**Curating the *NASS Bee Colony and Honey* *1974-2020* Dataset**

Beth Mawhinney

University of Denver, Library and Information Science

LIS 4220: Data Curation

Professor Matt Mayernik

March 16, 2021

***Project Description:***

This data curation project aims to review and assess the *NASS Bee Colony and Honey 1974-2020* dataset, describing the dataset, providing feedback for potential improvements of data and metadata, proposing an appropriate repository for deposit, and considerations for citation, long term preservation, and copyright.

**Data and Metadata Profile:**

NASS Bee Colony and Honey 1974-2020 (Version 2)

Combined from 2 NASS datasets into cleaned up json/csv

https://www.kaggle.com/kenshoimpactteam/nassbeecolonyandhoney

***Description:***

This data and metadata profile serves as a review of the *NASS Bee Colony and Honey 1974-2020* dataset, intending to identify the data, metadata, and room for possible improvements of the dataset.

***Data:***

The *NASS Bee Colony and Honey 1974-2020* dataset contains honey production and bee colony data across different states in the US, annually. The dataset combines data from two initial datasets published by The United States Department of Agriculture (USDA) National Agriculture Statistics Service (NASS). The authors have combined the original data into three new datasets, containing statistics in regard to honey production and bee colony numbers that are meant to be cross referenced. The first dataset is titled “honey” and contains data that details the amount of honey produced by producing bee colonies, by state, and the value of the yields dating from the years 1989-2019. The second data file is titled “numbers” and contains data of increases, decreases, and changes to bee colonies quarterly, by state, from the years 2015 to 2018. The final dataset is titled “stressors” and contains data of factors that affect changes in bee colony numbers, by state, quarterly from the years 2015 to 2018. This is the second version of the dataset, however there are no files provided containing the original version, or any directions on how to find or access the original version, making it unclear what changes have been made or how the dataset as been updated.

The original datasets from which the *NASS Bee Colony and Honey 1974-2020* dataset is derived from were published as part of two annual NASS reports titled, *Honey* and *Honey Bee Colonies*, that provide information for changes in honey bee colonies, honey production, honey prices, and honey value. Key stakeholders of the dataset may be the owner and authors of the dataset, USDA, NASS, honey farmers and beekeepers, honey importers and buyers, and stockholders.

The dataset contains six files, three CSV files and three JSON files. The CSV files each contain one spreadsheet either “honey”, “stressors”, or “numbers”; the JSON files contain the same data/information formatted in JSON. There are no usage restrictions or specific licensing for this dataset; it is in the public domain. There is not software required to open or analyze the data files. The CSV files can be fully viewed online through Kaggle or downloaded and opened on a desktop. The JSON files require to be downloaded in order to view them; once downloaded a new web page is automatically opened containing that data from the JSON file and is completely viewable.

***Metadata:***

The dataset comes with limited metadata; the Kaggle Metadata tab is completed for this dataset including sections for Usage Information, Provenance, Maintainers, Authors, Updates, and Data Coverage. A brief statement of description is provided on the main data tab along with two keyword tags, biology and agriculture. There is no additional README provided with this dataset. A brief description of what data was intentionally excluded from the dataset due to the formatting of original documents and ease of processing that information, along with a note intending to process the missing data in the future are included. Links are provided to two NASS publications that contain the original datasets which were used and combined for the new dataset. The limited metadata is housed on the website metadata tab, except the description which can be found on the data tab. The data files provided do not contain any further description or metadata.

The metadata does not seem comprehensive. The title and the description of the data provides only a vague understanding of what the data is or why it would be chosen for use, only referencing that the dataset is concerned with honey and bee colonies. Once the dataset files are reviewed it becomes clear that the dataset is concerned specifically with honey production and outside factors that affect production. The metadata was structured as recommended by Kaggle providing information for the categories: Usage Information, Provenance, Maintainers, Authors, Updates, Data Coverage, but does not follow any specific metadata standards or schema.

