Crafting a Sustainable Regional Food System Inventory: a guide for collecting, managing, and maintaining food systems inventory data

Laura Valentine and Gillian Revenis

The Center for Regional Agriculture, Food, and Transformation (CRAFT)

Chatham University

Pittsburgh, PA

craft@chatham.edu

Presented at:

2023 ASFS-AFHVS Conference "Knowing Food"

May 31-June 3, 2023

Summary

In 2018, the Center for Regional Agriculture, Food, and Transformation (CRAFT) at Chatham University received funding to create a regional food system inventory, collecting data about the Pittsburgh food shed defined for the purposes of the project as a 200-mile radius from downtown Pittsburgh. This project subsequently formed the foundation for our Regional Food System Inventory of Pennsylvania, Ohio, and West Virginia, which has supported our region's farmers, food businesses, and policymakers since.

CRAFT assessed and revised the methodology utilized for this initial project in an effort to create a sustainable, updatable, food system inventory that could continue to provide this information long-term. Such an inventory has the potential to connect food and farm businesses across the value chain and underpin regional research that shapes more effective policies and programming for a more robust regional food system. We have developed a flexible methodology for sourcing, aggregating, and cleaning data about regional producers, processors, aggregators, and outlets that we believe to be feasible and sustainable for organizations of varying capacity.

Our methodology prioritizes producing a final data set that can be used for supporting smaller producers, grant funding efforts, connecting to farms, food systems analysis, and finding products or services, as well as decreasing the amount of time needed for data preparation. The resulting entity-level data can be used to support decision-making, understanding regional food systems, mapping, and outreach.

Inventories developed using our regional food systems inventory methodology can broaden the availability of useful data to food organizations. For example, someone working on developing a grain network could use this kind of data to create a map of regional independent mills or reach out to farmers growing heritage grains to better understand their needs.

This methodology is designed to be adaptable to changing conditions, flexible, and uses commonly-available tools to clean and collate data. In this paper, we review the general method, explore challenges, and offer lessons learned from our regional food system inventories of Pennsylvania, Ohio, and West Virginia, as well as the Mississippi Delta region.

History & Approach

In 2018, the Center for Regional Agriculture, Food, and Transformation (CRAFT) at Chatham University received funding to create a regional food system inventory, collecting data about the Pittsburgh food shed (defined for the purposes of the project as a 200-mile radius from downtown Pittsburgh). This project required partial data from six states, and creation of the inventory was time- and labor-intensive. The resulting inventory was originally intended as a one-off, but its utility to food systems professionals quickly became apparent.

In order to create a sustainable process and a more useful data set, CRAFT reviewed the pain points of the project, developed plans to address those issues, interviewed people who had used the data set, and surveyed potential users. CRAFT's interviews of those who used the initial inventory revealed that the most common use of the inventory was to support smaller producers, followed by grant funding efforts, connecting to farms, food system analysis, and finding products or services. The assessment also revealed that people who used the data were most often interested in the information about producers, outlets, and aggregators, followed in order by processors, other support entities, and agricultural input providers, and in more public-facing resources, increasing specificity, updating more regularly, clarifying data sources, and making data available in various formats.

Major pain points included the difficulty in locating source data, and the time-intensive methods of preparing the data and organizing it into a final, useful data set. CRAFT revised source selection methods, analyzed tools, and developed processes for addressing these pain points.

The original inventory and revised process formed the foundation for our Regional Food System Inventory of Pennsylvania, Ohio, and West Virginia, which now supports our region’s farmers, food businesses, and policymakers. Since that time, CRAFT has adapted its methodology to create a similar inventory for the Mississippi Delta Region. A final inventory is a set of entity-level data for food system businesses of various types, including business names, addresses, geospatial information, and additional information about the type of production or operation as available. Our general method is: source selection, data collection, data preparation, categorization, final checking, and publication. We also created a data management plan, developed internal documentation, and a public methodology.

Stages of developing an aggregated regional food systems inventory data set.

Challenges

Most challenges faced in developing a food systems inventory fall into three categories: sourcing, data issues, and technical challenges.

