Data Rescue: An assessment framework for legacy research collections

Kelly M. Hoffman
Cooper T. Clarke
Hilary Szu Yin Shiue
Phillip Nichols
Miranda Shaw
&
Katrina Fenlon (PI)

September 30, 2020

Report of the Data Rescue Project of the Digital Curation Fellows Program

National Agricultural Library

University of Maryland College of Information Studies

COLLEGE OF INFORMATION STUDIES

USDA National Agricultural Library
U.S. DEPARTMENT OF AGRICULTURE

Acknowledgements

This white paper is the product of research conducted under a Non-Assistance Cooperative Agreement #58-8260-6-003 between the University of Maryland and the United States Department of Agriculture (USDA), Agricultural Research Service (ARS), National Agricultural Library with funding provided by the USDA, ARS, Office of National Programs.

We are deeply grateful to Carol Barford (University of Wisconsin-Madison), Jeffrey J. Volenec (Purdue University), and Sylvie Brouder (Purdue University), who provided feedback on a draft version of the processing guide and assessment framework described below, and to the many domain experts and National Agricultural Library staff who provided mentorship and assistance to the Digital Curation Fellows who conducted this research.

Abstract

Widespread investments in the reproducibility and reuse of scientific data have spurred an increasing recognition of the potential value of data biding in unpublished records and collections of legacy research materials, such as scientists' papers, historical publications, and working files. Recovering usable scientific data from legacy collections constitutes one kind of *data rescue*: the application of selected data curation processes to data at imminent risk of loss. Given the growing interest in data-intensive science and growing movement toward computationally amenable collections in memory institutions, the National Agricultural Library and other curation institutions need systematic approaches to processing legacy collections with the specific goal of retrieving reusable or historically valuable scientific data. This white paper reports on research conducted under the auspices of the Digital Curation Fellows Program, a collaborative research initiative of the United States Department of Agriculture's National Agricultural Library and the University of Maryland College of Information Studies. We offer a framework for assessing collections of scientific records for the purpose of data rescue, developed through research on three case studies of agricultural research collections. This framework aims to guide data rescue initiatives at the National Agricultural Library and other agricultural research centers, and to provide conceptual and practical framing for emerging conversations around data rescue in the agricultural research community and across disciplines.

1. Data rescue at the National Agricultural Library

Widespread investments in the reproducibility and reuse of scientific data have led to a proliferation of publicly accessible data in the past decade. As the open science movement has gained traction—witnessed by the growth of research repositories, deposit requirements, data curation services, data journals, and data citation systems—it has focused on data generated by active and ongoing research efforts. However, stakeholders across the research landscape, including scientists, funders, and curators, share a growing recognition of the value and reuse potential of data biding in unpublished records and collections of legacy research materials, such as scientists' papers, historical publications, and working files (e.g., Downs & Chen, 2017; Wippich, 2012).

Recovering usable scientific data from legacy collections constitutes a kind of *data rescue*: the application of data curation processes to data at imminent risk of loss. *Data rescue* is a framing of the overarching concept of *data curation* to focus on the urgent or otherwise constrained application of selected curatorial processes to data that are particularly vulnerable to disappearance, corruption, or obsolescence. Data may be vulnerable for any number of reasons, including administrative, political, and organizational shifts; technical obsolescence; or the prohibitive costs of processing. Often used synonymously with *data recovery,* the specific entailments of data rescue depend on the condition of the collection and the context of the institution or project undertaking data rescue. Data rescue efforts may include digitization and other material or technical transformations; manual or automatic transcription or extraction of structured or unstructured data; and backup or migration of data into new storage. Each of these activities depends on foundational processes, including the selection and initial assessment of data for processing, which processes determine the priorities and goals of data rescue. This white paper addresses this foundational phase of data rescue work.

This paper arises from the need—an acute need at NAL and many collecting institutions and scientific research centers—for systematic approaches to processing legacy collections specifically to retrieve reusable or historically valuable scientific data from them. Like many collecting institutions and data centers, NAL is frequently confronted with the dual opportunity and challenge of donated collections: often extensive collections of scientific records, publications, and personal papers, donated in the wake of closures of scientific research centers and labs or the retirement of individual scientists. In addition, the NAL maintains accumulated historical collections of materials largely in physical (rather than digital) formats, which contain data and information pertinent to ongoing scientific research, or data documenting historically significant scientific advancements during the history of the USDA.

This white paper offers preliminary guidance on assessing the benefits and challenges of processing collections of scientific records for the purpose of data rescue. The two central objectives of this report are to:

- Inform data rescue initiatives at the National Agricultural Library and parallel work in other agricultural research centers;
- Establish conceptual and practical foundations for ongoing, field-wide conversations and research within the agricultural research community about data rescue.

This white paper is the product of research conducted by fellows in the Digital Curation Fellowship Program, a collaboration between the National Agricultural Library (NAL) and the

University of Maryland College of Information Studies (UMD iSchool). With support from the U.S. Department of Agriculture (USDA) Agricultural Research Service (ARS) Office of National Programs, and under the mentorship of NAL staff and faculty at the UMD iSchool, Data Rescue Fellows conducted three case studies of historical collections of scientists' papers and data, held by Special Collections at the NAL:

- *Frederick Vernon Coville Blueberry Records* (1907-1938): This collection of hand-written research notes and other documents represents the USDA blueberry records of Frederick Vernon Coville, documenting the earliest crosses of commercial blueberries.
- *Wilbur Olin Atwater Papers* (1891-1906): A collection of nutrition datasheets stemming from Atwater's research in the chemical composition of foods, dietary studies, and the respiration calorimeter.
- *The Rufus Chaney collection* (1989-2014): Donated to NAL in 2019 by retired USDA agronomist Rufus Chaney, this is a born-digital collection of Chaney's impactful soil science research, which includes raw data sets, related publications, and analysis files.

The goal of studying these diverse cases—two historical collections of paper materials, and one recent collection of born-digital materials—is to identify strategies for efficiently assessing potentially data-rich collections for data-rescue processing. Prior publications of this research include a complete report on the data rescue case studies (Clarke & Shiue, 2020b) and a processing guide for obtaining preservation-ready data from scientific legacy collections (Clarke & Shiue, 2020a). The data being 'rescued' in the course of case studies is intended for inclusion in the USDA's ARS open access data repository, Ag Data Commons.[1]

The assessment framework offered by this white paper builds on the case studies of agricultural research collections (Clarke & Shiue, 2020b) and the complete preservation-oriented processing guide (Clarke & Shiue, 2020a; discussed in section 4, below). This assessment framework defines a set of 18 factors that determine the value and difficulty of conducting data rescue for a legacy collection of research materials. The intended users of this framework include:

- *Data rescue initiatives in agriculture and beyond*: Having developed from case studies of agricultural research collections, the framework is intended for application in the agricultural domain. However, the assessment factors are not specific to agriculture, and may usefully apply to data rescue efforts in other fields.
- *Data rescue initiatives within and beyond curation institutions*: Because this framework is intended for use by NAL and other curation institutions that provide public access to data over the long term, several factors are predicated on the potential for open-ended data reuse. However, we also hope the framework and processing guide will benefit targeted data-rescue efforts conducted in the course of scientific research projects and led by domain experts rather than curatorial professionals.
- *Data producers*: Research centers, labs, and individual scientists whose collected research materials eventually become the legacy collections considered by this white paper may benefit from considering, from a long-term and open-ended reuse perspective, what it means for a data collection to be complete, fit-for-purpose, and reusable. Of course, the landscape of data sharing and reuse is shifting, along with the roles of data producers and curators in stewarding the reuse potential of data

[1] https://data.nal.usda.gov/

collections; we expect that data rescue for future legacy collections (in other words, for the collections that result from current and ongoing research) will confront very different challenges and opportunities.

