Final Report and Recommendations of the Data Rescue Project at the National Agricultural Library

Prepared for the National Agricultural Library of the U.S. Department of Agriculture by

Cooper T. Clarke & Hilary Szu Yin Shiue

Data Rescue Fellows at the University of Maryland's College of Information Studies

August 2020

Disclaimer

This document was prepared by Data Rescue Fellows from the University of Maryland, College Park College of Information Studies. This document was reviewed by faculty at the University of Maryland and staff at the National Agricultural Library, it has not been peer reviewed. Any recommendations, standards, or guidance do not necessarily reflect those of the National Agricultural Library, U.S. Department of Agriculture, or the University of Maryland.

Table of Contents

I. Executive Summary

The National Agricultural Library (NAL) identified a need for a framework of guidance to support rapid appraisal and processing for scientific researchers' collections after being offered collections of scientific data and data-rich materials that required immediate appraisal before acquisition. To this end, the NAL partnered with the University of Maryland's College of Information Studies (iSchool) to support two Data Rescue Digital Curation Fellows to investigate processes for efficiently identifying, appraising, and processing scientific data out of legacy collections, to support data use and reuse. The original objective of the NAL Data Rescue project was to, "develop a tested process to rapidly assess collections of data and papers of retiring scientists or multiple scientists at laboratories that are closing" (McCarthy, 2019). The Digital Curation Fellowship program is a partnership between the National Agricultural Library and the University of Maryland College of Information Studies (iSchool) to connect students from across iSchool programs to research projects that help solve real digital curation challenges at the NAL. Mentored by iSchool Assistant Professor Katrina Fenlon, and supervised by staff in the Knowledge Services Division (KSD) and the Digitization and Access Branch at NAL, we conducted this research over the 2019-2020 academic year. This Report and accompanying Data Rescue Processing Guide document the work and scholarship of the Data Rescue Project.

The project was first tasked with researching data rescue, data curation, rapid appraisal, and knowledge retention. This research resulted in a significant literature review and informed every aspect of the project. Published literature on data rescue addresses rescuing existing records that are at risk of loss due to their age or format but rarely address collections that contain both digital and analog materials. The NAL expects data-rich materials to be in both analog and digital formats; consequently, our processing guide reflects traditional archival principles and contemporary data management standards. Additionally, we reviewed literature related to data curation, rapid appraisal, Open Archival Information System (OAIS) repositories, and agricultural data creators.

In order to develop a guide for future rapid data appraisal, team members were to decide on either applying one processing approach to several collections or to test three processing approaches on one collection. Because agricultural research is so varied, there will never be a 'one size fits all' solution to data rescue and data appraisal. Considering this and our literature review, we decided to develop a framework based on the Open Archival Information System (OAIS), implementing a uniform framework would benefit both processes of data rescue and curation. During our research we identified Cornell University Library's "Digital Processing Framework" which incorporates the standards of OAIS and we created a data processing guide heavily based on this framework (Faulder, 2018). During the drafting of the processing guide we analyzed two collections of data-rich materials already held by the NAL's Special Collections for insights into appraising data. After drafting the processing guide, we then tested it on a collection of data files created by Rufus Chaney, a retired USDA researcher.

The data being 'rescued' is intended for inclusion in the USDA's Agricultural Research Service (ARS) open access data repository, Ag Data Commons. The repository broadly supports the Federal Data Strategy and the Data Rescue Project expresses one of the Strategy's guiding principles, to "harness existing data" (Federal Data Strategy, n.d.). Federally created or funded

data to be 'rescued' are not considered federal records within the scope of the U.S. National Archives.

Based on our work and research, our recommendation focuses on implementing digital preservation practices to ensure data integrity and long-term access. Future data rescue projects will require policy or standards for the amount of data cleaning and curation that will be conducted internally before inclusion on Ag Data Commons. This will be heavily dependent on the data reuse value compared with the work required to make the data usable. Additionally, establishing a network of USDA researchers and field experts for data curators to consult could also facilitate rapid data appraisal processes, their knowledge is indispensable to understanding data.

II. Literature Review

Data Rescue

The current literature on data rescue is primarily related to climate data projects, especially in the wake of the 2016 U.S. Presidential Election. Specific projects include the Data Rescue: Archive & Weather (DRAW) and the Millennium Ecosystem Assessment (MA) collection of scientific data by the NASA Socioeconomic Data and Applications Center. This scholarship emphasizes either the data rescue of born-digital records in obsolete formats or digitizing legacy/analog materials, but not both. Downs and Chen describe a case study of the rescue of digital data from the Millennium Ecosystem Assessment in "Curation of scientific data at risk of loss: Data rescue and dissemination" (2017). The Millennium Ecosystem Assessment data was initially federally funded through the U.S. Geological Survey (USGS) until a 2012 budget cut ended service and support for the online data repository. The process of 'rescuing' the data began by assessing the reuse possibilities and the SEDAC appraisers determined the data was primarily historical as the science had been superseded. The digital data files were organized by theme, the authorship and publication rights were determined along with other metadata for description on the Socio-Economic Data and Applications Center (SEDAC) digital repository at NASA. Additionally, the data quality was analyzed for usability by the designated community by internal and external scientists. While the Millennium Ecosystem Assessment data rescue was born digital, the case study described in "Data rescue archive weather (DRAW): Preserving the complexity of historical climate data" by Park et al. (2018) focuses on analog to digital rescue. The DRAW project 'rescued' climate data handwritten in ledgers using a purpose-written metadata schema and transcription process. The online repository for the 'rescued' data complies with OAIS. The data rescue process for DRAW was further detailed in "From books to bytes: A new data rescue tool" by Slonosky et al. (2019) with many of the original authors from DRAW.

In "Where have all the scientific data gone? LIS perspective on the data-at-risk predicament" (2014) Thompson et al. surveyed forty-three information custodians for preservation plans and data practices. Most of the study focuses on rescuing analog and obsolete digital data with only a brief discussion of born digital data in the context of rescue. They point to the National Science Foundation's data management standards for grant funded research and the U.S. Office of Science and Technology Policy 2013 memo directing federal agencies to increase public access

to federal digital data, as signs of open data mandates that information custodians should recognize and create data rescue or preservation plans (Holdren, 2013).

Rescuing scientific data often serves several purposes. The preservation of data and records are one of the main concerns. For analog materials, curators face the challenges of media decay and obsolescence (Downs & Chen, 2017; Park et al., 2018). With the ever-changing technology and data management methods, access to past data and its discoverability becomes more difficult over time (Downs & Chen, 2017). The consideration of data reuse is also a main purpose of data rescue efforts. For past data to be accessible, discoverable, reusable, analyzable, data migration and/or transformation is necessary (Slonosky et al., 2019; Brunet & Jones, 2011). For governmental scientific data, other risks exist to influence the accessibility and long-term preservation of data, including policy changes due to government administration changes (Allen et al., 2017; Janz, 2018; McGovern, 2017). McGovern (2017) further addresses that federal-funded research data are more at risk of being lost than data generated by the federal government, because the former is dispersed and complies with different policies for retention and transparency.

Several documented data rescue projects have shown these concerns and demonstrated common issues encountered and provide recommendations. The Data Rescue: Archives and Weather (DRAW) project digitally captured and transcribed handwritten weather logbooks and registers at the McGill Observatory into machine readable formats. Because the climate data were from the nineteenth and twentieth century, the organization of data, format changes, different weather letters and symbols used over time all pose challenges to data rescue efforts due to the lack of uniformity (Park et al. 2018; Slonosky et al. 2019). The lack of quality and homogeneity of data also reduce their reuse value. In Brunet and Jones's research on existing climate records and their usefulness, they notice that "some of the available and accessible data do not reach the required standards of quality and homogeneity, making their usage doubtful for undertaking any climate analysis, applications, or services" (Brunet & Jones, 2011).

Insufficient documentation of data management practices creates obstacles for data rescue efforts. Downs and Chen (2017) discuss rescuing the Millennium Ecosystem Assessment (MA) collection of scientific data by the NASA Socioeconomic Data and Applications Center (SEDAC) and address that additional documentation, provenance information and methodological details need to be traced back for data to be useful, but the files were not well-documented which hinders the process.

Complex data management and documentation methods are especially prevalent in long-tail science and small science (Hsu et al., 2015; Akmon et al., 2011). Long-tail data are "data produced by individuals and small teams for specific projects, that 'tend to be small in volume, local in character, intended for use only by these teams, and are less likely to be structured in ways that allow data to be transferred easily between teams or individuals'" (Hsu et al., 2015). Nonetheless, "synthesis and reuse of the data, including for purposes other than the original intent, is one of the great benefits of rescuing long-tail data" (Hsu et al., 2015). The use of less-established standards, formats, and diverse data types makes homogenizing data more difficult. Akmon et al. (2011) offer further insights working with the Bennett Laboratory, a material science laboratory. They notice that laboratory members have various documentation

methods, data management practices, and that the variation limits data reuse by others (Akmon et al., 2011).

Interdisciplinary efforts in data rescue are recommended by many scholars to overcome the aforementioned challenges. DRAW is one example that utilizes different domain experts to transform physical analog data to a database format. In the DRAW data rescue process, "experts in archives, data and information studies, programming, geography, historical climatology, and meteorology" are all crucial for understanding, organizing, structuring data to making them available on the crowdsourcing platform and disseminating the computer-readable data (Slonosky et al., 2019). Akmon et al. (2011) noticed that archivists often do not see scientific data within their professional purview, while scientists lack knowledge in data curation. They further propose that the gap provides a great opportunity for archivists and scientists to work together to preserve scientific data. In examining multiple climate data rescue projects, McGovern (2017) also considers cross-domain collaborations to be strengthening data rescue expertise. For instance, a digital preservation team can provide obsolescence management advice, an archives team can offer provenance and context information, and a data science team can provide data curation knowledge (McGovern, 2017).

Organizational partnership and/or community involvement are recommended in some data rescue projects. Formed by five major American social science data archives, the Data Preservation Alliance for the Social Sciences (Data-PASS) aims to acquire and preserve social science data at risk of being lost to the research community, and ensure materials collected remain "accessible, complete, uncorrupted, and usable over time" (Altman et al., 2009). Organizational partnerships allow the Data-PASS to systematically preserve social science data existing across multiple repositories (Gutmann et al., 2009). To ensure consistency, appraisal guidelines and processing guidelines were created (Altman et al., 2009). The shared catalog of the Data-PASS project is identified by Altman et al. (2009) as an essential infrastructure because it facilitates automated metadata crosswalks while each participating institution uses their own schema internally.

For scalable data, such as surface climate data, terrestrial and marine meteorological data, international collaboration of data rescue can facilitate research by consolidation of terrestrial observation records, improve quality of data and metadata, and so on (Brunet & Jones, 2011). The two international data rescue (DARE) initiatives, the Mediterranean climate data rescue (MEDARE) project and the Atmospheric Circulation Reconstructions over the Earth (ACRE) initiative, discussed by Brunet and Jones (2011) are examples of international data rescue efforts. The goal of MEDARE initiative is to "develop a comprehensive high-quality, high-resolution time series of instrumental climate data for the GMR (Greater Mediterranean Region)," which will enable the GMR countries to better manage climate-related risks and adapt to climate-related impacts (Brunet & Jones, 2011). The ACRE initiative, led by a consortium, aims at "facilitating the recovery of historical instrumental surface terrestrial and marine meteorological observations recorded worldwide in order to support 4-dimensional weather reconstructions over the past 200–250 year" (Brunet & Jones, 2011). Without concerted international efforts, the large-scale research would not be possible.

Involving different communities in the data rescue process is also recommended by scholars. Through working with the Bennett Laboratory, Akmon et al. (2011) understand current scientists' data management practices and their data preservation needs. To solve the MA data issues, such as processing data that are not well-documented, missing sufficient metadata, and limited understanding of provenance information or documentation, the SEDAC team formed the SEDAC User Working Group (UWG), an advisory group composed of "scientists, representatives users, and other experts" (Downs & Chen, 2017). With the involvement of the UWG, SEDAC team was able to create a plan approved by the UWG to "archive and disseminate the MA collection with limited additional value-added efforts" (Downs & Chen, 2017).

Data rescue efforts can also create opportunities for public engagement. Multiple data rescue projects were initiated after the 2016 U.S. Presidential Election in order to rescue data provided by the government that might become inaccessible in the future due to policy changes by the new administration, potential budget cuts, etc. (Janz, 2018). Allen et al. (2017) introduced three web archiving initiatives that held multiple events to assist in their data rescue efforts, the Data Rescue events by the Penn Libraries joined the Penn Program in Environmental Humanities (PPEH), the Twin Cities Data Rescue events by the University of Minnesota (UM) and community-building events by the Mozilla Science Lab. Data rescue efforts such as data harvesting, web archiving of open federal data and websites, were achieved with help of the general public (Allen et al., 2017). These data rescue events not only maintain the accessibility of open data held by the U.S. federal government, they also create opportunities to raise public awareness that data are not only for use in scientific research, but can contribute to decision-making by local organizations and individuals (Janz, 2018).

Data Curation

Expanding the concept of *data rescue* to include considerations in data curation helps conceptualize a framework that can be used in different stages of the curation lifecycle. Current literature about data rescue often focuses on rescuing existing data at risk of being lost due to poor data management, lapses in funding, retirements, and changes in political administration or policy. The National Agricultural Library is primarily concerned with the potential for loss due to the retirement of individual scientists and the subsequent loss of context and institutional knowledge of their data.

The continuous efforts to maintain and preserve data for long-term use are critical in data curation work. In *Curating Research Data: Volume One*, data curation is defined as "the active and ongoing management of data through its lifecycle of interest and usefulness to scholarship, science, and education. Data curation enables data discovery and retrieval, maintains data quality, adds value, and provides for re-use over time through activities including authentication, archiving, management, preservation, and representation" (Johnston, 2017) Specific to digital data, Yakel (2007) states that "[d]igital curation is the active involvement of information professionals in the management, including the preservation, of digital data for future use."

The DCC Curation Lifecycle Model developed by the Digital Curation Centre (DCC) provides a conceptual view of data's lifecycle that illustrates ongoing data management and preservation

endeavors. Though the DCC defines data in this model to be "any information in binary digital form," the high-level view still contributes to conceptual thinking for data residing in other formats. From this perspective, many data rescue actions taken on existing collections can be seen as "reappraise," "preservation action," and "migration" that appear in Figure 1, while processing retiring scientists' materials would involve more stages from "receive," "appraise & select," " ingest" to "transform."