Sourcing

The overwhelming lack of a single, authoritative source for producers, processors, aggregators, and outlets requires careful consideration of data sources. Selecting dependable sources is the best way to ensure the long-term viability of an inventory. When selecting data sources for a regional food systems inventory, considerations include:

- **Entity types:** what sort of functional entity should be captured. We chose four entity types based on their prevalence and structural importance to the larger food system, but appropriate breakdowns differ based on regional needs. CRAFT captures data on producers, processors, aggregators, and outlets. Of these, data on producers tends to be the most difficult to work with. Some businesses may be more than one type (e.g., a farm may both produce food and do on-farm value-added processing). The four entity types are defined as follows:
    - Producer: An organization or individual engaged directly in the growing of food crops or in livestock-based food farming (raising animals for meat, eggs, dairy, honey, etc.).
    - Processor: Entities who buy raw products and increase their value by processing them. Includes food manufacturing facilities, community kitchens and incubators, co-packers, packing sheds, butchers, bakeries, etc.
    - Aggregator: Businesses that collect and combine farm products from multiple sources and provide them to wholesalers or retailers. These entities assist

producers in bringing products to market by providing a connection point to larger markets. Aggregators may provide other services, such as co-packing.

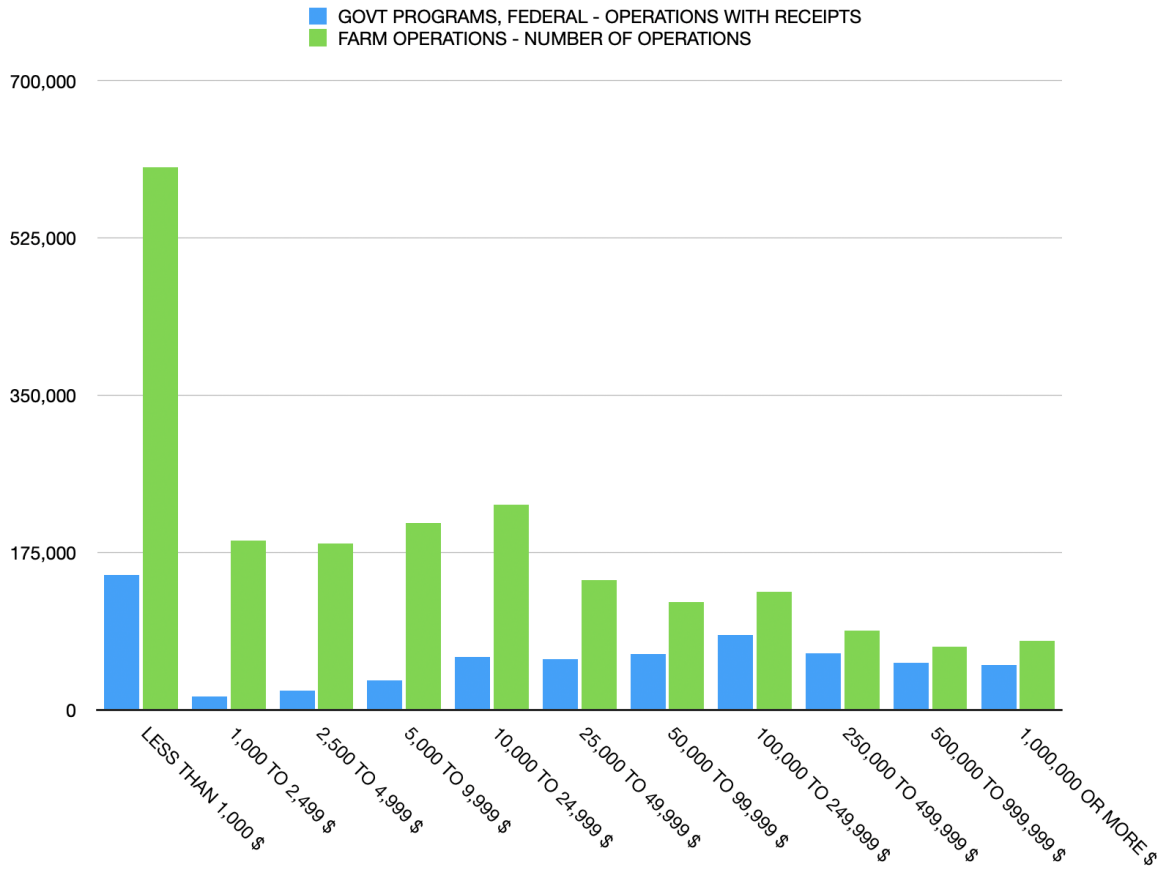
    - **Outlet:** Food retailers that may or may not purchase directly from local producers, depending on the needs of the inventory.