As data rescue initiatives in agriculture continue to emerge and evolve, this framework is intended to support ongoing conversations and research within the agricultural research community about data rescue. We hope the framework will be refined and expanded with the addition of community feedback and increased data rescue research over time.

2. Background: Agriculture and data rescue

Agriculture is characterized by distinctive research, communication, and data-sharing practices, which carry significant implications for data rescue initiatives. Research has found that agricultural researchers confront major barriers to accessing and reusing research data, including a lack of discovery tools for data; state-level budget constraints that reduce centralized data collection; and the prohibitive costs and proprietary nature of privatized data (Cooper et al., 2017). Nonetheless, agricultural researchers do rely on data produced by others, and they have been shown to generate large personal research collections over the course of their careers. These, like all personal research collections, tend to be organized idiosyncratically and maintained inconsistently; and upon the closure of a lab or the retirement of a scientist, they are either disposed of, forgotten, or donated to a collecting institution.

Across disciplines, data rescue initiatives—particularly initiatives that target legacy collections—face numerous common obstacles, including the sheer volume of potentially reusable data, the difficulty of assessing its value, the variability of data quality and documentation, and limited resources for retrospective data rescue given the priority placed on active data curation for ongoing research. Beyond these common obstacles, agriculture faces certain distinctive challenges:

Agricultural research encompasses work across a very broad disciplinary spectrum, and is characterized by a high degree of collaboration, including cross-institutional collaboration and partnerships with industry, non-governmental organizations, and government. Agricultural research includes work across the sciences and social sciences, on topics ranging from agronomy to nutrition to animal science, from natural resource management to rural sociology to horticulture (Cooper et al., 2017). For this reason, scientific research and data collections can be large, complex, heterogeneous, and distributed across many institutions. In addition, in some areas essential data are increasingly collected by private entities and are inaccessible for sharing or reuse. The complexity and distribution of projects could compromise the completeness of any collection originating from a single researcher or institution, as is often the case in legacy collection donations.

Data are highly diverse, both within and among projects. Agriculture researchers rely on diverse data types and formats, including quantitative data, such as physiological measurements or survey data; qualitative data, such as field notes; genetic sequencing data; lab notebooks and other metadata; and visual data, such as microscopy and photographs (Cooper et al., 2017). One project may include data of various types and formats; and the documentation and other materials that serve to contextualize the data are equally heterogeneous. While there is never a one-size-fits-all approach to data curation, it may prove difficult to effectively systematize processes for agricultural data rescue to any degree.

Agricultural research includes work done by extension services and programs oriented toward education and the application of research to agricultural practices. This widens the scope of communication and publishing practices in agriculture. Many of the outcomes of agricultural research are shared or published as grey literature, in the form of reports, blog and social media posts, educational materials, videos, etc. The field's reliance on grey literature may pose challenges for the completeness of legacy collections, as data and important documentary materials may be difficult to locate, access, verify, or reconcile outside of conventional systems of publication and preservation.

Agricultural research can carry particular replication concerns. Field work and observational research tend to produce data that are difficult or impossible to recreate or replicate, which lends them high potential value and priority in data rescue efforts bent toward reuse in new research contexts. However, work in certain domains has historically been shaped by the necessity of confirmatory research (e.g., research done to confirm experimental outcomes in different environmental contexts). Because confirmatory research, while necessary to scientific progress and agricultural practice, is not always accorded the same value as research resulting in fundamentally novel scientific outcomes, data producers may not save or thoroughly document confirmatory data for sharing or reuse, which could compromise the completeness of collections subject to data rescue.

2.1. Project context

Since its establishment NAL has been charged with preserving and providing access to research data. Indeed, the 1862 founding legislation for the USDA obliged the department's commissioner to "acquire and preserve in his [sic] Department all information concerning agriculture which he can obtain by means of books and correspondence, and by practical and scientific experiments, (*accurate records of which experiments shall be kept in his office*,)..."[2] (cited in Punzalan et al., 2016; emphasis added). Of course, the burgeoning mass of research collections that fall within NAL's remit long ago exceeded the capacity of anyone's office. Ongoing donations and collection development over the ensuing century and a half of USDA history and the acceleration of scientific production have inevitably produced an accumulation of legacy collections in various stages of active processing. This research on data rescue is motivated by the accumulation of agricultural legacy collections, the increase in data-intensive research across all disciplines, and the attendant movement toward computationally amenable collections in memory institutions (e.g., Padilla et al., 2019).

The NAL/iSchool Data Rescue Project is one strand of the Digital Curation Fellows Program, a multifaceted collaboration between the University of Maryland and the USDA's National Agricultural Library, which began with a cooperative agreement in 2014 to conduct research on curation at NAL and in the broader agricultural community. As part of this agreement the iSchool has placed student fellows from all of its undergraduate and graduate academic programs in positions across NAL divisions, doing research on topics ranging from web archiving to user experience design, from data science to creating historical digital collections. Prior research under the umbrella of the Digital Curation Fellows Program laid the groundwork for this study, particularly reports on research data curation and preservation infrastructures at NAL (Punzalan et al., 2016; Daniels, 2018). A growing collection of Digital Curation Fellows project outcomes can be found in the University of Maryland institutional repository.[3]

[2] "An Act to Establish a Department of Agriculture" (https://www.nal.usda.gov/act-establish-department-agriculture)

[3] https://drum.lib.umd.edu/handle/1903/26345

3. Review of data rescue literature

The following review of prior work on data rescue, stemming from the literatures of research and practice in data curation, library and information science, archives, and several scientific domains, is intended to provide an overview of the concept and major challenges, particularly for practitioners and scientific experts without professional expertise in data curation.

3.1. Data rescue

The phrase *data rescue* refers to systematically converting legacy data from an at-risk (often physical) format or medium to one that is more sustainable (Brunet & Jones, 2011). More broadly defined, data rescue "refers to efforts that enable the sustained use of data that otherwise might go unused" (Downs & Chen, 2017). Data rescue may entail various curation processes, including *digitization* (the conversion of materials from physical to digital format, e.g., scanning), *data recovery* (the retrieval of deleted or damaged digital data, e.g., as from a corrupted hard drive), metadata creation, digital preservation activities, and other processes to ensure the continued management, accessibility, and usability of data. Although physical formats are a major focus of data rescue, data rescue projects increasingly target aging digital data that are "remastered" (Wyborn et al., 2015) into more sustainable and robust digital formats.

The term data rescue gained prominence during distributed efforts to salvage data related to climate change, as administrative turnover led to the removal of some federal environmental websites containing important data after the 2016 presidential election. The crowd-sourced "Data Refuge" initiative to preserve and curate federal climate data (Janz, 2018)—while not the first large-scale data rescue initiative—served to highlight the vulnerability and value of data residing in federal collections and publication. Due to the inherently longitudinal nature of climate research and the relatively high value of even the oldest climate records, climatologists have led the way in data rescue efforts other than U.S. federal data as well. If lost, pre-digital climate data can never be recreated, leading to worldwide efforts to rescue as much climate data as possible (Park et al., 2018; Persaud et al., 2019; Png et al., 2019). However, climatologists are far from having a monopoly on data rescue. Fields as diverse as astronomy, geology, and pharmacology have recognized the benefits of data rescue.