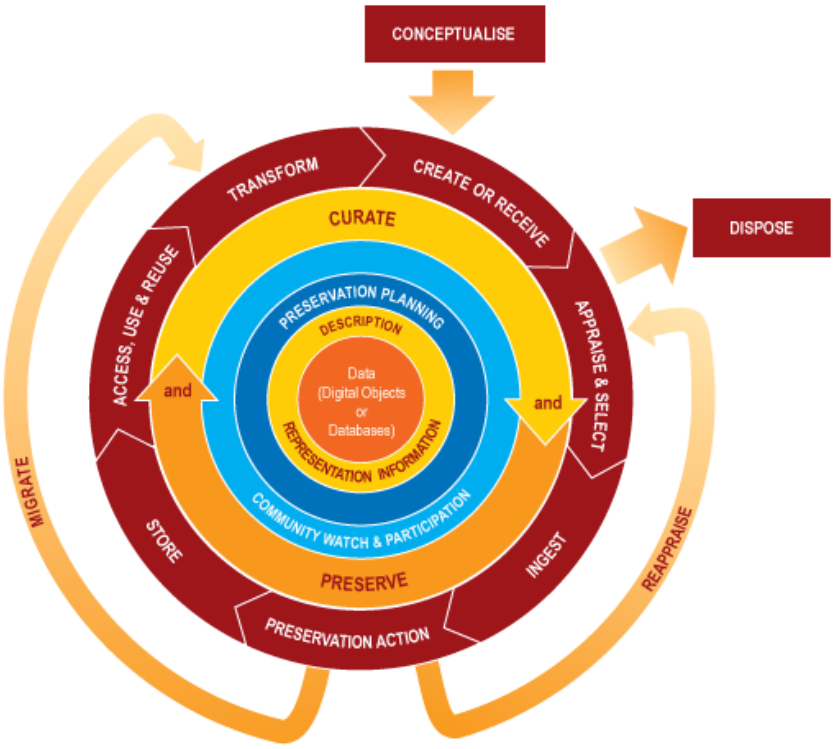


Figure 1: The DCC Curation Lifecycle Model

Notably, the "conceptualize" step that "conceives and plans the creation of data, including capture method and storage options" encourages data curators and repositories to understand data producers' data types, generation methods and the implications posed with ingest (Higgins, 2008). The gap between the archivist community and the scientist community in digital preservation is discussed by Akmon et al. (2011). They noticed archivists do not view scientific data within their professional purview, while scientists lack knowledge in curating data for long-term use, which hinders the digital preservation process. Specific to the agricultural field, "Supporting the changing research practices of agriculture scholars" (2017) reflects similar concerns between information professionals and practitioners. It further addresses the research data lifecycle in the agriculture field, including scholars' data discovery methods, data management, and dissemination. Both articles present the importance of interdisciplinary efforts to curate research data for long-term use.

Open Archival Information System (OAIS) Repositories

The Open Archival Information System (OAIS) model, though primarily applied to born-digital collections, can be "applied to the long-term preservation of items in any form" (Lavoie, 2014). For instance, using an OAIS-compliant model, the Data Rescue: Archive & Weather (DRAW) project aims at "rescuing the climate data buried in Observatory registers; develop a protocol by setting up a repository to preserve historical scientific data; and make the data usable for research and accessible to the next generation" (Park et al., 2018). Their application of the OAIS model aided in the process of digitally capturing useful data from century-old physical weather registries and allowed the project team to operate a crowdsourcing website for digitized records to be transcribed.

Other than the broader definition of data at the OAIS model, the functional view of the model also offers guidance on what metadata to gather and what content to preserve and/or disseminate in the form of the three information packages. Notably, the consideration of the designated community in the OAIS model is also essential and helpful when making processing decisions. Detailed discussion of the OAIS model is provided in our processing guide.

Appraisal

Assessing data quality, utility, and integrity are central to appraising data sets and scientific research material for acquisition decisions. The archival concept of More Product, Less Process (MPLP), described in the article of the same title by Greene and Meissner (2005), argues that archivists are spending too much time processing collections and should prioritize access over removing every fastener, refoldering, and describing beyond the folder level. Greene and Meissner advocate this approach to processing to reduce large backlogs that already overburdened institutions may never make available to the public. Several authors suggest that MPLP can be adapted to the rapid appraisal techniques of data that prioritize user access over lengthy description and arrangement. In the article, "More data, less process? The applicability of MPLP to research data," Lafferty-Hess and Christian (2017) propose that MPLP can be interpreted for data processing and appraisal by focusing on "understanding of the whole" for basic description and metadata that is not duplicative of information already in the data and leave the research to the patrons. They assert that file normalization is best for accessibility and preservation as well as preserving the original proprietary file. MPLP is also invoked by Belovari in "Expedited digital appraisal for regular archivists: An MPLP-type approach" (2017). Belovari argues that appraisal and arrangement should be done simultaneously and as either a "broad appraisal" which is conducted manually to remove duplicate or junk files quickly or "in-depth qualitative appraisal" which determines the archival value of individual files and can be done in consultation with processing software.

"The selection, appraisal, and retention of digital social science data" (2004) by Gutmann et al. provides a case study of the selection, appraisal, and retention processes used by two data archives. Similar to Belovari, Gutmann et al. describe varying levels of processing intensity based on the quality of original documentation, reuse value, sensitivity issues, uniqueness, and relation to other data already maintained. Most relevant to the NAL, Faundeen and Oleson describe the appraisal process utilized by the U.S. Geological Survey (USGS) Center for Earth

Resources Observation and Science (EROS) in their article, "Scientific data appraisals: The value driver for preservation efforts" (2007). After identifying a data set, their appraisal process gathers ad hoc appraisal teams that are knowledgeable in the data being appraised. The appraisal team reviews the data for authenticity, reliability, integrity, reusability, and sensitivity. The review process results in documentation that should be preserved alongside the data and can be used for description/metadata. This process allowed the USGS to appraise both new data sets and data sets already in their collection, some of which were found to be outside the scope of the collection and were deaccessioned.

Data Producers and Designated Communities

The formats and types of data created by any research discipline are incredibly varied, and agricultural research is no different. The NAL provides data to many stakeholders in fulfilment of its mission and that of the USDA. The Library primarily offers access through AGRICOLA and Ag Data Commons. The nuances of data creation and use affect their ultimate preservation and dissemination. Cooper et al. (2017) point out in their article "Supporting the changing research practices of agriculture scholars" that agricultural scholars have no trouble with digital discovery and access but struggle with information management. Based on 230 interviews with agriculture scholars, the authors revealed that most are not maintaining or publishing their data outside peer-reviewed journals. A lack of best practices and data management infrastructure leads to lost and mismanaged data.

In "The application of archival concepts to a data-intensive environment: working with scientists to understand data management and preservation needs," Akmon et al. (2011) argue that archivists are largely ignoring scientific data sets for preservation and the scientists creating this data are unable to systematically review, manage, and preserve their work. They created a case study of a material science laboratory at a large U.S. university, finding that the lab did not have any consistent data management infrastructure making data reuse difficult if not impossible. Akmon et al. concludes that data creators and archivists should meet in the middle to develop practical data management processes. McGovern reacts to the increasing politicization of federally funded and federally created data in "Data rescue: Observations from an archivist" (2017). McGovern describes the differences between persistent access and persistent preservation, the latter being the actions to ensure digital data remains meaningful across technological change while the former relates to immediate discovery of data and digital materials. While the data and formats vary by discipline, the catalysts for data preservation and open access remain the same.

III. Proposed Data Rescue Approach

Based on our extensive literature review we determined that the data rescue appraisal and processing guide should be primarily based on the Open Archival Information System (OAIS) because of its thorough considerations for data creators, users, access, metadata, and preservation. The OAIS preserves files as received, as preserved internally, and as disseminated in three seperate 'packages.' This allows for any data curation or processing actions to be tracked from package to package and can be reversed if necessary. For example, if a researcher deposits a collection of Excel, PDF, and Word files these are all retained in the Submission Information

Package (SIP). The Archival Information Package (AIP) may only retain the Excel file, a CSV of the data in the Excel file, and Word files if the PDFs are documents already widely available. While the Dissemination Information Package (DIP) only contains the prepared CSV for public reuse. The OAIS acknowledges the necessity of maintaining files internally that may not be appropriate for public dissemination.

The OAIS establishes six requirements the system is grounded to; this foundation ensures that the data/materials can be independently understood by the designated community while properly preserved and documented. The first requirement is to maintain acquisition policy that limits what is appropriate for long term preservation by the institution and to communicate this to data producers. Secondly, the archive must receive or arrange intellectual and physical control of the materials. The third requirement is to determine the designated community that the materials are expected to be reused by, this influences the amount of curation required for the fourth requirement, to make the materials independently understandable by the designated community. The fifth and sixth requirements focus on the preservation and dissemination of the materials. These six requirements form the high level steps of our processing guide (Lavoie, 2014).

The "Digital processing framework," written by staff at Cornell University Library, was the only detailed framework for processing digital materials we were able to identify. The Cornell Framework informed the structure of our Data Rescue Processing Guide which we intend to be reused by the NAL for future data rescue projects. Both the Cornell framework and our processing guide incorporate varying levels of processing 'intensity' that is informed by the data value, reuse expectations, and the designated community. The Processing Guide can be used for both analog and born-digital data-rich materials with a final product of a digital data set.

IV. Analog Data Rescue Evaluations

To better understand existing appraisal practices and analog/legacy data materials, we examined two collections of data-rich materials already held by special collections at the NAL. The legacy data we analyzed were the Coville Blueberry note collection and the W.O. Atwater nutrition data sheets. Hilary Szu Yin Shiue worked with the Coville Blueberry note collection and Cooper Clarke worked with the W.O. Atwater collection. These separate processes allowed us to refine our processing guide against already processed and appraised legacy collections. Using an initial draft of our OAIS processing guide, the steps below point to some essential considerations we take from the perspectives of data rescue.

A. *Frederick V. Coville Blueberry Records (MS 413)*

Acquired in 2007, the Frederick Vernon Coville's Blueberry Notes collection is considered an important collection by the National Agricultural Library (NAL) because of the reuse value of Coville's research data in the notes as well as his significant contributions to blueberry domestication. "Experiments in Blueberry Culture" published in 1910 documented Coville's discoveries for blueberry domestication, including the use of acidic soil, the plant's dormancy, the need for cross-pollination to produce better yields, and so on (Coville, 1910). In the *Yearbook of Agriculture*, Coville's article, "Improving the Wild Blueberry," described fifteen remarkable

cultivars (Coville, 1937). Notably, some cultivars published by Coville are still being planted today, such as the Jersey blueberry and the Rubel blueberry.

Coville's Blueberry Notes collection is about six linear feet in total. It is stored in twenty-four boxes arranged in the order they arrived at the NAL. Most of them are loose leaf pages, but there are also eight spiral-bound notebooks in the collection. The notes were dated from 1907 to 1938, including one notebook from 1938 kept by George Darrow, Coville's successor in blueberry research. They were mostly handwritten with some typed notes. Currently, the collection and existing documentation are held by Special Collections at the NAL. There are administrative files and a container list. As the collection has not been processed, there are no finding aids related to Coville's Blueberry Notes collection. The first two boxes of the notes, however, are available on the Internet Archive (Coville, 1907-1908).

Data Rescue in Coville's Blueberry Notes Collection

There are numerous types of data that exist in Coville's notes, including blueberry pedigree information, fieldnotes, descriptions of different characteristics of blueberry cultures, and more. Portions of these data and the resulting findings have been published in USDA bulletins and contributed greatly to the domestication of blueberries. Nonetheless, the unpublished longitudinal data in Coville's notes still have significant reuse value to blueberry researchers analyzing contemporary blueberry cultivars that were first grown by Coville, agricultural scholars studying the history of the USDA and blueberries, and the general public interested in the work of the USDA.

Coville's Blueberry Notes collection is in a fragile condition due to its age and format, which makes reuse and locating distinct data sets more difficult. Constantly handling the loose-leaf pages may damage them. The original order of the loose-leaf notes would also be difficult to maintain. Therefore, rescuing data from Coville's Blueberry Notes collection requires proper format migration, transferring data from the analog physical medium to more accessible formats, such as comma-separated values format, PDF format, and so forth. With limited time, our data rescue project focuses on finding out what types of data are of importance to the designated communities of this collection, and also to understand who the designated communities will be. This information will further inform how we process and make available the data from Coville's Blueberry Notes collection.

Surveying Coville's Blueberry Notes Collection

Surveying the collection helps us understand what types of data exist, how they were documented in Coville's notes as well as how different pieces of data may connect together. Coville's notes include daily entries of observations of blueberries cultures, which could be presented in tabular formats, descriptive texts, etc. Coville also kept longitudinal records of blueberry experiments, and provided detailed information about different cultures and cultivars. Nonetheless, the notes present some challenges for processing. Some examples are provided below.

- Inconsistent data documentation methods

Different types of data can be found in Coville's notes, tabular data, description texts, fieldnotes, pedigree information. In Figure 2, dated September 6, 1912, the tabular data documented past cross pollination on May 17, 1912. Coville also wrote down descriptive information about different cultures and his experiment decisions (Culture 620 and 621).