- **Source availability and reliability over time:** what data sets are available, and how can they be obtained? Is a currently-available source regularly updated? Is it likely to continue being updated in the future? Can it be updated in the future?

After deciding on entity types, we conducted a broad search of sources, including all of those from the original inventory plus anything additional we could find. Our searches included not just looking for data sets, but also looking for maps, directories, organizations involved with farmers' markets or with advocacy, and more. We reviewed the information we found, identifying who collected it and how, whether it was original or sourced from elsewhere, its most recent date of update, and any plans for maintaining it into the future.

We determined through this process that appropriate data sources can be formal or informal. A formal source might be something like a county health department that updates weekly, has a status flag indicating if a business is in operation or has closed permanently, and the source is inspections done by the department itself. An informal source could be a flier put out by the manager of a farmers' market at the start of every season, listing the farms selling at the market. Informal sources can be quite reliable – consider a market that has lasted for decades and always has a list of vendors available – but they can be more difficult to locate, and often are missing information such as addresses.

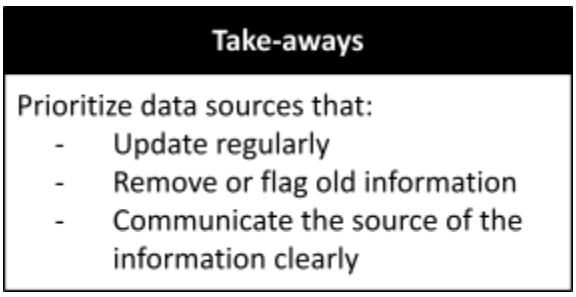
These informal sources are often very important, particularly for information on producers, which may be absent from any formal source. The single most complete set of publicly-available, entity-level data on producers is the USDA Farm Subsidy data. There are a lot of ways in which it is tempting to take this data set and use it as is – and there are legitimate uses for it that way. The Environmental Working Group uses it to show where subsidy money is going, and for what, over time. However, "farms that receive subsidy payments" is an incomplete picture of farming in the United States. Many farms are missing from this data set entirely, as can be seen by comparing the number of farms receiving subsidy payments to the number of total farms in the Census of Agriculture. The only way to capture any of those other farms is to use multiple sources, including informal ones.



Understanding how the data is maintained is equally important. It is very common to find exciting-looking datasets where someone did a project, but after finalization, the data is never updated again. For example, in our most recent source search for our inventory, we found what looked like a great farm data set in work done at another university. When we asked how they got that data, we were told it had been provided by the Pennsylvania Department of Agriculture – and when we asked the Department of Agriculture about it, it turned out to be a filtered portion of our original inventory. So, not something that they were collecting and maintaining, but something they had around because of a previous project, which was already a few years out of date. It can be very exciting to find data that looks exactly what is needed, but we asked ourselves questions about it, including:

- When was it last updated?
- Are updates on a regular schedule that can be identified?
- When was it last purged (or is there a way to filter out old information)?
- Where did it come from in the first place?

Data available for food systems is limited and often of poor quality, especially for producers. Better data can often be purchased from trade groups, but this is not always financially feasible. In many cases, data is available for download by anyone. In other situations, it must be requested via Freedom of Information Act (FOIA) request, Right to Know (RTK) law request, or other data request. Compliance with these requests may vary by state and may be contingent on payment per page or per entity fees. In some cases, we enter data manually from sources such as scanned, typewritten directories. Informal sources can require structured searches for farmer's market lists, producer affinity organizations, and so on.



We have opted to use only publicly-available, entity-level data to create our data sets, and selected data sets with clear inclusion criteria and a history of regular updates. Because of the limitations of the source data, the final data set may not be appropriate for statistical analysis.