The NAL's data rescue initiative is focused on scientific data, but encompasses data that are structured and unstructured, qualitative or quantitative, digital or analog, etc. Our emphasis is not on processing complex objects or artifacts (though we acknowledge that they, too, may serve as data or primary sources for ongoing research.) The following sidebar on rescuing institutional history at NAL describes a separate but closely related effort to devise systematic and efficient approaches to processing archival documents and files at NAL, in order to preserve institutional history (Shaw & Nicholas, 2020). Such a process requires a different approach and set of requirements than rescuing data for future analytic use; but the guiding principles, including the archival principle of *More Process Less Product* (discussed below), ensure relevance across data rescue and rapid appraisal efforts.

Sidebar: Rescuing Institutional History at the National Agricultural Library

In conjunction with the Data Rescue Project, additional UMD iSchool Digital Curation Fellows worked with a large collection of personal files generated by a senior administrator at the National Agricultural Library (NAL). The fellows were tasked with studying expedited appraisal and to supplement an institutional history collection with the employees' work files (Shaw & Nicholas, 2020). While not focused on scientific data, this project overlaps with data rescue because of its concentration on rapid appraisal and 'rescuing' potentially valuable materials at risk of loss. The necessity of the project emerged with the pending retirement of Susan McCarthy, the Associate Director for the NAL's Knowledge Services Division, who collected analog and digital materials over a thirty-year career. The NAL History Collection, intended to document the institutional history of the Library, had not been updated since 1994 and McCarthy anticipated the accession of her materials could supplement the institutional history of the NAL. In order to process such a large collection and devise procedures for efficient processing, Shaw and Nicholas worked with NAL Special Collections staff to develop a collections development policy. The paper argues that institutions can supplement their historical collections with materials gathered from employees' personal files in consultation with clear collections development policies. Shaw & Nicholas combined the newly created policy with the archival practice of "More Product, Less Process" (MPLP) to rapidly appraise and process the collection in search of materials with significant research value that document the work of the NAL as a whole. The application of MPLP to scientific data, discussed below, informed the assessment procedures described in this white paper.

Reference:
Shaw, M. & Nicholas, P. (2020). Maintaining Institutional Historical Collections through Rapid Appraisal of Employee Files. Digital Repository at the University of Maryland. http://hdl.handle.net/1903/26474 ; https://doi.org/10.13016/szlo-8e08

3.1.1. Why conduct data rescue?

There are many reasons for researchers and institutions to engage in data rescue, beyond preserving data for preservations' sake. The most important is that data rescue may enable better science, where it contributes to the reproducibility and reuse of data. When data accumulated over time is organized, complete, normalized, and documented with sufficient metadata, it becomes feasible to conduct longitudinal analysis (Rountree et al., 2002). Beyond longitudinal data, data from multiple sources, when brought together, can allow for more expansive research to broach "grand challenges" that rely on the confluence of data from various subdisciplines (Cragin et al., 2010; Oden et al., 2011). For instance, the eTOX project brought together clinical trial data from multiple pharmaceutical companies and other organizations to enable the creation of predictive models and more efficient pursuit of new medical drugs (Sanz et al., 2017; see sidebar). Additionally, legacy data enables long-term analysis that would not otherwise be possible (Hawkins, 2013), and although many researchers are content to use only as much historical data as is digitized, Griffin (2015) points out that this practice imposes an arbitrary limit on analysis for no scientific reason.

Sidebar: The eTOX Project

The eTOX project in the pharmaceutical field aimed to improve the situation of lacking experimental data by utilizing rich legacy preclinical drug safety data stored in paper and PDF formats. The project was conducted between 2010 and 2016 with a consortium formed by pharmaceutical companies, academic institutions, enterprises, and was sponsored by the European Innovative Medicines Initiative. The project created eTOXsys, which includes a database with a graphical user interface, and predictive models.

The database in eTOXsys can be used to search if a particular pharmacological target was pursued before. In some cases, the database provides sufficient information and the drug research can continue without animal studies. The data can also be used to “ analyse the correlation between the presence of chemical substructures and the occurrence of specific toxicities” (Sanz et al., 2017, p.812). It is useful for both early drug safety assessment and later assessment, such as “assessment of impurities in drug products” (p.811). Integration of eTOX with other data creates another type of useful object. The predictive models of eTOXsys were developed using “eTOX database and other data resources, such as RepDose, ChEMBL and DrugBank” (p.812). Both the database and predictive models are valuable for drug safety assessment and development. Currently on eTOXsys, the database is freely available to the public, while predictive models are sold separately.

These results would not be possible without the consortium, but the collaboration was one of the challenges faced by the eTOX project. For pharmaceutical companies to be willing to share proprietary preclinical data, multiple solutions were used in the process, including signing legal agreement, using "honest broker” concept, and data-sharing concepts such as FAIR (Findability, Accessibility, Interoperability and Reusability) (Sanz et al., 2017, p.812). After the consortium was formed, the project faced technical challenges to make legacy data available for search and analysis, which includes “lack of standardization” and that extracted raw data requires further transformation for use in modelling (p.812). Hackathons were organized to “define rules for summarizing related toxicological findings and identified underlying relationships between chemical structures and organ toxicities” (p.812). New tools were also created for the consortium to cooperate together (Ontobrowser), and to develop and sustain models (eTOXlab and ADAN).

Reference:

Sanz, F., Pognan, F., Steger-Hartmann, T., Díaz, C., eTOX, Cases, M., Pastor, M., Marc, P., Wichard, J., Briggs, K., Watson, D. K., Kleinöder, T., Yang, C., Amberg, A., Beaumont, M., Brookes, A. J., Brunak, S., Cronin, M., Ecker, G. F., Escher, S., … Zamora, I. (2017). Legacy data sharing to improve drug safety assessment: the eTOX project. *Nature reviews. Drug discovery*, *16*(12), 811–812. https://doi.org/10.1038/nrd.2017.177

For some data sets, especially in disciplines that do not experience rapid change—such as the study of bedrock, or large-scale crop experiments—conducting data rescue, though time-consuming, may still be more cost-effective than recreating the original study (Fallas et al., 2015). Other data rescue efforts involve enhancing the original data, or even transforming into a new form of information. This kind of “data upcycling”, as Vearncombe et al. (2017) put it, adds value to the data and/or uses it in a way that was not foreseen when it was collected. For instance, Piazza et al. (2019) used existing 2D videos of the ocean floor to construct 3D models.

3.1.2. Curation

While the term may have its origins outside of the professional field of data curation, the concept of *data rescue* essential reframes data curation to focus on the urgent or otherwise constrained application of selected curatorial processes to data that are particularly vulnerable to disappearance, corruption, or obsolescence. Data curation can be conceptualized in a lifecycle model, like the one developed by the Digital Curation Centre (DCC) in Figure 1, which focuses on the steps necessary to make sure the data of value produced by a research project or study is preserved and accessible (Higgins, 2008).