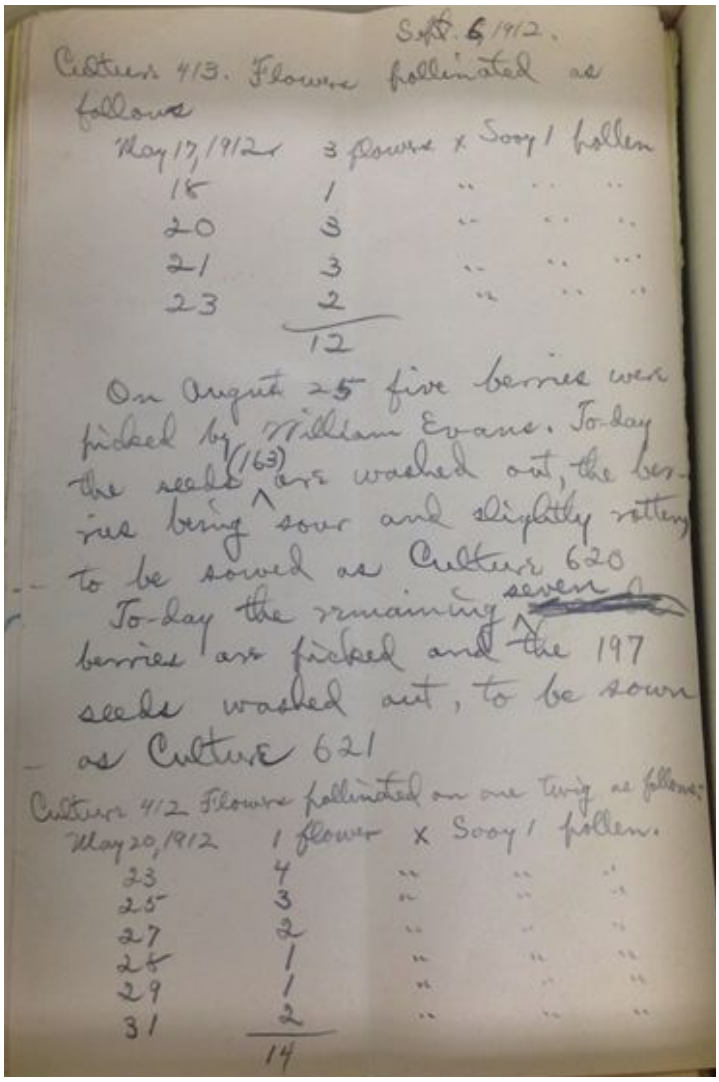

Sept. 6 1912.
Culture 413. Flowers pollinated as follows
May 17, 1912 3 flowers x Sooy 1 pollen
18 1 " " "
20 3 " " "
21 3 " " "
23 2 " " "
12
On August 25 five berries were picked by William Evans. To-day the seeds (163) are washed out, the berries being sour and slightly rotten to be sowed as Culture 620
To-day the remaining seven berries are picked and the 197 seeds washed out, to be sown as Culture 621
Culture 412 Flowers pollinated on one twig as follows:
May 20, 1912 1 flower x Sooy 1 pollen.
23 4 " " "
25 3 " " "
27 2 " " "
28 1 " " "
29 1 " " "
31 2 " " "
14

Figure 2: Photographed loose leaf note from "Frederick Vernon Coville Blueberry Notes | USDA Bureau of Plant Industry Horticultural and Pomological Investigations Records | 1912"

Figure 3 is a good example of what types of information Coville recorded. Figure 3 shows blueberries that were selected by Coville, which indicated they had good plant characteristics. The data that Coville documented include cultivars, fruit size, plant origin and their characteristics.

Nonetheless, the date of the loose leaf note is missing, though it is stored in the box "1913." The data was kept in tabular format with column headings with some empty fields.

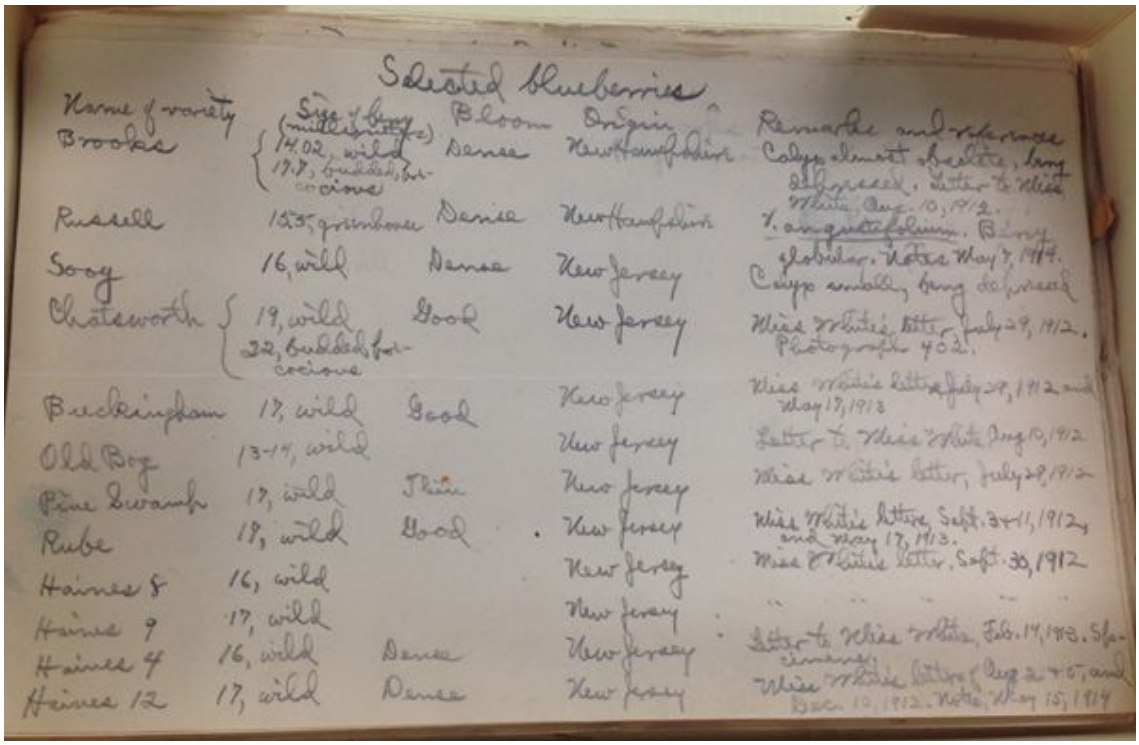
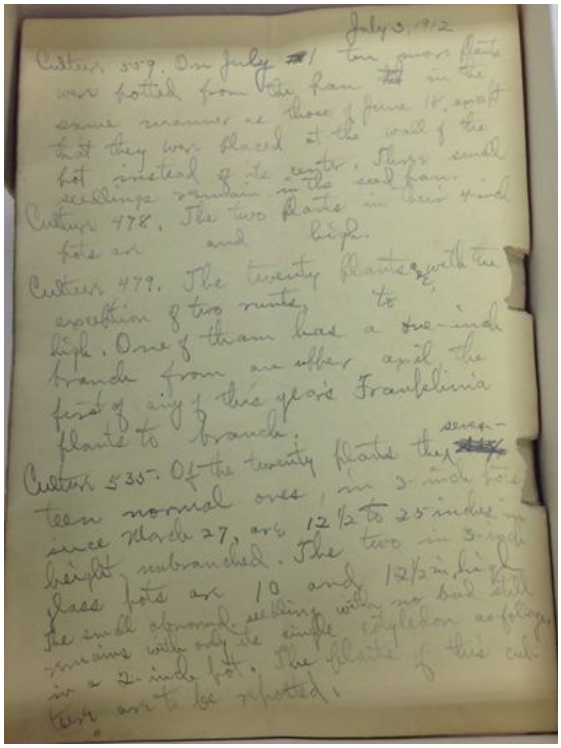


Figure 2: Photographed loose leaf note from "Frederick Vernon Coville Blueberry Notes | USDA Bureau of Plant Industry Horticultural and Pomological Investigations Records | 1913"

Figure 3 is an example of Coville's fieldnote. He documented in descriptive texts how different cultures were growing including the details, such as size of the planting pot, the height of the plant.



July 3, 1912.

Culture 529. On July 1 ten more plants were potted from the bank in the same manner as those of June 18, except that they were placed at the wall of the hot instead of its center. These small seedlings remain in the seed pans.

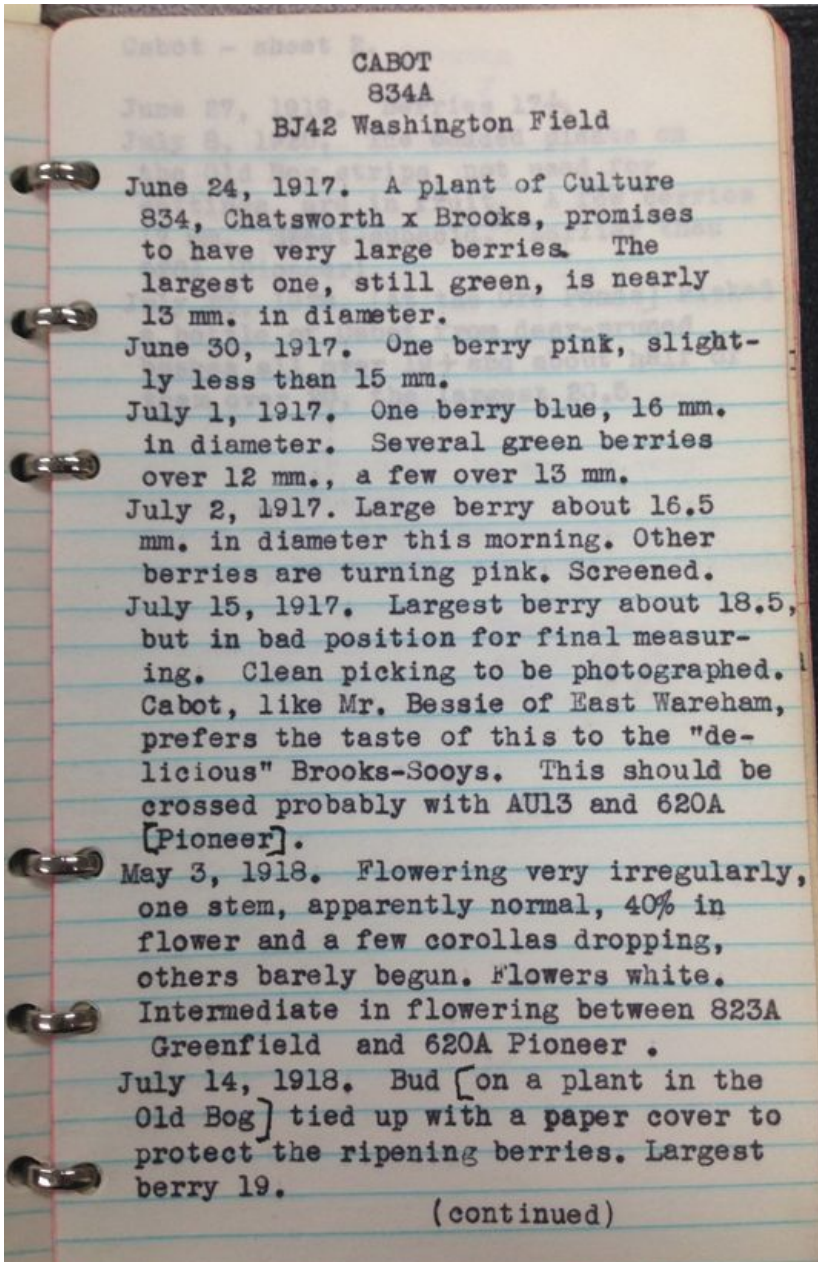
Culture 478. The two plants in [illegible]-inch pots are [illegible] and [illegible] high.

Culture 479. The twenty plants, with the exception of two runts, [illegible] to [illegible] high. One of them has a five-inch branch from an [illegible] axil the first of any of this year's Franklinia plants to branch.

Culture 530. Of the twenty plants the seven-teen normal ones, in 3-inch pots since March 27, are 12½ to 25 inches in height, unbranched. The two in 3-eighth glass pots are 10 and 12½ inches high. The small almond seedling with no bud still remains with only the simple cotyledon and foliage in a 2-inch pot. The plants of this culture are to be repotted.

Figure 4: Photographed loose leaf note from "Frederick Vernon Coville Blueberry Notes | USDA Bureau of Plant Industry Horticultural and Pomological Investigations Records | 1912"

Some of Coville's notes were typed with a typewriter. Figure 5 is an example. The cultivar CABOT was one of the fifteen releases published in the 1937 Yearbook of Agriculture. The listed dates and observations could potentially be traced back to Coville's daily entries.



CABOT
834A
BJ42 Washington Field

June 24, 1917. A plant of Culture 834, Chatsworth x Brooks, promises to have very large berries. The largest one, still green, is nearly 13 mm. in diameter.
June 30, 1917. One berry pink, slightly less than 15 mm.
July 1, 1917. One berry blue, 16 mm. in diameter. Several green berries over 12 mm., a few over 13 mm.
July 2, 1917. Large berry about 16.5 mm. in diameter this morning. Other berries are turning pink. Screened.
July 15, 1917. Largest berry about 18.5, but in bad position for final measuring. Clean picking to be photographed. Cabot, like Mr. Bessie of East Wareham, prefers the taste of this to the "delicious" Brooks-Sooys. This should be crossed probably with AU13 and 620A [Pioneer].
May 3, 1918. Flowering very irregularly, one stem, apparently normal, 40% in flower and a few corollas dropping, others barely begun. Flowers white. Intermediate in flowering between 823A Greenfield and 620A Pioneer.
July 14, 1918. Bud [on a plant in the Old Bog] tied up with a paper cover to protect the ripening berries. Largest berry 19.
(continued)

Figure 5: Photographed notebook page from "Frederick Vernon Coville Blueberry Notes | USDA Bureau of Plant Industry Horticultural and Pomological Investigations Records | Number 1"

- Missing column names for tabular data

  In Figure 6 the tabular data seem to include different cultivar crosses, culture numbers. However, the column names were not documented, so it is difficult to confirm the semantic meanings of the columns. The first and second column would be hard to determine. The semantic meaning of the last column of date would be difficult to comprehend, too.

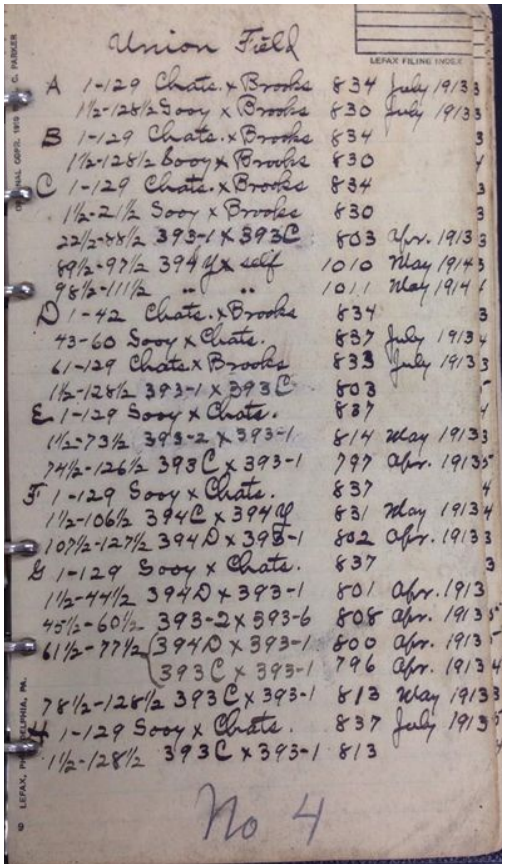


Figure 6: Photographed loose leaf page in notebook from ”Frederick Vernon Coville Blueberry Notes | USDA Bureau of Plant Industry Horticultural and Pomological Investigations Records | Number 4: Union Field, Whitesbog, New Jersey”

- Missing indications of plant species

Figure 7 is an example that some culture numbers documented in Coville’s notes may not be blueberries. On this page, it was written “2000 azalea arborescens x a. lutea R xxxx…..” “2000” seems to be a culture number, but it was azalea arborescens which seems to be a flowering plant.