***Possible Improvements:***

To improve on the data and metadata, a few considerations should be taken. Including descriptions of the datasets as part of the dataset files would provide more ease to interpreting that data, as well as more information on why the data was compiled or the intended use of the dataset. The links provided for the original two NASS datasets provide useful descriptions; consideration for adopting some of the information from these descriptions should be taken. Transferring some applicable information from the original datasets would support users more quickly understanding why the data was collected and how they might want to use it. Additionally, the tags for this dataset are limited to biology and agriculture, which do not seem to capture the main ideas or scope of this dataset. Again, it may be beneficial to adopt some of the tags from the original datasets such as, honey, productions, or bee colonies. Making these small additions may allow for better discoverability of this dataset. Consideration should be taken on providing information on the reason for conjoining these datasets and what further use or benefit for users the authors believed these changes would have, rather than just using the original datasets.

The data files are missing units of measurement in the column description, leaving the data unclear and ambiguous to users. As the data is compiled from other NASS datasets and the authors have provided the links to those publications, users are able to review the original datasets to identify the units of measurement, however this is inefficient and ultimately still unclear. The dataset will be more complete and usable if the units of measurement are updated.

A potential issue seems to be that multiple time frames for the data are indicated. In the title it states that the data will span from 1974-2020, however the metadata indicates that data coverage is from Jan. 1st, 2001 through Dec. 31st, 2019; there is no data that dates back to 1974. The “honey” files cover data from the years 1989-2019, but the “numbers” and “stressors” files cover data from the years 2015-2018. Reassessment of the time frames actually covered is needed and updates to the metadata and description will be necessary. Consideration to change the title should be taken; a potential new title may be, *NASS Bee Colony and Honey Production 1989-2019 with Compounding Factors*. Additionally, the data’s intended use seems to be for identifying how changes to amounts of bee colonies and stressors relate to and affect honey production stats for each year, but any information for numbers and stressors is missing prior to 2015 and after 2018. For those years, users will be unable to use the data as it seems to be intended. More information on the datasets’ value or potential use when used independently or in tandem may enlighten users on how the dataset can be used.

***Publications:***

There are no publications provided with the dataset. A general search using the DU library databases was completed to identify any publications that referenced or cited the dataset, none were found. A google search was completed to identify if the dataset was referenced or cited within anything published on the internet, none were found. At this time there are no known publications containing or referencing the dataset, beyond Kaggle.

**Repository Profile:**

United States Department of Agriculture National Agricultural Statistics Service

https://www.nass.usda.gov/

***Description:***

This repository profile is to serve as a review of the United States Department of Agriculture (USDA) National Agricultural Statistics Service (NASS) and provide explanation as to why USDA NASS would serve as an appropriate repository for deposit of the *NASS Bee Colony and Honey 1974-2020* dataset.

***Why NASS:***

In reviewing potential repositories to deposit the *NASS Bee Colony and Honey 1974-2020* dataset the USDA NASS repository was found to be appropriate for several reasons. First, the original datasets that were combined to create the *NASS Bee Colony and Honey 1974-2020* dataset are published in reports that used NASS data and can be found in the NASS repository. Making this dataset available to users where they can easily access and compare the original data with the new dataset, will make use of the data more efficient and functional. Next, the *NASS Bee Colony and Honey 1974-2020* dataset will appeal to audiences and users of NASS. “NASS reports include production and supplies of food and fiber, prices paid and received by farmers, farm labor and wages, farm finances, chemical use, and changes in the demographics of U.S. producers,” (United States Department of Agriculture (USDA) National Agricultural Statistics Services (NASS), 2021, p. 1). The *NASS Bee Colony and Honey 1974-2020* dataset contains honey production and bee colony data across different states in the US, annually and is directly related to beekeeping, honey farming, honey production, and the value of honey. The data falls easily within the domain of data that NASS collects and manages.

The mission statement of NASS is “The National Agricultural Statistics Service provides timely, accurate, and useful statistics in service to U.S. agriculture,” (USDA NASS, 2018, p.1).When reviewing NASS’s core values there is an emphasis on providing accurate and timely data to serve communities such as rural farmers that rely on the data for their livelihoods. Although there is no explicit description stating the intended use, the *NASS Bee Colony and Honey 1974-2020* dataset is directly related to agriculture and the people and communities doing that work, further making NASS an appropriate repository for the dataset. There is a clear and easy way to address any data issues and request changes ensuring that the quality of the data can be publicly audited, creating accountability for the quality of the data.