Data Issues

Once sources have been identified and data collected, the data itself can present a number of challenges. Data may be missing, incomplete, contain errata and duplicates, and/or be formatted in ways that make it difficult to work with. Our data preparation process includes cleaning (removing errors, duplicates, wrongly included entities, etc), aggregating, transforming it to fit various schemas, and other similar tasks. Preparation is tedious, time-consuming, and often the most resource-intensive part of the process. We focused our efforts in this area on pain points in that original inventory: very large data sets from the USDA, detecting duplicates, and organizing the information in such a way as to enable easy searching and mapping. The aim was to reduce reliance on using human eyes to process the data. Tools available to do this kind of processing include text editors, databases, scripts, UNIX command-line tools, spreadsheet filters, and, of course, human eyes.

Preparation is also *destructive*: it alters and removes information from the set. Our procedure is to always use a working copy; even with automatic saving and version control, it is surprisingly easy to lose the last-known-good version of a data set.

Missing data refers to the categories of data that are absent from the set, or which you know or suspect to be underrepresented in the sample for structural reasons, such as cottage food producers in places where such businesses do not require licensing. Missing data is best addressed through thorough investigation of sources before collection, and by understanding what sorts of entities are likely to be missing. For example, farms which don't a) receive

subsidies b) participate in local food economies or c) have on-farm processing are often going to be difficult to acquire, even through informal sources.

Incomplete data is data where some information that we need or want for the final set does not come with the source data by default, such as location information or products offered. Many entities have minimal online presence, contact information, or information about their goods and services. We address incomplete data with a variety of methods, including by geocoding, identifying duplicates, and using spreadsheet formulas to create best-guess filtering.

Geocoding is vulnerable to errors in the underlying data, such as typos or missing location information, PO boxes instead of street addresses, formatting problems, and errors in the tool's own database. Unless the data set is very clean to start with, we don't expect to be able to geocode everything, but we aim to be as complete as possible. Using multiple tools increases our success rate. Because of this vulnerability, we geocode late in our cleaning process.

To reduce errors in the final set, we use a de-duplicating process with several steps, including vocabulary standardization, matching on entity names and addresses to identify potential duplicates, and using fuzzy matching to catch unidentified duplicates. We also found that de-duplicating each original source, prior to combining data sets, made later de-duplication of the aggregated set simpler and faster.

Vocabulary standardization, completed before matching either exact or fuzzy duplicates, can reduce errors in de-duplication. Our source data often contains many variations on common abbreviations, for example, such as road, street, or avenue. Standardizing these terms allows more efficient and effective removal of duplicates later on in the cleaning process. We use a dictionary of replacements in order to reduce the variation that may occur, and this can be done either manually by searching for each replacement term, or by using a script. We use a small Python script, which allows for faster and more accurate processing, although it must be regularly altered due to differences in source data.

```
# Dictionary containing mapping of values to be replaced (on left) and replacement values (on right)
dictOfStrings = {'AVENUE AVE ': 'AVE,',
                 'BOULEVARD BLVD ': 'BLVD,',
                 'DRIVE DR ': 'DR,',
                 'HIGHWAY HWY ': 'HWY,',
                 'PIKE PIKE ': 'PIKE,',
                 'STREET ST ': 'ST,',
                 'ROAD RD ': 'RD,',
                 ' BLVD ': 'BLVD,',
                 ' DR ': 'DR,',
                 ' HIGHWAY ': 'HWY,',
                 ' HWY ': 'HWY,',
                 ' PIKE ': 'PIKE,',
                 ' ST ': 'ST,',
                 ' AVE ':' AVE,',
                 ' RD ':' RD,',
                 ' WAY ':' WAY,'
                }
```

Example of replacements from a version of our Python dictionary script, used for vocabulary standardization. This particular data set had repeated designations for the street suffix, and did not include commas between the street address and city name.