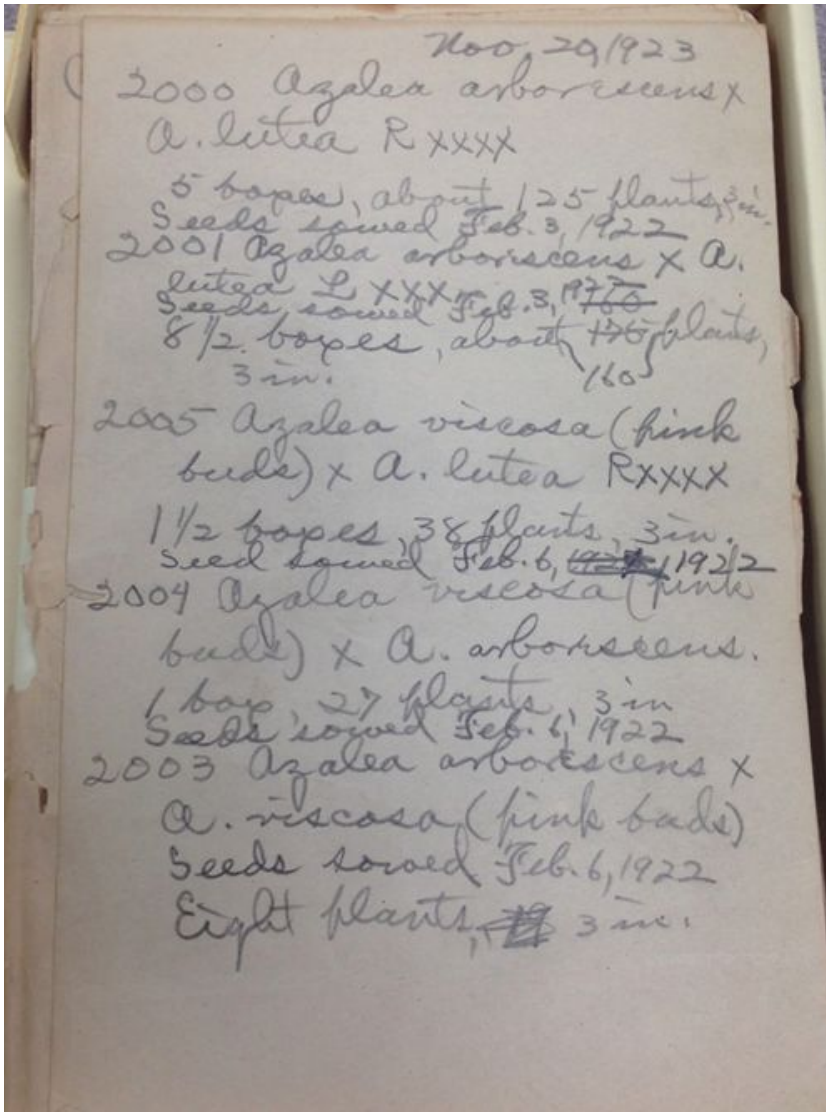


Figure 7: Photographed loose leaf page from "Frederick Vernon Coville Blueberry Notes | USDA Bureau of Plant Industry Horticultural and Pomological Investigations Records | 1922-1925"

Decision to Consult an Expert

As the century-old data residing in Coville's Blueberry Notes collection are difficult to understand, our supervisors from the NAL suggested we interview an expert to further understand the collection. More importantly, the expert consultation would focus on understanding if the data are unique in this field, if they can be reused, and who would be the designated communities.

Susan Fugate, the head of the Special Collection, assisted us to contact Robert Griesbach for an interview. Griesbach helped acquire this collection to the NAL, and was previously a research geneticist at the USDA. With Griesbach's familiarity with the collection and his background in research, interviewing him certainly aided us in how to process Coville's Blueberry Notes

collection. Our in-person interview scheduled on March 23, 2020, however, was moved to a phone conference presentation instead, due to the hit of the COVID-19 pandemic.

Though brief, Griesbach's presentation and short interview session during the conference call has provided us with useful processing considerations. In the brief interview, he mentioned three types of designated communities for Coville's notes: (1) the general public, (2) horticultural scholars, and (3) genetic scientists.

The reuse value of the Coville Blueberry Notes collection varies between the designated communities; this bears implications for data formats for dissemination. For example, both the general public and genetic scientists are interested in pedigree information, but the general public may be interested in knowing the parent cultivar names of well-known cultivars, and in what years they were released, while the genetic scientists would need to understand detailed pedigree information and characteristics of plants, even if some cultivars were never released. Table 1 presents the three types of designated communities for Coville's Blueberry Notes collection, the types of data interesting to them and the reasons.

Designated Communities	**Data of Interest**	**Reasons**
The General Public	Blueberry pedigrees of known cultivars.	People are more and more interested in knowing where their food comes from.
Horticultural Scholars	Information about horticulture in Coville's notes, such as fertilizers, cold treatments of plants, etc.	Some of the first fertilizers can be found in Coville's notes. For instance, Coville discovered that using acidic soil was key to blueberry cultivation.
Genetic Scientists	Pedigree information and documentation of characteristics of cultivars.	Using this information, geneticists would be able to research how certain characteristics of plants were inherited, which could be used to improve future cultivars.

Table 1: Designated communities for Coville's Blueberry Notes collection, data types that interest them and the reasons.

After Griesbach's presentation, we produced three documents for future reference, namely, presentation transcript, presentation key points, and presentation memo. Presentation key points, included as Appendix 1, offer the presentation outline in a chronological order. Lastly, written from the context of data rescue, the presentation memo in Appendix 2 documents our reflections on the presentation and recommendations for future data curators.

Current Process and Next Steps

The identified designated communities will all need to access these materials digitally. At the time of writing, the National Agricultural Library is working on transcribing the two digitized boxes of Coville's Blueberry Notes using the platform FromThePage as a test run. The lead researcher for this collection created a transcription style guide (available in Appendix 3) based on the transcription guide from the Smithsonian Transcription Center (Transcribing on the Transcription Center, 2020). Transcribing the notes can inform how to digitally capture and transform handwritten unnormalized data for dissemination. Once the preliminary test is complete, digitizing and transcribing other volumes of the collection may be the next steps. However, prioritization of which type of data, such as tabular, textual, pedigree data, etc., to transcribe is not yet decided, and may be decided after the test run is finished.

B. *Wilbur O. Atwater Papers (MS 261)*

The Wilbur Olin Atwater Papers (MS 261) held by the USDA's Special Collections contain over 900 handwritten data sheets documenting Atwater's studies of food nutrition and caloric composition. The studies were conducted for the USDA by the Office of Experiment Stations from the mid-1890s to 1906. Atwater's nutrition investigations fundamentally changed the department's approach to nutrition and food composition. The data sheets match several Atwater USDA publications and often contain further data, primarily, "The chemical composition of American food materials" (Atwater & Bryant, 1906). The data sheets are organized by food type and document the percent of protein, water, carbohydrates, "refuse," and "ash" per pound as calculated by Atwater and other researchers using bomb calorimeters. Some data sheets contain information on the sourcing of the food type tested as well as margin notes (see Data Sheet Example below).

Validity and Appraisal

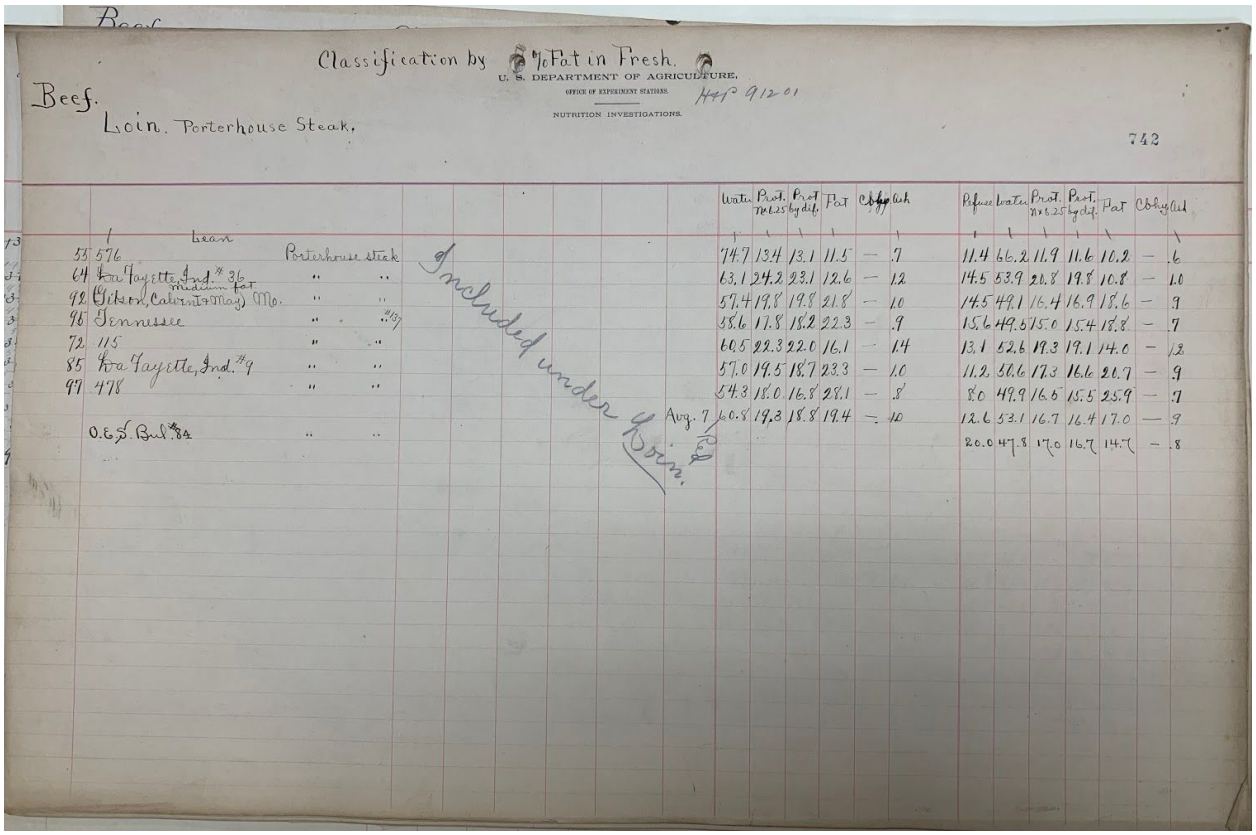
Atwater's research is still frequently cited by researchers in the field; his publications have been cited over 100 times in the past decade according to Web of Science. Atwater's data is even the subject of journal articles that critique current use of formulas first developed during his research. The 2016 article, "Calculating the metabolizable energy of macronutrients: A critical review of Atwater's results," by Sánchez-Peña et al. argues the Atwater system is flawed and in need of being updated. The Atwater system is a method of calculating the available energy of food, developed as part of his respiration calorimeter studies (Sánchez-Peña et al., 2016). The 2019 article, "Heats of combustion representative of the carbohydrate mass contained in fruits, vegetables, or cereals," analyzes Atwater's still utilized formula for calculating the carbohydrates of vegetable sources and concludes it overestimates the heat of combustion (Martínez-Navarro, 2019). Access to this raw data would offer insights into Atwater's process and methodology used to create nutrition formulas still in use today. Contemporary researchers could make use of this data to compare with contemporary quantitative studies.

Collection Survey:

Box	Data sheet # START	Data sheet # END	Contents	Page count
3	1	19	Prepared foods for infants and invalids	30
3	20	59	Fruits, canned, cooked, dried, juices, jelly, preserves	39
3	60	92	Fruits, canned, cooked, dried, juices, jelly, preserves	31
3	93	106	Nuts	14
3	107	116	Beverages, yeast	10
3	117	160	Grains, meals, flours, breakfast foods	45
3	161	200	Grains, meals, flours, breakfast foods	40
3	201	242	Breads	42
3	243	273[1]	Pastries, honey, sugar, molasses, and starches	30
4	275	310	Vegetables, fresh, cooked, canned, and dried	35
4	311	350	Vegetables, fresh, cooked, canned, and dried	40
4	351	382	Vegetables, fresh, cooked, canned, and dried	30
4	383	389	Condiments, pickles	7
4	390	420	Fish, fresh, cooked, pressed, canned, etc.	30
4	421	455	Fish, fresh, cooked, pressed, canned, etc.	35
4	456	470	Frog legs & shellfish	15
4	471	496	Egg & egg substitute	26
5	497	545	Dairy products	50
5	546	574	Condensed food & soups	30
5	575	587	Miscellaneous: gelatin, sturgeon, mincemeat, lard, cottolene, oleo margarine, sandwiches, hash, dendeng, bones	15
5	588	630	Lamb & mutton	45
5	631	659	Pork	30
5	660	690	Pork	30
5	691	711	Sausage	23
5	712	724	Poultry and game	15
6	725	762	Beef	51
6	763	813	Beef	68
6	814	858	Beef	46
6	859	883	Veal	31

[1] Data sheet 274 does not exist or is missing (missing from original finding aid as well)

Figure 8: data sheet example (no. 742)



Notice the general food product type in the top left corner, beef, followed by the cut of meat, loin porterhouse steak. The data sheet Bates number in the top right corner. This data sheet lists the sourcing of the food product in the left rows and the results of analysis on the right; listing the calculated "refuse," water, protein n*6.25, protein by dif., fat, carbohydrates, "ash." This particular page does not list the calculated calories (kcal), however, most do. In this example, each data column is labeled, most data sheet columns are not as well labeled but follow a similar layout.

OAIS Framework

Using our initial draft framework, the primary researcher for the Atwater collection briefly describes how the appraisal of the Atwater collection either fulfills or does not apply to each step in the three OAIS packages. Some of the steps were not applicable because of the analog nature of the collection and some steps can only be completed after the collection has been digitized.