***Data Submissions to NASS:***

NASS is a closed repository, containing data gathered for the Census of Agriculture, which is conducted every five years, and special surveys which are reoccurring on a weekly, monthly, quarterly, or annual basis depending on need. No submissions from the public are accepted. The last census was completed in 2017. Data providers that contribute to the census and special surveys, include “large ag businesses, small-to-medium sized family-owned operations, and niche farms and ranches, whether rural or urban farmers,” (USDA NASS, 2020, p. 4). Additionally, NASS partners with field offices at local governments or universities to conduct additional surveys as unique or specific needs may arise (USDA NASS, 2020, p. 17). NASS only collects data that provides statistics regarding US agriculture and submissions of data are restricted to data collected by NASS itself or an identified partner.

***Accessing NASS data:***

NASS is a public repository; anyone is able to access, download, and use NASS data. There are two main ways to access the data, directly through the website Quick Stats database or obtaining an API. Quick Stats allows users to complete queries and generate datasets. Each query return is limited to 50,000 data items. Database queries allow users to conduct searches to organize data by source, commodity, data item, sector, domain category, and location to create specified datasets which are downloadable and will open in programs like excel or numbers. Users have the option to name the dataset generated prior to downloading it. Using the Quick Stats database there is no need to create a login or account. The alternative to downloading data through the NASS Quick Stats database is to request an application programming interface (API) [here](https://quickstats.nass.usda.gov/api). The NASS website indicates that obtaining an API will allow users direct access to NASS data. A user must provide a valid email address to request an API but does not have to create a login or account. A valid email address is necessary to obtain the API and receive the link to download data files. Once the request is complete the API key is emailed to the user and the user will be able to download data files directly to their computer. By doing so a user is able to access more than 50,000 data items per query, differing from user experience when seeking access to data through the Quick Stats database feature.

There are no specific metadata standards used or displayed when downloading NASS datasets. As users are able to directly access and download data, generated datasets, and published reports, surveys, or censuses via the website or using an API key, there is no option to formally request information from NASS. A NASS dissemination information package (DIP) will only consist of any information downloaded by a user from the NASS website or using the API. When downloading data using the Quick Stats database no additional information is provided other than generated dataset. Similarly, when obtaining data utilizing the API key only files containing raw data are provided to users, no additional information or files accompany the data files.

**Considerations:**

***Suggested Citation:***

El-Husayni, N., Harwig, J., Neogy, R., & Osowski, F., (2020). *NASS Bee Colony and Honey*

*1974-2020.* Version 2. Kaggle. https://guides.lib.umich.edu/c.php?g=282964&p=3285995

***Preservation Considerations:***

Long-term preservation of the *NASS Bee Colony and Honey 1974-2020* dataset does not appear to be an issue at this time. The dataset is currently published on Kaggle, which is storing copies of the dataset. Kaggle is an online platform so there is no current concern for losing the files. Additionally, the dataset is within the public domain and can be viewed, downloaded, copied, or manipulated in common programs such as numbers, excel, or google docs, ensuring that there will not be future issues with data or software obsolescence. The CSV files can be viewed directly on Kaggle without downloading or requiring any software, further protecting against concerns of obsolescence.

***Suggested Copyright Statement:***

The content of this dataset is under Universal Public Domain and are free to the public and not protected by copyright. You may copy, modify, and distribute all material found in this dataset without asking for permissions.

***Human Subject Considerations:***

There were no human subjects used in regard to the data of the *NASS Bee Colony and Honey 1974-2020* dataset. Human subject considerations are not applicable to this project.

**References:**

https://www.kaggle.com/kenshoimpactteam/nassbeecolonyandhoney

United States Department of Agriculture. (Aug. 2020). Honey Bee Colonies. *National*

*Agricultural Statistics Service,* 1-18.

https://usda.library.cornell.edu/concern/publications/rn301137d?locale=en

United States Department of Agriculture. (March 2020). Honey. *National Agricultural Statistics*

*Service,* 1-6. https://usda.library.cornell.edu/concern/publications/hd76s004z?locale=en

United States Department of Agriculture National Agricultural Statistics Service. (Jan. 2021).

*Agency Overview.* https://www.nass.usda.gov/About\_NASS/index.php

United States Department of Agriculture National Agricultural Statistics Service. (May 2018).

*Mission and Core Values.* https://www.nass.usda.gov/About\_NASS/Mission\_Statement/index.php

United States Department of Agriculture National Agricultural Statistics Service. (July 2020).

*Talking about NASS Version 2.,* 1- 24.

https://www.nass.usda.gov/About\_NASS/pdf/Talking-About-NASS-Guide.pdf