Take-aways

- Always use a working copy before you do anything destructive to the data
- Each source data set needs its own cleanup before you join the data
- Be able to understand and articulate what is missing or incomplete

Name and address matching allows basic identification of entities that are exact duplicates, or which may be exact duplicates. There are different strategies that can be appropriate for handling these, depending on the exact information that is needed. Removing duplicates is destructive, and should be done on a working data set, not the original. *Fuzzy matching* uses the Levenshtein distance (a measure of the number of single-character edits) between two addresses; this enables the identification of typographical errors (e.g. "POBox" instead of "PO Box") as well as addresses that are substantially similar and need to be checked by hand.

Data issues in general are where we run into confusion, and that confusion can persist throughout the whole process. Before publishing or using the data set, we always conduct final checks. These checks are designed to ensure that the data is ready to be used. Final checks can be very quick, or they can uncover errors that take considerable time to fix.

- De-duplicate again – among other things, it can help to make sure that nothing was accidentally merged that wasn't meant to be.
- Sort categorization columns and look for obvious errors, places where the formula did not fill properly, and so on.

- Check to make sure columns contain what they are supposed to contain – it is surprisingly easy to end up with information one cell off from where it should be, which can affect formula results and maps.
- Double-check geocoding. It may not be able to identify locations for everything, but the set of uncoded entities should be as small as possible.

Note on USDA Subsidy data

We developed special procedures to deal with our largest data set, the farm subsidy data acquired by the Freedom of Information Act from the USDA. Such data almost always provides the bulk of data on farmers in any region in the US, and includes names and mailing addresses of businesses and individuals who receive subsidies. Sometimes these subsidies go to a farm business, at the farm location, but many times they do not. Furthermore, multiple farming businesses may use the same mailing address. For example, farmers may operate two farms but only receive mail at one, or family members may each operate a farming business out of the same location. Cross-county farms, where the same farming operation covers land in multiple counties, may or may not be duplicated in the USDA data, depending on how the farmer received the subsidies. The aim of these cleaning decisions was to de-duplicate the data as much as possible while minimizing the chance of combining separate operations.

Some aspects of this are difficult to automate and time-consuming to perform by hand. Each particular case is different and it should be carefully considered whether or not this is needed. Our strategy in the past has been to perform eyes-on deduplication for the following situations:

- One address, same subsidy recipient, different counties: keep all. We have found that these are usually either cross-county farms or multiple locations sharing one address.
- One address, different subsidy recipient: We usually combine these when farm ownership patterns suggest it is most likely joint owners of one farm are receiving subsidies individually, e.g. “Jennifer R Karcher” and “Robert G Karcher” would combine into "Jennifer R Karcher and Robert G Karcher". Otherwise, we do not combine.

Technical Challenges

While our methodology lays out a general path, there is simply no start-to-end pipeline that fast-tracks this process without considerable manual effort. Creating the final data set requires decisions about information structure and organization, data storage and maintenance, and understanding of available tools.

Data issues feed technical challenges, and understanding the structure of each source set and the strengths and weaknesses of various tools is vital to resolving those data issues to the extent possible.

Structural and organizational choices include what parts of the data to make public or keep internal, and why, and whether or not to provide the ability to filter the data using additional columns not in the source data. Our outreach to data users indicated that such filters were important, and our prior process for these was manual, time-consuming, and burn-out inducing. This then becomes a technical challenge: how do you create columns that people can filter on, that are accurate enough to be useful, based on the information you already have?

To address this, we developed a categorization process. We started off by developing a set of keywords: what terms were relevant to a category like "Meat Processing"? What about "Produce"? We used these keywords to structure initial spreadsheet formulas, which searched columns such as the name, subsidy type (if applicable), and products or services offered, using COUNTIFs nested inside of IF(OR). These formulas return a YES or NO, and we then reviewed our results to find false positives and false negatives, altered the formulas, and repeated this process until we could no longer identify either readily.

M	N	O	P	Q	R
Bakery	**Dairy**	**Eggs**	**Grain**	**Poultry Slaught**	**Meat Slaughter**
NO	NO	NO	NO	NO	NO
YES	NO	NO	NO	NO	NO
YES	NO	NO	NO	NO	NO
NO	NO	NO	NO	NO	NO
YES	NO	NO	NO	NO	NO
YES	NO	NO	NO	NO	NO
YES	NO	NO	NO	NO	NO
NO	NO	NO	NO	NO	NO
NO	NO	NO	NO	NO	NO
NO	NO	NO	NO	NO	NO
YES	NO	NO	NO	NO	NO

Results of formulas in the 2022 Pennsylvania, Ohio, and West Virginia Processors Regional Food Systems Inventory data set.