1. Create Submission Information Package (SIP)
    a. Identify collection and documentation - Atwater collection contains textual data sets
    b. Survey the collection - reviewed every data sheet box and entire contents, over 900 pages with handwritten data in rows and columns
    c. Identify restricted material - not applicable, public domain

    d. Manage PII - not applicable
2. Create Archival Information Package (AIP)
    a. Survey the collection - complete, over 900 data sheets contained in 3 oversize archival boxes
    b. Create processing plan - plan for digitization and/or transcription
    c. Determine level of description - level of metadata description necessary and file naming conventions
    d. Address presence of duplicate content - *post digitization* likely not applicable
    e. Record technical metadata - *post digitization*
    f. Gather metadata for description - *post digitization*
    g. Organize files/materials - complete for physical files already accessioned and organized; further organization require post digitization
3. Create Dissemination Information Package (DIP)
    a. Write or edit description and final metadata - *post digitization*
    b. Publish finding aid - *post digitization*
    c. Add description about electronic material to finding aid - *post digitization*
    d. Publish/update catalog record - *post digitization*

Designated Community

Because the data requires a basic understanding of food composition and nutrition research, the designated community will primarily be composed of scientific researchers looking for legacy data. The data could also be useful for historians analyzing scientific research methodology at the turn of the 20th century or researching Atwater specifically. These researchers expect to access the data digitally, scientists likely would prefer a complete transcription. However, it may be simpler for interested parties to transcribe the portions they are interested in independently. Simply having access to scans of the data sheets should be sufficient for the majority of researchers.

Data Quality Analysis

To assess the quality of the data a variety of questions should be considered; some are broad that apply to all data rescue cases and others are specific to the data in question and should be further developed with knowledge of the field or subject matter.

Appraisal questions:

1. *Is the data used in a published article/study?* Yes, however, the data varies slightly and appears in several similar articles/bulletins published by the USDA.
2. *Potential for reuse in the scientific community?* Unknown, possible for researchers to analyze historical methodology and conclusions.
3. *Can this data be reproduced?*
4. *Is the methodology documented?*
5. *What is/was the intended use?*
6. *Are there any use limitations (e.g. potential issues a user removed from the original researcher may not understand)?*
7. *What format would be preferred by a researcher? Not preferred?*

8. *Would changing the data's appearance (e.g. formatting, layout, structure) change its meaning or value?*
9. *How much work is required to transform the data (digitize and/or transcribe)?* Extensive, all pages are handwritten, and some have strikethroughs complicating the transcription process. Option to transcribe the typed data from the published articles, categories and headings are the same between the data sheets and published data.
10. *How could Atwater's raw data be used by a contemporary researcher?*
11. *Has nutrition science methodology changed? Would that change be documented in this collection?*
12. *How does a nutrition researcher measure data quality?*
13. *Is there a specific food type studied by Atwater that would be of particular interest to nutrition researchers if the entire collection could not be transcribed (e.g. beef, potatoes)?*

Field experts for possible data reuse consultation:

- Beltsville Human Nutrition Research Center
- Methods and Application of Food Composition Laboratory (MAFCL) - Beltsville
- Food Surveys Research Group - Beltsville
- Food Components and Health Laboratory - Beltsville
- Food Quality Laboratory - Beltsville

Next Steps

The simplest option to "rescue" the data would be transcribing the data contained in Atwater's USDA publication, "The chemical composition of American food materials" from 1906 in the Office of Experiment Stations Bulletin No. 28. The data sheets match the data in this publication and it would be simple to transcribe the text from the already digitized publication. The transcription could be easily formatted in a format suitable for Ag Data Commons. While this data is already widely available on archive.org and Google Books, it would require data cleaning because the automatic OCR contains errors and cannot be manipulated.

A more thorough option would be to scan the front page of each sheet and arrange the sheets according to the Bates numbering and arrangement already in place. Because the data sheets are handwritten, manual transcription would be required to make the data machine-readable, this could be achieved with the NAL's newly implemented FromThePage transcription system. Scanning the data sheets is the minimum approach necessary for their data "rescue" because they cannot be properly transcribed before digitization. Scans could be suitable for NAL Digital Collections and/or Ag Data Commons, with varying levels of metadata and descriptive file names. Ag Data Commons does accept image files (e.g. .JPG, .TIF) and hosts several scanned historic data collections; the Pomological Watercolor Collection is linked, and there is a North Dakota Aerial Image Dataset.

V. Digital Data Rescue Evaluation

A. *Rufus Chaney Collection*

Rufus Chaney, a now retired USDA research agronomist, studied the effects of elements (primarily heavy metals) on soil and crops for human consumption. According to Google Scholar, Chaney is listed as an author on over 500 publications and has over 35,000 citations (as of June 2020) (Google Scholar, n.d.). Chaney identified the collection in question as significant and as having a potential for scientific reuse, he gave the files to the NAL for inclusion in Ag Data Commons. However, the collection was not in a condition for immediate upload, it required curation, cleaning, and description.

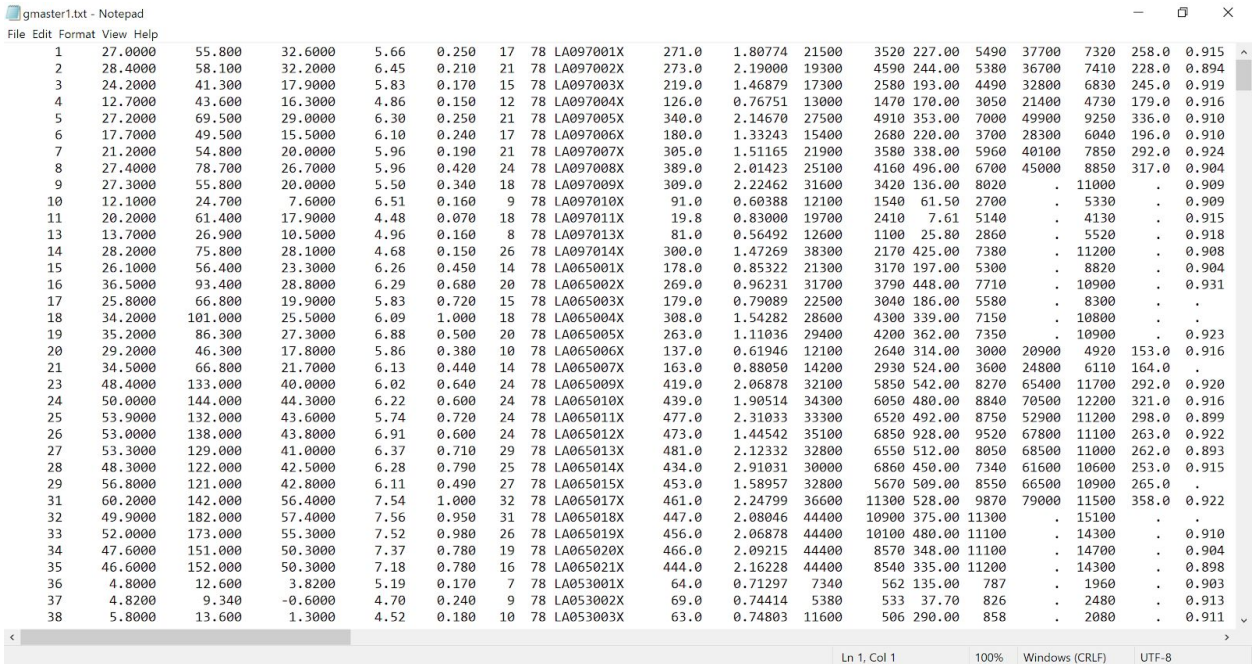
We received the Chaney collection on a removable drive which copied the files from the NAL's network drive. Chaney identified and transferred the files to the NAL for their eventual inclusion on Ag Data Commons, the USDA's digital data repository. Upon receiving the files, our first step was to create a compressed archive file to preserve everything exactly as they were received. We used 7-Zip, an open-source file archiving and compression software. However, creating a TAR or ZIP file would also suffice. This archive file is eventually used for the Submission Information Package (SIP) and acts as a backup if files or data are lost in processing. This copy should ideally be saved to a different system (e.g. network, removable storage) in case of data loss.

A working folder containing the existing folder structure and files was then created for assessment and processing. To create a complete inventory of the collection, we used a file directory command (dir /ogn /s >file_list.txt) with Command Prompt. This creates a TXT file listing every folder, file, and associated metadata. The command will also reveal any hidden files. From this TXT file a spreadsheet inventory was created listing each file with variables for associated metadata (folder name, original date, original size in bytes, original file name, original file format). In this inventory, we also kept notes on individual files and information describing how the file was converted to a sustainable format.

As received, the Chaney collection consisted of 262 files spread across 14 folders. Generally, the files were organized by crop type and date, for example, there is a 'Wheat-2012' folder and a 'WHEAT-2008' (file and folder naming stems from the original creator and was preserved as received). The files were in a variety of formats, both proprietary and open: 77 WPD, 52 SAS, 30 LOG, 28 PDF, 21 LST, 19 DOC/DOCX, 15 HTM/MHT, 15 TXT, and 1 ASC, 1OW, ASV, OPJ. We began appraising and assessing the collection by combing through each file within a folder looking for similarities in variables and consistently formatted data. We eventually realized that the files in each subfolder were extracted data from the 'gmaster' files (1 through 6) that were in the top folder. We believe that the files in each subfolder were data exports and dialog commands from the statistical analysis software, SAS. We decided that combining the 'gmaster' files to create a unified data set would be the best option to facilitate reuse. The additional files we deemed supplementary and may not be suitable for dissemination because they document Chaney's individual data analysis steps without sufficient contextual information

about the process. However, the supplementary files should be retained internally as part of the SIP and AIP in accordance with the OAIS framework.

The ‘gmaster’ data set was spread across six separate TXT files with tab-delimited values, however, the columns were not identified. Because portions of the data in the gmaster files were further described in the supplementary files, we were able to identify most of the variables. Below is the representation of the file, gmaster1.txt, open in Notepad.



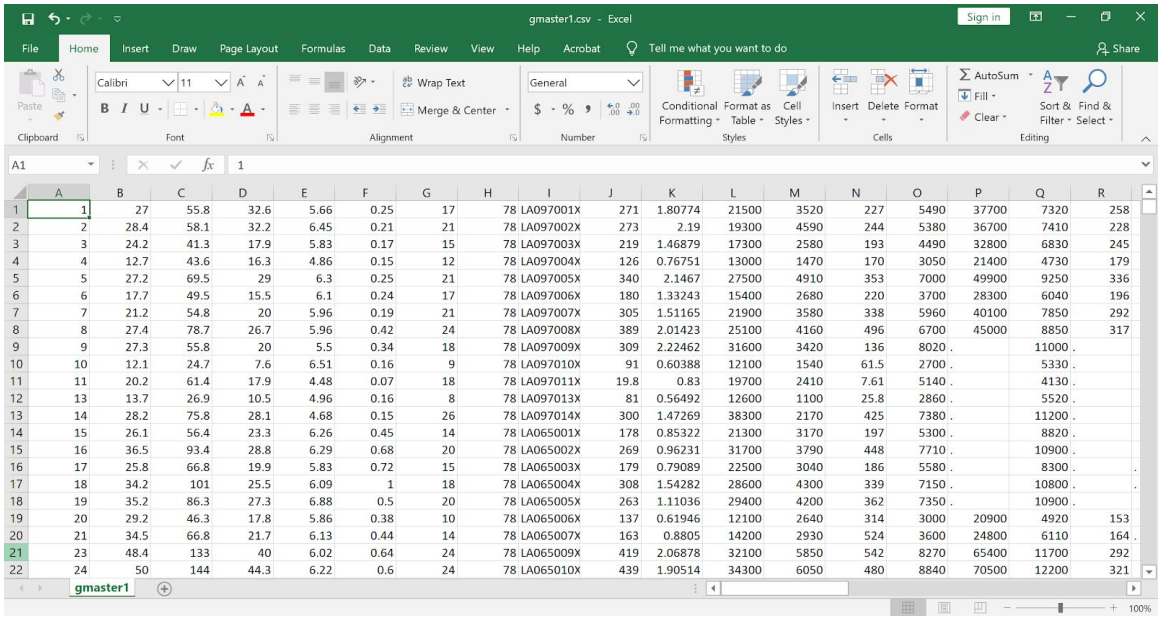
gmaster1.txt - Notepad

File Edit Format View Help

1	27.0000	55.800	32.6000	5.66	0.250	17	78	LA097001X	271.0	1.80774	21500	3520	227.00	5490	37700	7320	258.0	0.915
2	28.4000	58.100	32.2000	6.45	0.210	21	78	LA097002X	273.0	2.19000	19300	4590	244.00	5380	36700	7410	228.0	0.894
3	24.2000	41.300	17.9000	5.83	0.170	15	78	LA097003X	219.0	1.46879	17300	2580	193.00	4490	32800	6830	245.0	0.919
4	12.7000	43.600	16.3000	4.86	0.150	12	78	LA097004X	126.0	0.76751	13000	1470	170.00	3050	21400	4730	179.0	0.916
5	27.2000	69.500	29.0000	6.30	0.250	21	78	LA097005X	340.0	2.14670	27500	4910	353.00	7000	49900	9250	336.0	0.910
6	17.7000	49.500	15.5000	6.10	0.240	17	78	LA097006X	180.0	1.33243	15400	2680	220.00	3700	28300	6040	196.0	0.910
7	21.2000	54.800	20.0000	5.96	0.190	21	78	LA097007X	305.0	1.51165	21900	3580	338.00	5960	40100	7850	292.0	0.924
8	27.4000	78.700	26.7000	5.96	0.420	24	78	LA097008X	389.0	2.01423	25100	4160	496.00	6700	45000	8850	317.0	0.904
9	27.3000	55.800	20.0000	5.50	0.340	18	78	LA097009X	309.0	2.22462	31600	3420	136.00	8020	.	11000	.	0.909
10	12.1000	24.700	7.6000	6.51	0.160	9	78	LA097010X	91.0	0.60388	12100	1540	61.50	2700	.	5330	.	0.909
11	20.2000	61.400	17.9000	4.48	0.070	18	78	LA097011X	19.8	0.83000	19700	2410	7.61	5140	.	4130	.	0.915
13	13.7000	26.900	10.5000	4.96	0.160	8	78	LA097013X	81.0	0.56492	12600	1100	25.80	2860	.	5520	.	0.918
14	28.2000	75.800	28.1000	4.68	0.150	26	78	LA097014X	300.0	1.47269	38300	2170	425.00	7380	.	11200	.	0.908
15	26.1000	56.400	23.3000	6.26	0.450	14	78	LA065001X	178.0	0.85322	21300	3170	197.00	5300	.	8820	.	0.904
16	36.5000	93.400	28.8000	6.29	0.680	20	78	LA065002X	269.0	0.96231	31700	3790	448.00	7710	.	10900	.	0.931
17	25.8000	66.800	19.9000	5.83	0.720	15	78	LA065003X	179.0	0.79089	22500	3040	186.00	5580	.	8300	.	.
18	34.2000	101.000	25.5000	6.09	1.000	18	78	LA065004X	308.0	1.54282	28600	4300	339.00	7150	.	10800	.	.
19	35.2000	86.300	27.3000	6.88	0.500	20	78	LA065005X	263.0	1.11036	29400	4200	362.00	7350	.	10900	.	0.923
20	29.2000	46.300	17.8000	5.86	0.380	10	78	LA065006X	137.0	0.61946	12100	2640	314.00	3000	20900	4920	153.0	0.916
21	34.5000	66.800	21.7000	6.13	0.440	14	78	LA065007X	163.0	0.88050	14200	2930	524.00	3600	24800	6110	164.0	.
23	48.4000	133.000	40.0000	6.02	0.640	24	78	LA065009X	419.0	2.06878	32100	5850	542.00	8270	65400	11700	292.0	0.920
24	50.0000	144.000	44.3000	6.22	0.600	24	78	LA065010X	439.0	1.90514	34300	6050	480.00	8840	70500	12200	321.0	0.916
25	53.9000	132.000	43.6000	5.74	0.720	24	78	LA065011X	477.0	2.31033	33300	6520	492.00	8750	52900	11200	298.0	0.899
26	53.0000	138.000	43.8000	6.91	0.600	24	78	LA065012X	473.0	1.44542	35100	6850	928.00	9520	67800	11100	263.0	0.922
27	53.3000	129.000	41.0000	6.37	0.710	29	78	LA065013X	481.0	2.12332	32800	6550	512.00	8050	68500	11000	262.0	0.893
28	48.3000	122.000	42.5000	6.28	0.790	25	78	LA065014X	434.0	2.91031	30000	6860	450.00	7340	61600	10600	253.0	0.915
29	56.8000	121.000	42.8000	6.11	0.490	27	78	LA065015X	453.0	1.58957	32800	5670	509.00	8550	66500	10900	265.0	.
31	60.2000	142.000	56.4000	7.54	1.000	32	78	LA065017X	461.0	2.24799	36600	11300	528.00	9870	79000	11500	358.0	0.922
32	49.9000	182.000	57.4000	7.56	0.950	31	78	LA065018X	447.0	2.08046	44400	10900	375.00	11300	.	15100	.	.
33	52.0000	173.000	55.3000	7.52	0.980	26	78	LA065019X	456.0	2.06878	44400	10100	480.00	11100	.	14300	.	0.910
34	47.6000	151.000	50.3000	7.37	0.780	19	78	LA065020X	466.0	2.09215	44400	8570	348.00	11100	.	14700	.	0.904
35	46.6000	152.000	50.3000	7.18	0.780	16	78	LA065021X	444.0	2.16228	44400	8540	335.00	11200	.	14300	.	0.898
36	4.8000	12.600	3.8200	5.19	0.170	7	78	LA053001X	64.0	0.71297	7340	562	135.00	787	.	1960	.	0.903
37	4.8200	9.340	-0.6000	4.70	0.240	9	78	LA053002X	69.0	0.74414	5380	533	37.70	826	.	2480	.	0.913
38	5.8000	13.600	1.3000	4.52	0.180	10	78	LA053003X	63.0	0.74803	11600	506	290.00	858	.	2080	.	0.911