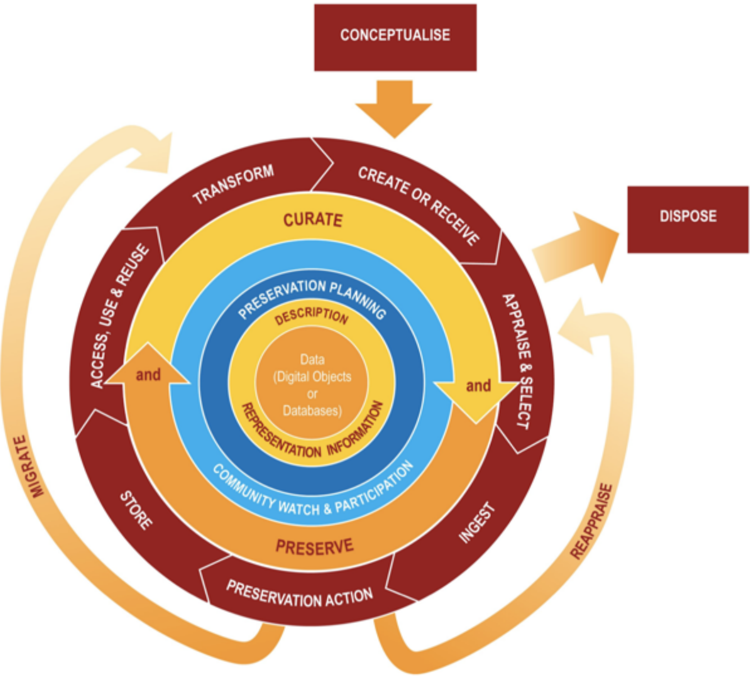

Figure 1, the DCC Curation Lifecycle (Higgins, 2008)

Data rescue may be necessary at any point in the data curation lifecycle and may entail multiple lifecycle activities or processes (Figure 1). While data rescue projects may not address the full lifecycle of data (particularly if they are focused on a specific research objective, and conducted by domain experts rather than curators), partnership between data producers, domain experts, and data curators can lead to more effective data rescue initiatives (Palmer et al., 2011; Pryor, 2012). For this reason, a data rescue project may involve librarians, archivists, domain experts and data scientists—a balance of roles, with curators and data scientists advising on the structure and organization of data and information, while subject matter experts provide insight into the past and future applications of the data. The combination of expertise will make it much more likely that rescued data will be put to future use.

Sidebar: Collaboration in Data Rescue Projects

The McGill University project Data Rescue: Archives and Weather (DRAW) transforms paper weather records into a database. It is an interdisciplinary effort with experts from "archival practices, information studies, data management, public participation, historical climatology, and software design" (Slonosky et al., 2019, p.60). Archivists understand archival principles in organization of records; climate scientists aid in understanding historical weather data; creating the database would require expertise from data management. As DRAW developed an online crowdsourcing transcription tool, software design, programming and public participation are also indispensable for their data rescue project.

McGovern (2017) sheds light on the "cumulative strengths" (p.25) of cross-domain expertise in discussing data refuge efforts, which involves data and web archiving. She identifies seven domains that usefully collaborate on data rescue: Libraries, Archives, Records Management, Digital Preservation, Museums, Software Development, and Data Science. Each of these has specific areas of expertise. For instance, libraries can provide discovery services; archivists are experts in provenance; records management team is familiar with retention schedules, etc. Three examples of collaboration include: (1) ensuring persistent access, (2) completing a gap analysis of needs, and (3) monitoring data and associated information (p.25).

Reference:

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

Slonosky, V., Sieber, R., Burr, G., Podolsky, L., Smith, R., Bartlett, M., Park, E., Cullen, J., & 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

3.1.3. Why are data at risk?

Although the proximate cause of risk to data is often due to environmental factors and/or the passage of time (e.g. deteriorating media or obsolete formats) the distal cause is typically a failure in the curation lifecycle, whether from lack of resources or lack of valuing the data. Lack of contextual information that enables interpretation of the data can also cause the data to be effectively lost. Physical storage carriers deteriorate, and hardware and software become obsolete. Along with other factors like difficult-to-read handwriting or the lack of required specialized equipment, such as equipment that can scan large maps, physical obstacles can stop a data rescue project in its tracks (Fallas et al., 2015; Persaud, 2019; Williams et al., 2019).

There are also contextual challenges. Although some fields, like astronomy, tend to organize data in standardized ways, others are far more heterogeneous (Specht et al., 2018). This is particularly true in diverse fields such as ecology (Specht et al., 2018), where long-tail data from "small science" projects may vary widely from one dataset to another. Not only does this make it difficult to standardize a data rescue workflow for a single field, or even a single data rescue project, metadata from one dataset to another may be challenging to reconcile and homogenize. Longitudinal data, particularly data that was recorded over years or decades, may pose additional challenges to homogenization. The format or structure of the records may have been changed over time; it is even possible that numbers were rounded differently at different times (Png et al., 2019). Even in cases where standardized taxonomies were being used, such as species names, changes to those taxonomies may have occurred while the data was being recorded or in the time since it was recorded (Specht et al., 2018).

The challenges of data rescue mean that much time and effort can be saved if data is properly documented and managed proactively, from the outset of a research lifecycle, with data management planned from the beginning (Yu, 2020; Johnston et al., 2019; Kaufman, 2018; Specht et al., 2018). As Griffin (2015) points out, a great deal of legacy data has survived "more by circumstance than by design" — it has frequently not been stored or documented with an eye towards future use by new communities. Data are sometimes lost because their enduring value simply is not perceived by data producers, or because the time and effort needed to rescue data may not be considered worthwhile since data rescue tends to be resource-intensive (Fallas et al., 2015; Specht et al., 2018). The incentives for creating new data from research projects is often greater than rescuing existing data (Griffin, 2015), and data rescue, curation, and planning for future efforts often compete for funding (Gallaher & Diggs, 2017).

Sidebar: The NUS Republican China Weather Database

Png et al.'s (2019) project gathered 463,530 instrumental climate observations from 319 stations from 1912 to 1951 and consolidated them into a data set. Weather records are useful for a wide variety of research interests, including studying climate variability, events of extreme weather, and analysis of changes. Nonetheless, historical weather data of the first half of the 20th century in China is limited (p.2). It is dispersed in multiple library collections (both physical and digital collections) in China, Taiwan, Japan, and the U.S. The project utilized 64 sources from 36 libraries and online collections (Png et al., 2019, p.4) to create the NUS Republican China Weather Database.

Challenges of consolidating data mostly originated from "changes in the observation procedure, personnel, instruments, or monitoring stations" (Png et al., 2019, p.5), but could also come from "errors in typesetting and digitization" (p.12). Png et al. (2019) detail how they detected errors, such as checking whether daily maximum temperature is higher than the minimum and comparing observation with the previous day. The penalized maximal F test and packaged code RHtestV4 was used "to detect change points, or undocumented mean shifts that are *not* associated with sudden change in the linear trend of a time series" (p.8). They also noticed excessive zeros and fives due to re-rounding of temperature from Fahrenheit to Celsius and utilized the Hidden Markov Model to resolve this issue (p.6). The new data set was also validated by comparing it with three existing data sets.

The process of Png et al.'s project demonstrates potential challenges for data rescue projects, and how they can be resolved by using existing methods, such as the Hidden Markov Model and RHtestV4 codes.

Reference:
Png, Ivan P. L. and Chen, Yeh-Ning and Chu, Junhong and Feng, Yikang and Lin, Elaine Kuan-hui and Tseng, Wan-ling, Temperature, precipitation and sunshine across China, 1912-

51: A new daily instrumental dataset (September 16, 2019). Available at SSRN: https://ssrn.com/abstract=3454857 or http://dx.doi.org/10.2139/ssrn.3454857

3.2. Minimal Standards Data Rescue

Incentivizing data rescue projects is one approach to the problem of data rescue (Griffin, 2015; Hsu et al., 2015). Another approach is to reduce the time and effort required to complete a particular data rescue project. For guidance on this approach we can turn to the field of archival science, particularly the *More Product, Less Process* (MPLP) concept.

3.2.1. MPLP

MPLP refers to principles outlined in a seminal 2005 work, "More Product, Less Process: Revamping Traditional Archival Processing," in which Mark Greene and Dennis Meissner argue that it is better to achieve "minimal or partial processing" (p. 239) of archival collections than to allow them sit inaccessible and unused in a backlog waiting for comprehensive preservation, arrangement, and description. As Greene says, "we must accept that 'good enough' is better than 'one of these days'" (2010, p. 178).