This process is not 100% accurate. There is no way that we have been able to discover to classify data of this quality, with this many gaps, with 100% accuracy, even with many hours of human effort. The goal of the formula work was to approach what a human could do with the available information in a fraction of the time, and without the wear and tear and fatigue that this kind of categorization causes in humans.

Producers	Dairy	`=IF(OR(COUNTIF($V2,"*dairy*"),COUNTIF($V2,"*milk*"),COUNTIF($V2,"*cheese*"),COUNTIF($AJ2,"*DAIRY MARGIN*"),COUNTIF($A2,"*DAIRY*")),"YES","NO")`
Producers	Grain	`=IF(OR(COUNTIF($V2,"*grain*"),COUNTIF($V2,"*wheat*"),COUNTIF($V2,"*barley*"),COUNTIF($V2,"*rye*"),COUNTIF($V2,"*oat*"),COUNTIF($V2,"*buckw*"),COUNTIF($V2,"*teff*"),COUNTIF($V2,"*spelt*"),COUNTIF($V2,"*emmer*"),COUNTIF($V2,"*soy*"),COUNTIF($V2,"*sorghum*")),"YES","NO")`
Producers	Eggs	`=IF(OR(COUNTIF($V2,"*egg*"),COUNTIF($V2,"*laying*"),COUNTIF($V2,"*layer*")),"YES","NO")`

Example formulas for sorting subcategories of producers. These examples are from a one-time inventory with custom categories.

We use data management planning to address the storage and maintenance decisions: what we plan to retain, and for how long, and in what formats. Fortunately, because we are not keeping data such as continuous monitoring data, which can get very large, the available technical solutions are reasonably simple. We use DMPTool (https://dmptool.org/), a free, open-source online application, to create our data management plan.

Lessons Learned

Intentionality Matters

CRAFT's first regional inventory was created under circumstances that did not allow us to be as intentional as we would have liked. It was created for a specific audience and under time pressure. Taking a step back and thinking critically about our goals, taking the time to understand the use cases, investigating the complex of problems around data sources, and analyzing the available tools allowed us to develop a process that works for us and is much less intensive to complete.

Consider Regional Distinctions

CRAFT is based in western Pennsylvania, and our work has been focused on Pennsylvania, Ohio, and West Virginia. When we formulated a food system inventory for the Mississippi Delta Region, including select counties in Mississippi, Arkansas, and Tennessee, we found we needed to adapt our approach.

First, the places in which one requests public data will vary by region. No list of sources we can give could be complete and accurate across all locations. States, counties, and other regional entities collect data differently and have different regulations governing data. Furthermore, local or regional non-government organizations may or may not have relevant data that they may or may not be willing or able to share. If you are working in a multi-state area, you may find state governments are not obligated to provide information to non-state residents; we encountered this when working on the Mississippi Delta inventory.

Lastly, it is simply easier to create these inventories if you have some regional knowledge. Important crops or categories of entity can vary, and local organizations and contacts can be useful. It's possible to create these inventories without that regional knowledge, but it requires more time and research.

Document Everything

Developing this process involved many consequential decisions, and we suspected, going in, that we would have to make choices that would not seem obvious in retrospect. We made the decision to write our decisions down, and keep track. Often, we would go down a path, run into a stumbling block, and ask ourselves why we'd made that choice in the first place – being able to look back at why was invaluable.

Our documentation includes a public methodology, an internal source research document, reports on our outreach to data users and potential data users, a technical manual that contains detailed instructions and copies of our scripts and formulas, and a data management plan. All of these pieces enable us to explain why we made the choices we did, reliably repeat our food systems inventory for our region, and adapt it to other regions if necessary.

Resources

Our public methodology is available on the CRAFT website, at http://craft.chatham.edu/data.