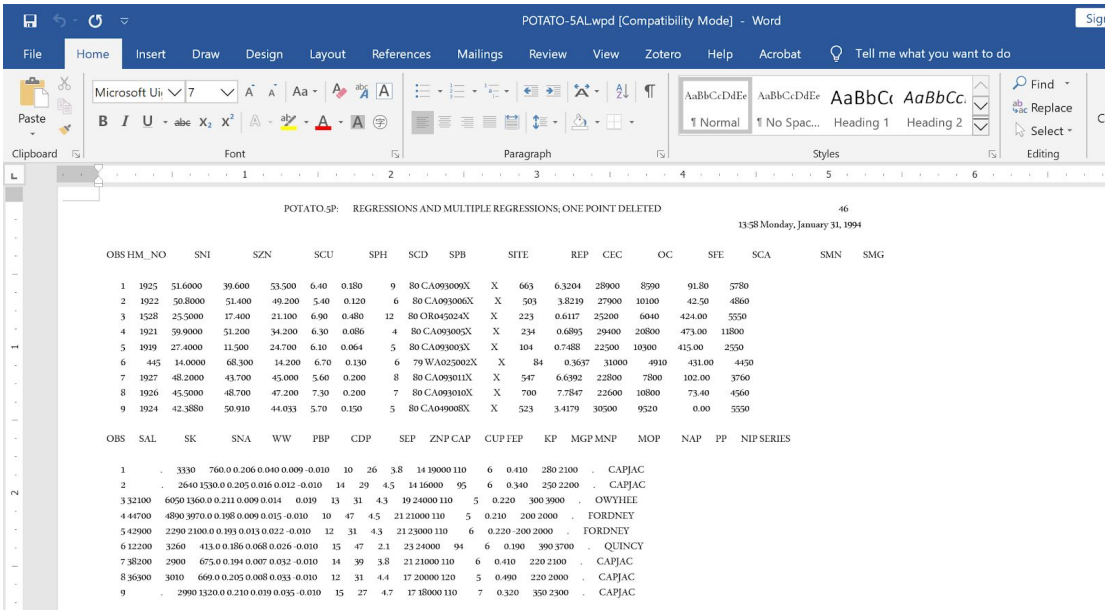
Ln 1, Col 1 100% Windows (CRLF) UTF-8

The representation looks like a table, so we used Microsoft Excel ‘Text to Columns’ tool to open and parse possible columns.



While doing the file inventory of Rufus Chaney's data set, we noticed a file, POTATO-5AL.wpd, contains a table with possible entity names. The file path of this file is: Rufus Chaney\Potato-2012\POTATO-5AL.wpd.

POTATO-5AL.wpd is a text file for Corel WordPerfect. It is recommended to use Microsoft Word to open it. Below is how Microsoft Word presents POTATO-5AL.wpd.



Eventually, we downloaded the trial version of Corel WordPerfect to confirm the contents, in one example, a graphic table was only visible when opened with WordPerfect.

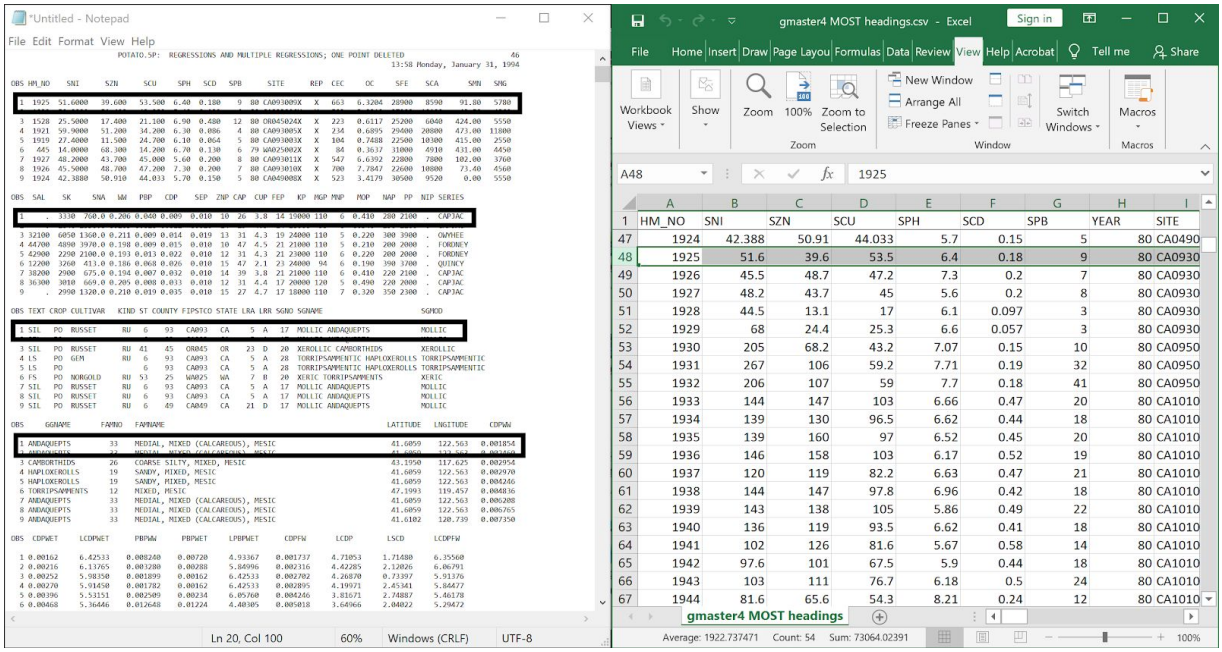
As presented above, the entity list does not line up with the data nicely. Because the data seems to be all text, we copied and pasted it into Notepad to see if it will present the data better. Below is the outcome. Please note that if opening POTATO-5AL.wpd file directly with Notepad, the content will not render in a legible way.

*Untitled - Notepad

File Edit Format View Help

```
                    POTATO.5P:  REGRESSIONS AND MULTIPLE REGRESSIONS; ONE POINT DELETED                    46
                                                                                  13:58 Monday, January 31, 1994

 OBS HM_NO    SNI      SZN      SCU     SPH   SCD   SPB      SITE      REP  CEC     OC     SFE    SCA      SMN    SMG

   1  1925  51.6000   39.600   53.500  6.40  0.180    9  80 CA093009X   X   663   6.3204  28900   8590    91.80   5780
   2  1922  50.8000   51.400   49.200  5.40  0.120    6  80 CA093006X   X   503   3.8219  27900  10100    42.50   4860
   3  1528  25.5000   17.400   21.100  6.90  0.480   12  80 OR045024X   X   223   0.6117  25200   6040   424.00   5550
   4  1921  59.9000   51.200   34.200  6.30  0.086    4  80 CA093005X   X   234   0.6895  29400  20800   473.00  11800
   5  1919  27.4000   11.500   24.700  6.10  0.064    5  80 CA093003X   X   104   0.7488  22500  10300   415.00   2550
   6   445  14.0000   68.300   14.200  6.70  0.130    6  79 WA025002X   X    84   0.3637  31000   4910   431.00   4450
   7  1927  48.2000   43.700   45.000  5.60  0.200    8  80 CA093011X   X   547   6.6392  22800   7800   102.00   3760
   8  1926  45.5000   48.700   47.200  7.30  0.200    7  80 CA093010X   X   700   7.7847  22600  10800    73.40   4560
   9  1924  42.3880   50.910   44.033  5.70  0.150    5  80 CA049008X   X   523   3.4179  30500   9520     0.00   5550

 OBS  SAL     SK      SNA   WW   PBP    CDP     SEP   ZNP CAP  CUP FEP   KP   MGP MNP   MOP   NAP   PP  NIP SERIES

   1    .    3330   760.0 0.206 0.040 0.009  0.010   10  26   3.8  14 19000 110   6  0.410  280 2100   .  CAPJAC
   2    .    2640  1530.0 0.205 0.016 0.012  0.010   14  29   4.5  14 16000  95   6  0.340  250 2200   .  CAPJAC
   3 32100   6050  1360.0 0.211 0.009 0.014  0.019   13  31   4.3  19 24000 110   5  0.220  300 3900   .  OWYHEE
   4 44700   4890  3970.0 0.198 0.009 0.015  0.010   10  47   4.5  21 21000 110   5  0.210  200 2000   .  FORDNEY
   5 42900   2290  2100.0 0.193 0.013 0.022  0.010   12  31   4.3  21 23000 110   6  0.220  200 2000   .  FORDNEY
   6 12200   3260   413.0 0.186 0.068 0.026  0.010   15  47   2.1  23 24000  94   6  0.190  390 3700   .  QUINCY
   7 38200   2900   675.0 0.194 0.007 0.032  0.010   14  39   3.8  21 21000 110   6  0.410  220 2100   .  CAPJAC
   8 36300   3010   669.0 0.205 0.008 0.033  0.010   12  31   4.4  17 20000 120   5  0.490  220 2000   .  CAPJAC
   9    .    2990  1320.0 0.210 0.019 0.035  0.010   15  27   4.7  17 18000 110   7  0.320  350 2300   .  CAPJAC

 OBS TEXT CROP CULTIVAR    KIND ST COUNTY FIPSTCO STATE LRA LRR SGNO SGNAME                          SGMOD

   1 SIL   PO  RUSSET      RU   6    93    CA093   CA     5   A   17  MOLLIC ANDAQUEPTS             MOLLIC
   2 SIL   PO                    6    93    CA093   CA     5   A   17  MOLLIC ANDAQUEPTS             MOLLIC
   3 SIL   PO  RUSSET      RU  41    45    OR045   OR    23   D   20  XEROLLIC CAMBORTHIDS          XEROLLIC
   4 LS    PO  GEM         RU   6    93    CA093   CA     5   A   28  TORRIPSAMMENTIC HAPLOXEROLLS  TORRIPSAMMENTIC
   5 LS    PO                    6    93    CA093   CA     5   A   28  TORRIPSAMMENTIC HAPLOXEROLLS  TORRIPSAMMENTIC
   6 FS    PO  NORGOLD     RU  53    25    WA025   WA     7   B   20  XERIC TORRIPSAMMENTS          XERIC
```

After this transformation, we were able to start matching data and fill in entity names in gmaster files. We presumed that the HM_NO matches with the number of the first column in gmaster files. The top row of the Microsoft Excel screen was created after matching the same values in both files. However, some columns in gmaster files are still unidentified as there are no matching values in POTATO-5AL.wpd or other supplementary files.



In order to create a single data set from the six gmaster files, the variables (or column headings) need to be in the same order and format to allow for future data manipulation. A consistent heading order was first established by assigning a number to each variable, for example, the HM_NO is the first variable in every gmaster file (see graphic below). Using this standard order, columns could be arranged using Excel. In some gmaster files, data variables were combined in one cell while in others they were not. To separate the clearly identifiable data variables, various Excel functions were used (RIGHT, LEFT, LEN).