MPLP was received enthusiastically by many archivists. Inspired by Greene and Meissner, the Yale University Library's Archives condensed the accessioning and processing phases into one—arranging and describing materials and creating simple finding aids as the materials were accessioned (Weideman, 2006). As a new archivist at the University of Montana, Donna E. McCrea used MPLP principles to confront a backlog of 3,000 linear feet at a rate of 2 hours per linear foot, rather than the previously estimated 8 hours per linear foot it would have taken using traditional processes (McCrea, 2006). At Humboldt State University, the MPLP guidelines were used as "a conceptual model" to help "navigate the tradeoffs between quantity and quality" while processing special collections (Harling, 2014).

However, MPLP was not without detractors. Ness (2010) argued that most archives in the United States were already practicing MPLP principles, without giving them a name. McCann (2013) criticized Greene and Meissner for implying that preservation is a competing goal to access. Cox (2010) demonstrated tempered enthusiasm for MPLP, describing how his archives strive for "maximal processing," which aims to "facilitate access to the greatest degree possible" (p. 145) which may mean planning to return to a minimally processed collection at a later time to process it more fully. In the 2019 paper "Toward Slow Archives," Christen and Anderson pointedly critique MPLP by stating that "the *process* is as essential as the *product*" (p. 111), and by "slowing down" archivists would be "focusing differently, listening carefully, and acting ethically" (p. 90).

Some of these criticisms may boil down to a fundamental disagreement about the purpose of archives and archival work. Meissner and Green (2010) have stated that "researcher use is the purpose of all archival effort" (p. 195), hence their emphasis on getting collections into the hands of users as quickly as possible. Others argue that "the history of collecting is the history of colonialism" (Christen & Anderson, 2019, p. 99) and that traditional archival practices were developed as tools to enforce "settler ambitions, practices, and assertions (Christen & Anderson, 2019, p. 91). They further argue that "the *process* is as essential as the *product*" (emphasis original) and that a "slow" process "creates a necessary space for emphasizing how knowledge is produced, circulated, contextualized, and exchanged through a series of relationships" (p. 90).

Meissner and Greene's response to criticisms has generally been to claim a misunderstanding of what MPLP really is. They and other MPLP defenders argue that MPLP and minimal processing are not interchangeable; rather, that MPLP is a "conceptual model" (Harling, 2014, p. 497) that is intended to guide "resource management" (Meissner & Greene, 2010). Regardless of what the ultimate purpose of archiving is, approaching MPLP as a reminder to consider goals and objectives in relation to available resources and potential impact can be a useful perspective.

3.2.2. MPLP and Data

The field of data curation is beginning to turn its attention to MPLP as well. Curating data is somewhat different than curating traditional archival materials because context and metadata are much more necessary to understand it -- for instance, how the data was collected, how it is structured, or what particular terms or categories mean (Lafferty-Hess & Christian, 2017).

Applying the MPLP principles to data, Lafferty-Hess and Christian (2017) propose a "minimal data curation" pipeline that involves *arrangement* (ensuring completeness), *description* (basic metadata and identifiers), and *preservation* (converting files into "non-proprietary, software-agnostic" formats), with the fundamental end goal of "ensur[ing] that enough information is present for users to understand and interpret the data as a whole" (p. 10).

Recognizing that all data need not be processed the same way or to the same extent, Emory University developed a tiered system of processing for born-digital data (Waugh et al., 2016). Evaluating each dataset on quality of content, access restrictions (such as copyright), and expected level of use, they determine what tier of processing is appropriate, from Tier 1 (simple collections that are likely candidates for high automation) to Tier 3 (high complexity and manual effort).

3.2.3. Minimal Appraisal

Appraisal has been described as "the single most important function performed by an archivist" (Craig, 1992, p. 176), yet relatively little work has been done exploring MPLP specifically as it relates to archival appraisal. In 2010, Greene argued that MPLP could be used "not just for processing," describing how it might apply to appraisal, preservation, and reference, alongside processing.

What MPLP-related work that does exist regarding appraisal focuses on the initial step of determining whether a collection is in general an appropriate acquisition for an archive or repository. Both Greene (2010, 2011) and Ness (2010) suggest that overly liberal appraisal and acquisition decisions (often made without the guidance of collection development policies) have contributed to the archival backlog.

However, appraisal happens at multiple points during the acquisition and processing of archival materials. Not only is there some level of appraisal before materials are selected and acquired, appraisal at different levels (such as the series or item level) can occur throughout processing (Cross, 2011; Searcy, 2017). Greene (2010) argues for more appraisal occurring "at the site of origin" and "on the loading dock" in order to cut down on the total amount of materials that need to be processed.

In order to fully appraise scientific data for reuse potential — such as when determining the amount of resources and processing that should be allocated to a data rescue project — the appraiser must understand not just the primary user community (often the originating research sub-discipline) and their potential uses of the data, but also the likelihood that other potential user communities exist and whether the data in its current form suits the purposes of those communities (Palmer et al., 2011), or if it would need to undergo additional transformation first. This is another stage of data curation benefitted by collaboration between data specialists (such as archivists, data scientists, or librarians) and subject matter experts (with at least some knowledge of multiple disciplines). Experts with knowledge of multiple disciplines are better able to identify potential future uses of the data outside of the immediate designated community; data specialists can determine whether transformation of the data to become fit for the purposes of another community is possible and appropriate.

4. Collections assessment and processing

The following assessment framework elaborates on the assessment phase of a more complete, preservation-oriented processing guide, provided in Clarke and Shiue (2020a) and discussed in more detail in section 4, below.

The framework is intended for use in the initial, exploratory phase of data rescue, to determine priorities for data rescue, explain the potential value of data rescue processes, anticipate potential obstacles to processing, and begin to assess the labor and resources required for different levels of processing. This framework defines a set of 18 factors that together determine the costs and value of processing a collection to recover research data for reuse. The primary audience for this guide is data curation professional selecting and assessing collections for data rescue oriented toward long-term, open-ended preservation and reuse of resulting data. However, the guide may also benefit researchers with specific, near-term research use of the data and no stakes in the longer data lifecycle.

Section 4.2. demonstrates the application of the framework to three agricultural research collections—our three case studies, detailed in Section 1, above. The factors aim to reveal a rich set of potential "pros" and "cons" to data rescue. Because ultimate data processing decisions are necessarily contextualized by the resources and priorities of the institutions undertaking data rescue, this framework does not offer any prescriptive guidance on how to weigh these factors against one another, or formal approach to applying them to curation decisions. The factors below are oriented toward a preliminary assessment in the context of the NAL data rescue initiative. This framework is not comprehensive or exhaustive; there are contextual, collection-, project-, and institution-specific factors confronting any data rescue effort, which we cannot anticipate. It is intended as a starting point for ongoing refinement and expansion in conversation with the agricultural community.

The following assessment factors derive from our experiences evaluating the case study collections (reported in Clarke & Shiue, 2020b) in combination with factors identified in several sources stemming from research and practice in data curation, including:

- Ag Data Commons Data Submission Manual[4]
- Data Curation Profiles Toolkit (Carlson, 2010)
- The Digital Processing Framework (Faulder et al., 2018)
- Curating Research Data: Volume II (Johnston, 2017)
- "Scientific data appraisals: The value driver for preservation efforts" (Faundeen & Oleson, 2007)
- "The analytic potential of scientific data: Understanding re-use value" (Palmer et al., 2011)

[4] https://data.nal.usda.gov/ag-data-commons-data-submission-manual#description-fields

4.1. Assessment factor definitions

The following 18 assessment factors are defined in Table 1, below. For each factor, we provide a set of guiding questions that shed light on the reasoning behind the inclusion of the factor and its implications for data rescue.