We use DMPTool (https://dmptool.org/), a free, open-source online application, to create our data management plan.

For fuzzy matching, we use the Google Sheets add-on Find Fuzzy Matches by Ablebits.

Geocoding Tools

Tool	Pros	Cons
Geocode for Awesome Table (Google Sheets)	- Easy to use - Concatenating data is the only structural change needed	- Can only process a limited number of addresses per day - Does not handle address variability well
Excel Geography data type	- Available with recent versions of Excel - Good with incomplete data - Provides some fuzzy matching capability - No limit on number of addresses	- Requires data restructuring - Requires careful processes to avoid altering original data
US Census Geocoder	- Very fast - Good with incomplete data - No limit on number of addresses - Provides some fuzzy matching capability - API support	- Requires data restructuring - Need to set up separate working files, process them, and then re-integrate

Example Data Sources

Entity type	Sources
Producers	FOIA request to USDA for farms receiving subsidies. LocalHarvest data download, courtesy LocalHarvest (farms and farm markets) USDA Organic Integrity Database FarmFresh WV Local Food Directories: National Farmer's Market Directory (USDA)
Processors	State Departments of Agriculture Local Harvest USDA Meat, Poultry, and Egg Inspection Directory Meat & Poultry Inspection programs
Aggregators	State Departments of Agriculture
Outlets	State Departments of Agriculture Local Food Directories: National Farmer's Market Directory (USDA) State and county Departments of Health SNAP Retailer Locator

Example Python Dictionary Script

This is a version of the dictionary script we use for cleaning; in this case, the source data contained what is likely to be a database conversion error that resulted in duplicate street abbreviations, and was formatted to combine parts of the address that we wanted to be in

separate columns. The dictionary script ran through a copy of the file and replaced the errors in the abbreviations, then added the missing commas. A copy of this python script is available upon request.

```python
# imports a library that lets us use regular expressions to search
import re

# opens the file to clean in read-only mode. use a csv file for this always
# because the dictionary uses the commas to help it search
data = open("[pathtofile].csv","r")

# reads the contents of the file into a string
strValue = data.read()

# closes the file
data.close()

# Dictionary containing mapping of values to be replaced (on left) and replacement values (on right)
dictOfStrings = {'AVENUE AVE ': 'AVE,',
                 'BOULEVARD BLVD ': 'BLVD,',
                 'DRIVE DR ': 'DR,',
                 'HIGHWAY HWY ': 'HWY,',
                 'PIKE PIKE ': 'PIKE,',
                 'STREET ST ': 'ST,',
                 'ROAD RD ': 'RD,',
                 ' BLVD ': 'BLVD,',
                 ' DR ': 'DR,',
                 ' HIGHWAY ': 'HWY,',
                 ' HWY ': 'HWY,',
                 ' PIKE ': 'PIKE,',
                 ' ST ': 'ST,',
                 ' AVE ':' AVE,',
                 ' RD ':' RD,',
                 ' WAY ':' WAY,'
                 }

# Iterate over all key-value pairs in dict and replace each key by the value in the string
for word, replacement in dictOfStrings.items():
    strValue = re.sub(word, replacement, strValue)

# prints the new version of strValue to the computer terminal screen.
# Should probably replace this with feeding it directly to a file but right now
# I am using the shell command "cat" to take the output and put it into a file instead
print(strValue)

```

Example Categorization Columns for Filtering

We have developed spreadsheet formulas and keyword vocabularies to support the creation of all of these columns. Copies of formulas are available upon request, but may be region-specific and should not be used as-is.

Entity type	Filter Columns
Producer	Produce, Meat, Dairy, Grain, Eggs, Aquaculture, Livestock, Misc, Agritourism, CSA, U-pick, Events, Education, Online Retail, Direct to Consumer, Wholesale
Processor	Bakery, Dairy, Eggs, Grain Mill, Poultry Slaughter, Meat Slaughter, Poultry Processing, Seafood, Candy/Snacks, Honey, Beverage, Other Manufacturing, National Chain
Aggregator	Distributor, Warehouse
Outlet	Farmers' Market, Farm Store/Stand, Grocery Store, Convenience Store, National Chain, SNAP Retailer