	A	B	C	D	E	F	G	H	I	J	K	L	M
1		1	2	3	4	5	6	7	8	9	10	11	12
2		HM_NO	SNI	SZN	SCU	SPH	SCD	SPB	YEAR	SITE	CEC	OC	SFE	SC
3	GMASTER 1	1	27	55.8	32.6	5.66	0.25	17	78	LA097001X	271	1.80774	21500	
4	GMASTER 1	2	28.4	58.1	32.2	6.45	0.21	21	78	LA097002X	273	2.19	19300	
5	GMASTER 1	3	24.2	41.3	17.9	5.83	0.17	15	78	LA097003X	219	1.46879	17300	
6	GMASTER 1	4	12.7	43.6	16.3	4.86	0.15	12	78	LA097004X	126	0.76751	13000	
7	GMASTER 1	5	27.2	69.5	29	6.3	0.25	21	78	LA097005X	340	2.1467	27500	
8	GMASTER 1	6	17.7	49.5	15.5	6.1	0.24	17	78	LA097006X	180	1.33243	15400	
9	GMASTER 1	7	21.2	54.8	20	5.96	0.19	21	78	LA097007X	305	1.51165	21900	
10	GMASTER 1	8	27.4	78.7	26.7	5.96	0.42	24	78	LA097008X	389	2.01423	25100	
11	GMASTER 1	9	27.3	55.8	20	5.5	0.34	18	78	LA097009X	309	2.22462	31600	
12	GMASTER 1	10	12.1	24.7	7.6	6.51	0.16	9	78	LA097010X	91	0.60388	12100	
13	GMASTER 1	11	20.2	61.4	17.9	4.48	0.07	18	78	LA097011X	19.8	0.83	19700	

While we were able to identify most of the variables across the data set, there were several that we could not identify. We have left this data on the far right of the combined file and clearly labeled the headings as unidentified. The gmaster 6 file was particularly difficult to identify and many of the variables from the file are unidentified because the data could not be matched to any other documentation. However, some of the gmaster 6 unidentified variables appear consistent with identified variables, we are unsure if this is enough evidence to label them.

Because the majority of the files in the collection were in a proprietary format that may not be accessible in the future or are already difficult to access. The LOG, LST, SAS can be converted to TXT by changing the file extension. This only works with plain text files already encoded with UTF-8, ASCII or similar encoding, but it will not work with advanced markup encoding like that in DOCX, PDF, etc. Batch conversion can be done quickly using Command Prompt with the following command: “forfiles /S /M *.sas /C "cmd /c rename @file @fname.txt”. In this command, switch .sas with the desired file type (e.g. LOG, LST, XML, etc.). There was one issue with a converted file name becoming the same as an existing file name. To remedy this, the original file type was added to the converted file name, for example, SP-1-PROG.lst became SP-1-PROG-LST.txt.

The process of converting the files to sustainable formats required a variety of software: Microsoft Word to convert DOC/DOCX files to TXT, Microsoft Excel to convert uniform data to CSV, Adobe Acrobat Pro to convert PDF to PDF/A, and Corel WordPerfect to open and convert WPD files to TXT. Converting 262 files took roughly three full days to complete, however, this timeframe was heavily dependent on the mix of file types and could increase or decrease based on the complexity of file formats. After processing, the final collection contains 237 files in four major formats, 20 CSV, 164 TXT, 36 PDF/A, 15 HTM. Only 23 of the original 262 files were already in a sustainable format, 26 files were deleted because they were exact

duplicates of other files. There are two 'mystery' file formats (OPJ, 1OW) that could not be analyzed but have been retained for possible future exploitation.

While the NAL has not implemented digital preservation practices, to fulfill the OAIS component of our Processing Guide, we prepared the collection with a BagIt application. BagIt is an archival file structure and hierarchy standard for preserving digital content designed by the Library of Congress. It attaches checksums and preservation and description metadata to the files being preserved in a standard file system which can easily be navigated and understood without advanced knowledge of BagIt standards (Kunze et al., 2016). The prepared files and file structure can be crosswalked to any file system. We used BDBag, a simple GUI, that automatically creates the required documentation and structure in accordance with BagIt (FAIR Research, 2019). BDBag is one of several software that can create the BagIt package.

The Chaney collection still needs description, we were unfortunately unable to interview Chaney due to COVID-19 restrictions. Chaney could clear up uncertainty relating to unidentified variables in the data set, provide description/metadata for the data set, and identify publications that resulted from these data.

VI. Recommended Next Steps

The National Agricultural Library has prioritized digital data access to best serve the information needs of agricultural scholars and the Department. Currently, the NAL has not implemented digital preservation practices for Ag Data Commons or the Digital Collections. A previous UMD digital curation fellow researched the NAL's digital preservation system and recommended best practices in the report, "Digital workflows at the National Agricultural Library and implications for preservation." The NAL is currently relying on weekly and annual tape backups, however, this does not address preservation concerns such as fixity (ensuring that data does not change over time) and geographic distribution of backups (Daniels, 2018). The report recommends updating the Fedora software that supports the NAL's Unified Repository to a version with built in preservation tools. To prepare the OAIS packages for individual data sets, we recommend the Library of Congress designed BagIt file standards, this will allow users to verify data integrity. We prepared the Chaney collection files with BDBag, an open source software that automatically creates the necessary metadata file structure, and checksums in accordance with BagIt standards (FAIR Research, 2019). There are several GUIs and Python based programs for BagIt implementation.

Policy or standards should be created in regard to the level of data cleaning and curation that will be conducted internally for data rescue projects before inclusion on Ag Data Commons. It is important to consider how much work will be done by staff to enable reuse and how much data cleaning or manipulation will be left to researchers. When processing the Chaney collection, we generally only took steps to reformat data to sustainable formats and did not curate and thoroughly describe every file because of time constraints and reuse expectations. The main data set derived from the six gmaster files, however, we significantly organized, cleaned, and converted the files to a single CSV. Balancing the value of potential data reuse with the effort

required to make the data usable will be a significant consideration of future data rescue appraisal.

To assist appraisal of complex data, establishing a network of USDA researchers and field experts would enable quick consultation for appraisal decisions. The data created and used by the USDA are often complex with specific knowledge required to fully understand the data, this barrier to rapid appraisal and description can only be cleared with collaboration and communication with experts. This outreach could be done in conjunction with a campaign for data management best practices to reduce the need to rescue data in the future.

VII. References

Akmon, D., Zimmerman, A., Daniels, M., & Hedstrom, M. (2011). The application of archival concepts to a data-intensive environment: Working with scientists to understand data management and preservation needs. *Archival Science*, *11*(3–4), 329–348. https://doi.org/10.1007/s10502-011-9151-4

Allen, L., Stewart, C., & Wright, S. (2017). Strategic open data preservation: Roles and opportunities for broader engagement by librarians and the public. *College & Research Libraries News*, *78*(9). https://doi.org/10.5860/crln.78.9.482

Altman, M., Adams, M., Crabtree, J., Donakowski, D., Maynard, M., Pienta, A., & Young, C. (2009). Digital Preservation through archival collaboration: The Data Preservation Alliance for the Social Sciences. *The American Archivist*, *72*(1), 170–184. https://doi.org/10.17723/aarc.72.1.eu7252lhnrp7h188

Atwater, W. O., Bryant, A. P. (1906). The chemical composition of American food materials. Office of Experiment Stations - Bulletin No. 28. Available from Internet Archive: https://archive.org/details/CAT31006482

Belovari, S. (2017). Expedited digital appraisal for regular archivists: An MPLP-type approach. *Journal of Archival Organization*, *14*(1–2), 55–77. https://doi.org/10.1080/15332748.2018.1503014

Brunet, M., & Jones, P. (2011). Data rescue initiatives: Bringing historical climate data into the 21st century. *Climate Research*, *47*(1), 29–40. https://doi.org/10.3354/cr00960

Cooper, D., Bankston, S., Bracke, M. S., Callahan, B., Chang, H., Delserone, L. M., & Diekmann, F. (2017). Supporting the changing research practices of agriculture scholars. Ithaka S+R. https://doi.org/10.18665/sr.303663

Coville, F. V. (1907-1908). Frederick Vernon Coville blueberry records / box 1. Available from Internet Archive: https://archive.org/details/cat31321157001

Coville, F. V. (1910). Experiments in blueberry culture. Bureau of Plant Industry - Bulletin No. 193. U.S. Department of Agriculture. https://archive.org/details/experimentsinblu193covi

Coville, F. V. (1937). Improving the wild blueberry. In U.S. Department of Agriculture. (Pub.), *Yearbook of Agriculture, 1937* (pp. 559-574). Washington, D.C. Government Printing Office. https://archive.org/details/yoa1937

Daniels, M. (2018). Digital workflows at the National Agricultural Library and implications for preservation. University of Maryland. http://hdl.handle.net/1903/26356

Downs, R. R., & Chen, R. S. (2017). Curation of scientific data at risk of loss: Data rescue and dissemination. Association of College and Research Libraries. https://doi.org/10.7916/D8W09BMQ

FAIR Research. (2019). *Big data bag utilities (BDBag)*. GitHub. https://github.com/fair-research/bdbag

Faulder, E., et al. (2018). Digital processing framework. Cornell University Library. https://hdl.handle.net/1813/57659

Faundeen, J. L., & Oleson, L. R. (2007). Scientific data appraisals: The value driver for preservation efforts. In *Proceedings of PV 2007 International Conference*. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.94.8035&rep=rep1&type=pdf.

Federal data strategy: What are the principles?. (n.d.) Retrieved May 1, 2020, from https://strategy.data.gov/principles/

Google Scholar. (n.d.). *Rufus Chaney*. Retrieved June 1, 2020, from https://scholar.google.com/citations?user=ztOfDNoAAAAJ

Greene, M., & Meissner, D. (2005). More Product, Less Process: Revamping Traditional Archival Processing. *The American Archivist*, *68*(2), 208–263. https://doi.org/10.17723/aarc.68.2.c741823776k65863

Gutmann, M., et al. (2009). From preserving the past to preserving the future: The Data-PASS project and the challenges of preserving digital social science data. *Library Trends, 57*(3), 315-337. https://doi.org/10.1353/lib.0.0039.

Gutmann, M., Schürer, K., Donakowski, D., & Beedham, H. (2004). The selection, appraisal, and retention of digital social science data. *Data Science Journal, 3*, (209-221). http://www.digitalpreservation.gov/partners/documents/data-pass_selection_data.pdf

Higgins, S. (2008). The DCC curation lifecycle model. *International Journal of Digital Curation 3*(1). https://doi.org/10.2218/ijdc.v3i1.48.

Holdren, J. P. (2013). Memorandum for the heads of Executive Departments and Agencies. Executive Office of the President, Office of Science and Technology Policy. Retrieved May 1, 2020, from https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/ostp_public_acce

ss_memo_2013.pdf.

Hsu, L., Lehnert, K. A., Goodwillie, A., Delano, J. W., Gill, J. B., Tivey, M. A., . . . Arko, R. A. (2015). Rescue of long-tail data from the ocean bottom to the Moon: IEDA Data Rescue Mini-Awards. *GeoResJ*, *6*, 108–114. https://doi.org/10.1016/j.grj.2015.02.012

Janz, M. M. (2018). Maintaining access to public data: Lessons from data refuge [Preprint]. LIS Scholarship Archive. https://doi.org/10.31229/osf.io/yavzh

Johnston, L. R. (Ed.). (2017). *Curating research data: Volume one: Practical strategies for your digital repository* [ebook]. Association of College and Research Libraries, American Library Association. http://www.ala.org/acrl/sites/ala.org.acrl/files/content/publications/booksanddigitalresources/digital/9780838988596_crd_v1_OA.pdf

Kunze, J., Littman, J., Madden, L., Summers, E., Boyko, A., & Vargas, B. (2016). The BagIt file packaging format (V0.97) draft-kunze-bagit-14. Internet Engineering Task Force. https://tools.ietf.org/id/draft-kunze-bagit-14.txt

Lafferty-Hess, S., & Christian, T. M. (2017). More data, less process? The applicability of MPLP to research data. *IASSIST Quarterly*, *40*(4), 6. https://doi.org/10.29173/iq907

Lavoie, B. (2014). *The Open Archival Information System (OAIS) reference model: Introductory guide* (2nd ed.). Digital Preservation Coalition. https://doi.org/10.7207/twr14-02

Martínez-Navarro, A. G., et al. (2019). Heats of combustion representative of the carbohydrate mass contained in fruits, vegetables, or cereals. *Food Science & Nutrition 7*(9), 3119-3127. https://doi.org/10.1002/fsn3.1175

McCarthy, S. (2019). Data rescue project. USDA. National Agricultural Library.

McGovern, N. Y. (2017). Data rescue: Observations from an archivist. *ACM SIGCAS Computers and Society*, *47*(2), 19–26. https://doi.org/10.1145/3112644.3112648

Park, E. G., Burr, G., Slonosky, V., Sieber, R., & Podolsky, L. (2018). Data rescue archive weather (DRAW): Preserving the complexity of historical climate data. *Journal of Documentation*, *74*(4), 763–780. https://doi.org/10.1108/JD-10-2017-0150

Sánchez-Peña, M. J., et al. (2016). Calculating the metabolizable energy of macronutrients: A critical review of Atwater's results. *Nutrition reviews 75*(1), 37-48. https://doi.org/10.1093/nutrit/nuw044

Slonosky, V., Sieber, R., Burr, G., Podolsky, L., Smith, R., Bartlett, M., . . . Fabry, F. (2019). From books to bytes: A new data rescue tool. *Geoscience Data Journal*, *6*(1), 58–73. https://doi.org/10.1002/gdj3.62

Thompson, C. A., Davenport Robertson, W., & Greenberg, J. (2014). Where have all the scientific data gone? LIS perspective on the data-at-risk predicament. *College & Research Libraries*, *75*(6), 842–861. https://doi.org/10.5860/crl.75.6.842

Transcribing on the Transcription Center. (2020). Smithsonian Transcription Center. Retrieved March 1, 2020, from https://transcription.si.edu/sites/default/files/uploads/transcribing_on_the_transcription_center_-_quick_guide_1.pdf

Yakel, E. (2007). Digital curation. *OCLC Systems & Services: International Digital Library Perspectives*, *23*(4), 335–340. https://doi.org/10.1108/10650750710831466

VIII. Appendix

1. Rob Griesbach Presentation Re: Coville Blueberry Collection - Key Points

These notes provide a more detailed outline of Robert Griesbach's presentation about USDA blueberries on March 23, 2020. The complete transcript can be found below.