- Extent
- Data objects
- User communities
- Stakeholders
- Reuse value
- Reusable objects
- Historical value
- Historical objects
- Completeness
- Sensitivity
- Access and use constraints
- Rarity or uniqueness
- Reproducibility
- Relevant collections
- Associated publications
- Fit for purpose
- Obstacles to recovery
- Priorities

Assessment factor	***Guiding questions***
Extent	How large is the collection (characterized in terms of linear feet, number of boxes, number of digital files, digital file size, etc.)? Within the collection, how much data is present? To what extent is the collection or the data within the collection already *processed*? In other words, how much of the collection is already documented, organized, digitized, curated, or otherwise completed to the degree intended by the data rescue initiative?
Data objects	What kinds of data exist in the collection, and in what forms? What file formats or physical materials are the data in? How are the data related to other materials in the collection? For example, do the data exist as stand-alone files or datasheets, or are the data embedded in other documents, such as published articles or hand-written field notes?
User communities	What groups of potential users should be able to understand and use the data in this collection? If data rescue is being conducted for the purposes of a specific research project, the project team and surrounding research community are likely the primary community for the data. If data rescue is being conducted for open-ended future reuse, user communities should be evaluated based on the originating community of the data, any explicitly indicated audiences for the research, and a meta-analysis of related fields conducted by a curator in consultation with domain experts (Palmer et al., 2011).
Stakeholders	Other than direct users, what groups, institutions, or communities have or could have an ongoing interest in the data? Who has invested in the data or in the research it supports? Who would be affected by use or reuse of the data?
Reuse value	What are the intended, demonstrated, anticipated, or plausible reuse opportunities for the collection? What new uses could the data and associated documentation be put to, or what analyses are planned for the data, once rescued? Note that this question is *not* limited to the data alone, but also encompasses potential reuse of other facets of the collection, including methodological or contextual documentation, tools, or protocols. Note that this factor focuses on novel uses of the data and materials; for a distinct but related factor, see *reproducibility* (below), which addresses the reuse of the data for reproduction or replication of scientific results.

Reusable objects	Are there specific components of the collection that carry reuse opportunities? Are there specific components that are amenable to reuse? Components may be material or abstract; they may correspond to a subset of *data objects* (above).
Historical value	What is the potential historical value of the collection? What important or noteworthy scientific approaches, results, or advances are documented or evidenced by the data? Note that this question is *not* limited to the data alone, but also encompasses other facets of the collection, including methodological or contextual documentation, tools, or protocols.
Historical objects	Are there specific components of the collection that carry historical value? Components may be material or abstract; they may correspond to a subset of *data objects* (above). While historical value and reuse potential are certainly interwoven, this factor distinguishes data as potential evidence for *science* from data as potential evidence for the *history of science*, as such data may have different processing entailments. For example, it may be sufficient for data intended as *historical objects* (and not *reuse* objects) to be digitally accessible and readable, without being transformed into a format amenable to statistical analysis.
Completeness	How complete or incomplete is the collection? In other words, are there gaps in the collection that would limit either reuse or historical value? This may be understood as a facet of *fit for purpose* (below).
Sensitivity	Are there aspects of the collection that may be considered *sensitive* to unintended or undesirable access, use, or interpretations, whether from the standpoint of privacy, ethics, security, or scientific accuracy? For example, is there personally identifiable information in the collection? Do historical data represent results that run counter to more recent advances in science and policy, or use oppressive or offensive metrics or methods? If so, data rescue must consider how the data will be stored, managed, or represented with appropriate access controls or contextual information.
Access and use constraints	What constraints will be placed on access to and use of the data? For example, are there intellectual property constraints or factors in *sensitivity* (above) that affect how the data should be made accessible?
Rarity or uniqueness	Is any part of the collection or data within the collection duplicated elsewhere, or actively stewarded, curated, or maintained by another other group or institution? This question is particularly relevant for digital data. This factor may also be used to address other, distinctive strands of *rarity*: whether the data are fundamentally irreplaceable, or whether aspects of them could be recreated (e.g., through experimental or computational replication). In addition, this factor overlaps with *reproducibility* (below) as an opportunity

	to consider whether the data attest to scientifically novel outcomes, or whether they dispute, confirm, or replicate existing research.
Reproducibility factors	In what ways, if any, are the data within the collection *reproducible*? Based on the contents of the collection (including the completeness of the data and documentation), are the data amenable to the reproduction or replication of scientific results?
Relevant collections	Are there other collections of research materials that are relevant to this collection, and which demonstrate a wider network of interest or investment in the research documented by the collection? (Note that this question considers collections that do not duplicate the data or other materials, as duplicative collections are considered under *rarity or uniqueness*, above.)
Associated publications	Are there identifiable publications associated with the collection, such as scientific journal articles that report, rely on, or cite the data or methods represented by the collection?
Fit for purpose[5]	To what extent are the data ready or suitable for actual or potential uses identified in *reuse value*, *historical value*, and *reproducibility* (above)? How much additional documentation, interpretation, and processing are required to prepare data either for reuse, or adequately to serve as historical evidence? And what level of scientific, technical, or research expertise would additional processing entail? For example, are there data represented by unstructured text and graphical representations that would need to be extracted or translated into structured form for future computational analysis? This factor takes into account other factors including *data objects*, *completeness*, and *access and use constraints.*
Obstacles to recovery	What are the anticipated or observed obstacles to recovering data from the collection? This question builds on *fit for purpose* and other factors (above) to invite data rescuers to inventory potential obstacles. Obstacles to data recovery may result from a very wide range of properties of the collection, including its physical condition; the quality, completeness, and forms of data in the collection; digital file formats; extant documentation; and the approachability or understandability of the collection to unfamiliar or non-expert users or curators.

[5] This factor is adapted from the analysis of *analytic potential* in Palmer et al. (2011).

	Answers to this question should reflect the unique context and objectives of the group, institution, or organization undertaking data rescue.
Priorities	What are the most immediate priorities for data recovery, as opposed to the optimal or long-term objectives of recovery?

Table 1. Assessment factor definitions and guiding questions

4.2. Assessment factors applied to collections

This section exemplifies how assessment factors may be applied to real-world collections in order to lay the groundwork for formal assessment and appraisal, as early stages in processing for data rescue. Table 2 provides a brief, structured summary of each case study collection, using standard descriptive metadata fields (specifically, a very small subset of data fields employed in Ag Data Commons). The goal of this table is simply to provide context for understanding the exemplary assessment of each collection, given in Table x.

Table 3 provides an exemplary assessment of each case study. The assessment is provided in tabular form for ready comparison of how the assessment factors might be wielded differently across collections of different scopes, sizes, and shapes.

For a full analysis of all case studies, and information on how the preliminary assessment led to processing decisions for each collection, see Clarke and Shiue (2020b).

Case summaries

	Case study 1: Coville	Case study 2: Atwater	Case study 3: Chaney
Description	The Frederick Vernon Coville's Blueberry Notes Collection documents Coville's seminal research into blueberry cultivation, through Coville's handwritten and typed notes on blueberry pedigree information, fieldnotes, descriptions of characteristics of blueberry cultivars, and more. It includes administrative files and a container list. The collection was acquired by NAL in 2007 and is held by Special Collections.	The Wilbur Olin Atwater Papers (MS 261) held by the USDA's Special Collections contain 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. 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.	The Rufus Chaney collection was donated by retired USDA agronomist, Rufus Chaney to the NAL in 2019, in hope to preserve and make the data available. It is a born-digital collection, consisting of a variety of formats. The collection is largely organized by crop types. The content includes raw data sets, data subsets, related publications, analytics system files.

