternative collection methods for states with limited online presence
   3. Standardize collection approaches where possible
3. Aggregate and normalize collected data
   1. Create a unified data structure for provider information
   2. Normalize address formats and other key data points
   3. Implement data quality checks and validation
4. Develop visualization and distribution tools
   1. Create an interactive map visualization of provider locations
   2. Produce exportable datasets in standard formats
   3. Document data dictionary and limitations
5. Establish a framework for periodic data updates
   1. Document collection processes for repeatability
   2. Recommend update frequency and maintenance approaches

## Scope and Approach

This project will focus on collecting business names and addresses of CCDF-funded childcare providers across all states and territories. The approach will vary by state based on data availability:

* **Web Data Collection:** For states with interactive maps or downloadable datasets, automated collection methods will be developed using Python-based web scraping tools.
* **Manual Research:** For states with limited online data, manual research and direct outreach may be necessary, though this will be limited in scope for the initial project phase.
* **Data Processing Pipeline:** Once collected, data will be processed through a standardized pipeline that will:
  + Clean and validate addresses and business names
  + Normalize data formats across different state sources
  + Geocode locations for mapping purposes
  + Identify and flag potential duplicates or errors
* **Visualization:** The processed data will be visualized through:
  + An interactive map showing provider locations
    - Mapping will be done with Tableau or [Plotly](https://plotly.com/python/tile-scatter-maps/) (Python)
  + Summary statistics by state/region
  + Exportable datasets for further analysis

## Partnership Model

**Sprint syncs:** Held weekly throughout the project duration.

**Working sessions:** Held weekly with the Data Surge Team and other stakeholders as needed.

**Asynchronous clarifications:** An MS Teams channel will be established with the members of the Data Surge Team and other relevant stakeholders. Teams will be used for informal communication. Email will be used for more formal communication and deliverables.

## Deliverables

* **Data Availability Assessment:** Documentation of data availability and access methods for all states/territories
* **Collection Code Repository:** Python-based tools for data collection and processing
* **Aggregated Dataset:** Comprehensive CSV file containing normalized provider data
* **Interactive Map:** Visualization tool showing the distribution of CCDF-funded providers
* **Documentation:** Technical documentation detailing the collection methods, data structure, and update processes
* **Update Framework:** Recommendations for maintaining data currency, including suggested frequency and methods

## Timeline

These timelines represent initial plans. Tasks may adjust based on findings from earlier weeks.

### Financial Data

* **Sprint 0: March 17-28, 2025**
  + Complete initial assessment of data availability across all state CCDF plans
  + Categorize states by data accessibility
  + Begin developing collection strategies for states with readily accessible data
  + Initial project sync with stakeholders
* **Sprint 1: March 31-April 11, 2025**
  + Complete data collection from states with accessible online data
  + Begin data normalization and cleaning
  + Develop initial data structure for aggregated dataset
  + Begin outreach to states with limited online data
* **Sprint 2: April 14-25 2025**
  + Continue data collection from remaining states
  + Implement data validation and quality checks
  + Begin development of visualization tools
  + Document collection methods and challenges
* **Sprint 3: April 28-May 1, 2025**
  + Finalize data collection and aggregation
  + Complete interactive map and visualization tools
  + Document data limitations and gaps
  + Develop recommendations for ongoing data maintenance
  + Final project review with stakeholders

## Assumptions

* State CCDF plans and related websites will remain accessible during the project period
* Most states have some form of provider data available, though the format and accessibility may vary
* Python and related data processing tools will be sufficient for the technical requirements
* The project scope is limited to data collection and organization
* Initial development of visualization tools will use open-source technologies due to potential delays in Tableau access

## Roles

* Governement Lead: Jane Yang
* Project Owner:
* Project Lead: Daina Bouquin
* Project Team:
  + When appropriate we will contact the Office of Planning, Research, and Evaluation and the Office of Childcare

## Future Considerations

If this initial data collection effort proves valuable, potential follow-on work could include: - Development of a public-facing dashboard - Integration with disaster planning tools - Expansion to include additional provider information beyond location data - Establishment of automated data refresh processes - Creation of APIs for programmatic data access