- There are a lot of resources for blueberry research in the NAL. USDA created the blueberry industry, and the starting point was Coville's blueberry research with his background in botany and taxonomy.
- Before the existence of commercial blueberry production: people picked wild blueberries. They were rare on the market.
- At the turn of the century (1900s), USDA was looking for what kinds of crops they can grow in the United States.
- Coville's methods:
    - Picking wild plants
    - There are records documenting detailed locations of blueberry plants. They were first planted at the Washington Mall, and later to the Arlington Farm, where the Pentagon is now.
- Coville's discovery:
    - Blueberries require acidic soil, which was documented in his notes. Fertilizers that can add acidity to soil were then needed for blueberries.
    - Propagating of blueberries using cuttings: using vegetative propagation so they get uniformity. Propagating by seeds creates a lot of variabilities.
    - Blueberry requires cold treatment
- Collaboration with Elizabeth White, a cranberry grower: After publishing bulletins of his discovery of blueberry cultivation, Coville received a lot of interest in blueberries and White was one of them.
- Coville and White's discovery:
    - Blueberries are self-sterile and require cross pollination.
    - In spite of valuable discoveries about blueberry cultivation, public interest still had not taken off. Coville and White started finding the best wild blueberry bushes and propagating those. They gave out prizes for people who brought good wild blueberry plants.
- Genetics wasn't a term until 1909. The cross pollination Coville was doing was pioneering. He was the first chairman of the American Genetics and American breeding Association, which was founded by the Secretary of Agriculture, and is now the American Genetics Society Association. He was also the senior editor of the Journal of Heredity, promoting plant breeding.
- Coville's notes keep detailed breeding records, including description of particular plants.
- By 1921, commercial production of blueberries had started and was getting more people's attention.
- Crosses took multiple years to come up with good cultivars and seedings. After Coville passed away, George Darrow continued the breeding program and also started taking the

notebooks over. Darrow started to look at chromosome numbers, and chromosomal compatibility which Coville wasn't doing before.

- Blue crop, released in 1941 and still widely grown today, was remarkable because it is disease-resistant and more hardy. It is also great in flavor, size and produces good yields. In the presentation, Dr. Griesbach showed us the pedigree of Blue crop, which includes cultivars that were released by Coville.
- USDA blueberry varieties are quite important in the blueberry industry. For instance, they took up 75% of commercial acreage of blueberry in 2010. Some of the varieties grown today were some of the early releases.
- Modern breeding and geneticists: Don Scott took over Darrow. In the 1970s, USDA started expanding blueberry production in non-conventional areas, such as heat-tolerant blueberries, blueberries that don't require cold treatment, etc. Scientists were experimenting with blueberry cultivation across the US. In the 1990s, Jeannie Rowland, a molecular biologist, started to look at genetic markers for gene selection. In 2000, USDA moved research to genomic and molecular areas, identifying specific genes that are linked to the traits of plants. This approach shortens the experimentation time, as scientists don't need to wait for and document the culture over time.
- Collaboration between USDA and universities: Collaboration with university collaborators started in the 1960s and continues to date. USDA scientists make crosses of blueberry in Beltsville and New Jersey and send good seedlings to different collaborators. USDA can be seen as the blueberry breeder for most of the country.
- Acquisition of Coville's blueberry notes: It took three or four years for Dr. Griesbach to convince the original holding institution, the Small Fruit Lab, to give the notes to the NAL. He convinced them that long-term preservation is needed for the blueberry notes. Another collection that was acquired during the same time was the negatives.
- Dr. Griesbach also brought up other related resources, such as newsletters, but currently they are scattered in the library and are not compiled together if a user needs it.
- Pedigree information that resides in Coville's notes appears to be quite important, especially for cultivars that are still prevalent in use today, such as the cultivar, Blue crop. Locating all parentage of Blue crop in the notebooks may gain a lot of interest from the public and can get resources for the project.
- Potential users for Coville's blueberry notes:
    - The general public: plants pedigree of known cultivars. People start getting interested in where their food comes from.
    - Horticultural field: some of the first fertilizers can be found in the notebooks, which can be devoted to horticultural research.
    - Scientists: genetics field, and how certain traits of the plant were inherited. Scientists may be interested to see how they can further improve the next generation of blueberries using genetic research.
- Other areas of research for blueberries in general: the economic impact of blueberry research, such as the jobs the blueberry industry has created, etc.
- Dr. Griesbach mentioned that it really depends on the target audience how the NAL makes certain information available. This may be a question to be answered by current members at the NAL.

2. *Rob Griesbach Presentation Re: Coville Blueberry Collection - Memo*

Presentation overview:

Due to COVID-19, our scheduled interview on March 23, 2020 was moved from in-person meeting to an online presentation by Rob Griesbach. All participants received a presentation file, and the presentation was conducted over phone dial-in. The presentation took about an hour, 80% presentation and the rest was open for comment.

The presentation focused on how USDA started the blueberry industry. Coville's effort was the initial start, because he found out how to cultivate blueberries, instead of picking them in the wild. His findings include using acid soil, cold treatments, cross pollination which were recorded in his blueberry notes. George Darrow was mentioned in the presentation as he continued Coville's research after 1937.

Coville's meticulous note-taking was mentioned by Dr. Griesbach in the presentation, such as location, pedigrees, etc. Many parentages Coville discovered are still used by the contemporary blueberry industry and growers.

After the presentation, we were able to ask a couple of questions in regards to the data rescue project: what and how to digitally capture information on the notes for the purpose of reuse, and who the potential users are.

1. In terms of digitally capturing data in Coville's notebooks, Dr. Griesbach emphasized again the important values of the pedigree information. Dr. Griesbach suggested that only portions of the notes would be of interest such as certain cultivars, and did not advise digitizing the entire collection. However, it would likely be easiest (and reduce handling) to organize and highlight significant notes if we have digital representations of Coville's notebooks.
2. Designated community: Dr. Griesbach brought out three types of potential users: (1) The general public, because the public is more interested in blueberries now, and may be interested to know the history. (2) Horticultural industry, like who made fertilizers, as Coville's notes provide in-depth information on different horticultural knowledge. (3) Scientists and geneticists, as the detailed pedigrees were documented.

Recommendations for processing Coville's blueberry notes:

- Defining the designated community:

  Dr. Griesbach emphasized that it really depends on who the audience is, in order to decide how and what to make available. These decisions would have to be made by the National Agricultural Library.

- Digitization of complete Coville's blueberry notes collection:

  To capture the complete pedigree data, and to reduce physically handling the fragile materials, we recommend digitizing the full collection of the notes.

- Produce transcription of Coville's blueberry notes:

Along with or after the completion of digitizing the notes, transcribing the notes would be the following steps. A transcription guide was generated during the data rescue project. It can be a useful guide for people who work on the transcription. Crowdsourcing tool may be considered for this project as well, because of the growing public interests in blueberries.

- Dissemination of data:

  After complete pedigree information of certain cultivars, such as Bluecrop, is identified throughout Coville's notes, a digital collection with the pedigree and interactive links to the individual notes as a source would be a great way to display the materials.

- Public engagement:

  Dr. Griesbach noted that some of the first USDA blueberry research was conducted on the National Mall, which would be a great connection to engage the general public. The aforementioned digital collection of blueberry notes can be used for public engagement as well. By doing so, the NAL can raise awareness and promote the reuse of this valuable collection.

- Comparison and integration of contemporary data:

  As research on blueberries is still a prominent field at the USDA, scientific data of blueberry research are still being generated. It would be worth speaking with current blueberry researchers to understand their practices and capture their data and notes to add to the collection.

From the experiences of this communication, what can be done better:

1. Topics of the interview should be broadened and can have more in-depth and detailed conversation on reuse and designated communities (potential users).

   The time for the interview in the call was very short. Dr. Griesbach did mention it depends on who the user group is in order to decide what and how to digitally capture information from the notes. If we had more time, we could possibly have Dr. Griesbach to elaborate more on different user groups' needs. Different information is important for different users.

   - General public: What's their current interest and specifically what aspects of blueberries? Formats of presentation may have to be more user-friendly.
   - Horticultural industry: Specifically what aspects of blueberry research are more important to them? How do they access data?
   - Scientists: what's important for current research questions. What format would be more useful? How do they access data?

2. Establish direct communication with the interviewee from the start.

   As the connection and communication was established by the head of special collections, Susan Fugate, she was the only person communicating with Rob before the virtual meeting took place. Many people from different divisions at the NAL were also invited to the call, so interests of blueberry and understanding of Coville's notes varied widely.

Asking questions specific to physical blueberry notes and about data rescue was not quite suitable in the context.

3. Communicate with the interviewee what data rescue is before the meeting. Possibly let the interviewee review the questions. Data rescue focuses on reuse of the data, how it can be reused and in what forms. Dr. Griesbach presented the “importance” of the note, but how we can reappraise would need to be derived from the presentation and short Q&A.
4. In-person meeting is most preferred.

   The COVID-19 pandemic hindered us from meeting in-person, so it was difficult to ask about specific details without seeing the actual blueberry notes.

3. *Coville Blueberry Collection Transcription Style Guide*

This style guide was created with references to the transcribing guide from the Smithsonian Transcription Center.

About Frederick V. Coville's Blueberry Notebooks:

The nutrient values of blueberries have been known by the general public for years. What we often take for granted, however, is how the blueberry industry came into existence. Frederick V. Coville (1867-1937) was one of the first breeders who found out how to cultivate blueberries. Coville kept detailed research and field notes that include some of the earliest discoveries about blueberries cultivation, including its requirements of acidic soil, cold treatment, cross-pollination, detailed pedigrees etc. Coville's successor, George Darrow continued his notebooks. These notes, six linear feet in total, are currently held by the Special Collection of the National Agricultural Library. Ranging from 1907 to 1938, the blueberry notebooks provide abundant information for scientists, geneticists, horticultural scholars and people who are interested in blueberries. As the notebooks were handwritten, and are in fragile conditions, data migration from the analog to the digital would greatly facilitate the use of the materials.

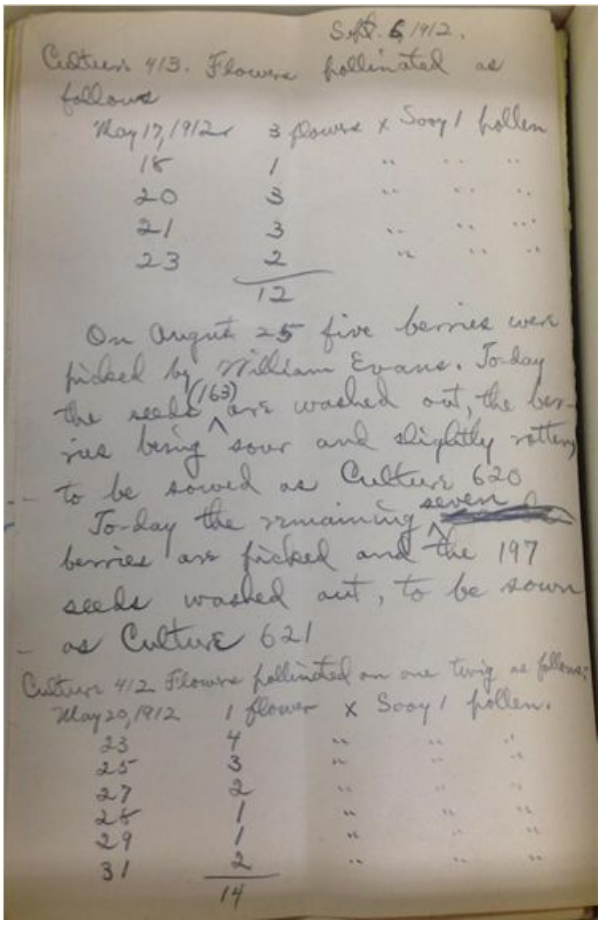
General rules:

- Type what you see: document strikethrough, spelling, grammar errors, etc.
- Transcribe from left to right; from top to down.
- Keep it simple: do not worry about formatting.

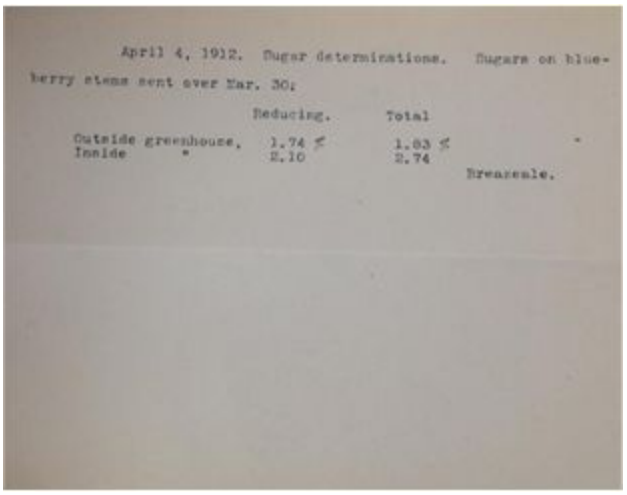
Original text	**Transcription guide**
Letterhead	Use [[letterhead]] BEFORE the letterhead texts and [[/letterhead]] AFTER the letterhead text.
Blank page	Use [[blank page]] to indicate.
Page number	Use [[page]] BEFORE the number and [[/page]] AFTER the number.
Column data or tables	Use pipe symbol (|) to indicate the number of columns. Use three hyphens (---) to indicate the value of the field. For example: | name of variety | Size of berry | Bloom | Origin | Remarks and reference | | Rubel | 17, wild | --- | new jersey | Miss White letter, Sept. 30, 1912 |
Crossed-out words	Type [[strikethrough]] BEFORE and [[/strikethrough]] AFTER the word(s).
Added word(s) with ^	Include the word(s) directly in the transcription.

Words that split in two line with a hyphen	Type the full word.
Ditto marks (“)	Type out the word(s) that ditto mark represents.
Separation line	Use six hyphens (------) to indicate.
Illegible word	Use [[?]] to indicate.
Image	Use [[image]] to indicate. Type out the caption, if applicable.
Margin and footnotes	Transcribe and indicate they are additional notes, using [[margin]] & [[/margin]], [[footnote]] & [[/footnote]] For example, adding [[margin]] BEFORE the margin note, and adding [[/margin]] AFTER the margin note.
Word formatting (bold, italic, underlined)	Do not indicate

Example:



Sept. 6, 1912.
Culture 413. Flower pollinated as follows
May 17, 1912	3 flowers x Sooy 1 pollen
18	1 flower x Sooy 1 pollen
20	3 flowers x Sooy 1 pollen
21	3 flowers x Sooy 1 pollen
23	2 flowers x Sooy 1 pollen
---	12
On August 25 five berries were picked by William Evans. To-day the seeds (163) are washed out, the berries being sour and slightly rotten, to be sowed as Culture 620 To-day the remaining [[strikethrough]] seeds [[/strikethrough]] seven berries are picked and the 197 seeds washed out, to be sown as Culture 621.
Culture 412 flowers pollinated on one twig as follows:
May 20, 1912	1 flower x Sooy 1 pollen
23	4 flowers x Sooy 1 pollen
25	3 flowers x Sooy 1 pollen
27	2 flowers x Sooy 1 pollen
28	1 flower x Sooy 1 pollen
29	1 flower x Sooy 1 pollen
31	2 flower x Sooy 1 pollen
---	14



April 4, 1912. Sugar determination. Sugars on blueberry stems sent over Mar. 30:
| --- | Reducing. | Total |
| Outside greenhouse, | 1.74% | 1.83 % |
| Inside greenhouse, | 2.10 | 2.74 | Breazeable.