Subject	• Coville, Frederick V. (Frederick Vernon), 1867-1937--Manuscripts. • Blueberries--United States. Blueberries--Soils--United States. Blueberries--Diseases and pests--United States. • Blueberries--Varieties--United States. • Blueberries--Prices--United States. • Fruit breeders--United States--History--Manuscripts. • Blueberries--Breeding--United States--History--Manuscripts.	• Food--Composition--Research. • Food--Analysis--Research. • Processed foods--Nutritional aspects--Research. • Dietaries--Research. • Nutrition--Research. • Nutrition surveys--United States--Archival resources. • Nutrition--United States--History--Archival resources.	• Soil chemistry. • Soil science. • Soils--Heavy metal content. • Crop science.
Temporal coverage	1907-1938[6]	1891-1906	1989-2014

Table 2. Case summaries for case studies detailed in Cooper and Shiue (2020b)

Application of assessment factors

	Case study 1: Coville	Case study 2: Atwater	Case study 3: Chaney
Extent	6 linear feet; 24 customized boxes. Part of the collection (2 boxes) is digitized. Not fully processed. No finding aids, but administrative files are available	900 handwritten sheets	262 files
Data objects	• Data present on paper as handwritten or typed notes. • Qualitative: narrative descriptions and field notes • Quantitative: tabular data (no. of cultures, temperature, flower and fruit size, etc.)	• Quantitative: tabular data • Qualitative: legacy research methods	• Quantitative: tabular (amounts of chemicals recorded, ph level, geospatial data) • Qualitative: analytics steps

[6] The notebook for 1938 was kept by George Darrow, Coville's successor, after Coville passed away in 1937.

	• Longitudinal: Observations and pedigree developments		
User communities	Horticultural scholars; genetic scientists; general public	Agriculture scholars; nutrition scientists	Soil scientists; plant scientists; crop scientists; environmental scientists; biosolids scientists
Stakeholders	USDA Agricultural Research Service (ARS), commercial blueberry growers.	USDA ARS	USDA ARS, the Food and Drug Administration (FDA), and the Environmental Protection Agency (EPA)
Reuse value	Certain cultivars still in contemporary cultivation; genetics research using longitudinal pedigrees information; confirmatory research of blueberry cultivation practices	Re-analyze the data used to create the Atwater formula; food composition longitudinal study	Confirmatory research, e.g. analytical steps; longitudinal study in soil science; genetics research using crop cultivars information; interoperate with other crop data (e.g. wheatinitiative.org)
Reusable objects	Detailed pedigree information for both released and unreleased cultivars, including parent cultivar names of well-known cultivars, years of release, and plant characteristics and inheritance	Raw data set (subject to ongoing scientific citations)	Raw data set Analytics system files (.sas)
Historical value	Significant contributions to blueberry domestication: early fertilizers, use of acidic soil and cold treatment for blueberry cultivation	Fundamentally changed USDA approach to nutrition and food composition, formed the basis of the Atwater formula still in use today	During Chaney's 48-year career as an agronomist at USDA-ARS, his research made significant contributions to the study of heavy metals present in soil and their uptake in crops, the application of biosolids to cropland

			(collaboration with FDA and EPA), and phytoextraction of contaminated soil.
Historical objects	The whole collection of Coville's blueberry notes are of historical value as it is a century-old collection.		
Completeness	The completeness of data in the collection is difficult to measure until it is fully digitized. Nonetheless, from initial inventory, the collection exhibits detailed observations in its temporal coverage.	The data sheets can be assumed to be nearly complete when compared with Atwater publications of the data.	The data set of Chaney's collection is quite extensive, although verifying variable names would be necessary before making It available. Potential gap in Chaney's collection exists in the connection between the data set and other digital objects, such as SAS system files.
Sensitivity	N/A	N/A	The data report findings of heavy metal uptake in crops used for human consumption, so could have controversial or alarming implications
Access and use constraints	N/A	N/A	Undetermined
Rarity or uniqueness	This is the original collection; a small subset (2 out of 24 boxes) has been digitized and is available through the Internet Archive (but not transcribed) (Coville, 1907-1908).	This is the original collection; none have been digitized or transcribed.	These are likely the only copies of this data, some results have been published.
Reproducibility factors	The qualitative nature of the field notes may complicate reproducibility.	Quantitative data in analog format that requires	Born-digital data set

		transcription before it can be readily reproduced.	
Relevant collections	Potentially related field notes, bulletins, journal publications, and correspondence attributed to Coville located in Smithsonian Institution Archives; Harvard Botany Libraries; Biodiversity Heritage Library; Biostor; JSTOR; New York Botanical Gardens; Wellcome Library; USDA; Library of Congress; United States Forest Service (aggregated at Internet Archive)	Wilbur Olin Atwater Papers [analog] at Wesleyan University Special Collections and Archives; The Medical Heritage Library; Augustus C. Long Health Sciences Library (Columbia University) (aggregated at Internet Archive)	Mostly publications disseminated through electronic versions of academic journals, e.g. Journal of Environmental Quality, Annual review of plant physiology, Soil Science Society of America Journal, etc. EPA research data: "Bioaccessibility tests accurately estimate bioavailability of lead to quail[7]" (Beyer et al., 2016)
Associated publications	USDA Bulletins: "Experiments in Blueberry Culture" (Coville, 1910); "Taming the Wild Blueberry" (Coville, 1911); "The Agricultural Utilization of Acid Lands by Means of Acid-tolerant Crops" (Coville, 1913); "Directions for Blueberry Culture" (Coville, 1916 & 1921) USDA Yearbook of 1937:	"The chemical composition of American food materials" (Atwater & Bryant, 1906); "Calculating the metabolizable energy of macronutrients: A critical review of Atwater's results" (Sánchez-Peña et al., 2016) "Heats of combustion representative of the carbohydrate mass contained in fruits, vegetables, or cereals" (Martínez-Navarro, 2019)	"Elements in Major Raw Agricultural Crops in the United States. 1. Cadmium and Lead in Lettuce, Peanuts, Potatoes, Soybeans, Sweet Corn, and Wheat" (Wolnik et al., 1983); "Elements in Major Raw Agricultural Crops in the United States. 2. Other Elements in Lettuce, Peanuts, Potatoes, Soybeans, Sweet Corn, and Wheat (Wolnik et al., 1983); "Elements in majoy Raw Agricultural Crops in the United States. 3. Cadmium, Lead, and Eleven Other Elements in Carrots, Field Corn, Onions, Rice, Spinach, and Tomatoes" (Wolnik et al., 1985); "Cadmium, Lead, Zinc, Copper, and Nickel in Agricultural Soils of the United States of America" (Holmgren et al., 1993)

[7] DOI: https://doi.org/10.1002/etc.3399

	“Improving the Wild Blueberry” (Coville, 1937)		
Fit for purpose	The fragility of century-old papers and the absence of finding aids are some of the obstacles for reusing the collection and locating specific data types. Additional processing and transcription are also necessary to migrate data from paper to machine-readable formats to fit modern research practices.	Transcription is necessary to migrate data from paper to machine readable formats to fit modern research practices.	For the raw data set, verification of variable names would be necessary. For other digital objects in the collection and their connection to the data set and publications, further interpretation and documentation may be necessary.
Obstacles to recovery	● Mix of analog and digitized materials ● Fragility of analog materials ● Loose leaf pages vulnerable to loss of original order ● Determining number and completeness of datasets within documents ● Mix of handwritten and typed data ● Mixed data types ● Structured data embedded in unstructured text ● Inconsistent or incomplete metadata and data within files/papers, including missing column headers, empty fields ● Determining processing	● Fragility of analog materials ● Handwritten tabular data Inclusion of handwritten margin notes, handwritten strike-throughs	● Linking data to relevant publications ● Missing context and metadata require expert consultation ● Determining completeness within files Access to outmoded software originally used to create the files

	priorities (based on reuse and historical value of different documents) requires expert consultation ● Unprocessed collection absent documentation and finding aid Linking data to relevant publications		
Priorities	1. Digitization of the other 22 boxes 2. Determine priority of creating machine-readable data 3. Creating transcription according to the priority Make data available on Ag Data Commons in suitable reuse formats	1. Digitization of the data sheets 2. Determine priority for transcription 3. Create transcription style guide 4. Make data available on Ag Data Commons of Digital Collections in suitable reuse formats	1. Appraise data files and identify major data types. 2. Process data files into a unified data set 3. Consult with Chaney for description and metadata information 4. Make data available on Ag Data Commons in suitable reuse formats

Table 3. Assessment factors as manifested in case study collections

4.2. Preservation-ready data rescue guide

The *Data Rescue Processing Guide*, available as a separate document (Clarke & Shiue, 2020a),[8] describes a comprehensive set of processes to produce data that are *preservation-ready,* or adequately curated and documented to support archival preservation and long-term use of the data (Lavoie, 2014; Palmer et al., 2011).

The guide is oriented toward data rescue and archiving in the context of preservation institutions, to support long-term, open-ended data use and reuse. It offers recommendations for different levels of archival processing (baseline, moderate, and intensive) that the NAL and other curation institutions may undertake to align processing work with available resources. The aspects of that guide include:

- An introduction to the Open Archival Information Systems (OAIS) reference model for understanding the roles and responsibilities of a data repository for preserving access to scientific data over time.
- As assessment of the designated community of the National Agricultural Library, or the communities for which data rescue decisions should be made.
- An initial set of appraisal questions for assessing data, which has been expanded for the framework in this white paper.
- An adaptation of the Cornell University Library Digital Processing Framework by (Faulder et al., 2018) to establish tiers of processing and OAIS-informed processing steps to best suit data-rich materials, both analog and digital.

While the guide is oriented toward professional curators and curation institutions (including libraries, special collections and archives, and data repositories), it acknowledges essential roles for data producers, would-be data reusers, and domain experts in the data rescue process. In addition, the guide aims to help domain experts conducting data rescue initiatives without the assistance of a curation institution. Whether scientists and researchers undertaking data rescue for their own purposes plan to implement processes for long-term data preservation, the guide may nonetheless elucidate what long-term data management entails, and what roles curation institutions can play in partnership with domain experts to ensure scientific data remain accessible and useful over time.

In general, successful data rescue initiatives in any domain will depend on collaboration between domain experts (researchers in the sub-disciplines from which the data originated) and data-curation professionals with a meta-disciplinary perspective and expertise in library and information science. The need for partnership and multiple perspectives in all data curation work has long been acknowledged. Data producers and domain experts are uniquely qualified to determine whether the quality, documentation, and forms of data collections are amenable to use in their own domain. But cross-disciplinary research on data sharing has suggested that data producers struggle to anticipate how their data may be used in research in other areas; and data producers are not well positioned to generate adequate descriptive metadata or documentation to support wide-ranging reuse possibilities (Cragin et al., 2010; Baker &Bowker, 2007). Whatever their expertise, users of this guide will need to assess which roles they can assume in the data rescue process, and which would benefit from partnership with curation institutions or scientific domain experts.

[8] http://hdl.handle.net/1903/26473

5. Future work

The assessment framework defined above, and the complete preservation-oriented processing guide detailed in Clarke and Shiue (2020a), provide a foundation for data rescue work at NAL, along with some conceptual and practical framing for emerging conversations around data rescue in the agricultural research community and across disciplines. We hope that the assessment framework and processing guide will be refined and expanded over time, both through:

- Further case studies to apply and evaluate the assessment framework in more institutional and research contexts;
- Ongoing conversations with and feedback from researchers and curation professional across agriculture and related fields, about this and related data rescue efforts.

Future work will also aim to contribute additional guidance on the specific roles of and collaboration between curators and domain experts in data rescue workflows, and to investigate the outcomes impact of data rescue initiatives on scientific reproducibility, data reuse, and public access to science.

References

Ball, A. (2012). *Review of data management lifecycle models*. University of Bath. https://researchportal.bath.ac.uk/en/publications/review-of-data-management-lifecycle-models

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

Carlson, J. (2010). *The Data Curation Profile Toolkit: The Profile Template*. Purdue University Libraries / Distributed Data Curation Center. https://doi.org/10.5703/1288284315653

Christen, K., & Anderson, J. (2019). Toward slow archives. *Archival Science*, *19*(2), 87–116. https://doi.org/10.1007/s10502-019-09307-x

Clarke, C. T., & Shiue, H. S. Y. (2020a). *Data Rescue Processing Guide: A Practical Guide to Processing Preservation-Ready Data from Research Data Collections*. National Agricultural Library and University of Marland College of Information Studies. https://doi.org/10.13016/dif5-arr2

Clarke, C. T., & Shiue, H. S. Y. (2020b). *Final report and recommendations of the data rescue project at the National Agricultural Library*. National Agricultural Library and University of Marland College of Information Studies. https://doi.org/10.13016/kpt7-cqgr

Cox, R. S. (2010). Maximal processing, or, archivist on a pale horse. *Journal of Archival Organization*, *8*(2), 134–148. https://doi.org/10.1080/15332748.2010.526086

Cragin, M. H., Palmer, C. L., Carlson, J. R., & Witt, M. (2010). Data sharing, small science and institutional repositories. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *368*(1926), 4023–4038. https://doi.org/10.1098/rsta.2010.0165

Craig, B. L. (1992). The acts of the appraisers: The context, the plan and the record. *Archivaria*. https://archivaria.ca/index.php/archivaria/article/view/11848

Cross, S. N. (2011). *Appraising archivists: Documentation and the need for accountability in the appraisal process* [Master's Thesis]. Western Washington University.

Daniels, M. (2018). *Digital Workflows at the National Agricultural Library and Implications for Preservation*. National Agricultural Library and University of Marland College of Information Studies. https://doi.org/10.13016/cgxo-pmvc

Downs, R. R., & Chen, R. S. (2017). Curation of scientific data at risk of loss: Data rescue and dissemination. In *Curating Research Data Volume One: Practical Strategies for Your Digital Repository* (pp. 263–277). Association of College and Research Libraries.

Fallas, K. M., MacNaughton, R. B., & Sommers, M. J. (2015). Maximizing the value of historical bedrock field observations: An example from northwest Canada. *GeoResJ*, *6*, 30–43. https://doi.org/10.1016/j.grj.2015.01.004

Faulder, E., Annand, S., DeBauche, S., Gengenbach, M., Irwin, K., Musson, J., Peltzman, S., Tasker, K., Uglean Jackson, L., & Waugh, D. (2018). *Digital Processing Framework* [Report]. https://ecommons.cornell.edu/handle/1813/57659

Faundeen, J. L., & Oleson, L. R. (2007). *Scientific Data Appraisals: The value driver for preservation efforts*. 6. http://www.pv2007.dlr.de/Papers/Faundeen_AppraisalsValue_for_Preservation.pdf

Gallaher, D., & Diggs, S. (2017, April 6). *Data rescue IG meeting*. RDA 9th Plenary Meeting, Barcelona, Spain. https://www.rd-alliance.org/system/files/documents/data%20rescue%20RDA9.pdf

Greene, M. (2010). MPLP: It's not just for processing anymore. *The American Archivist*, *73*(1), 175–203. https://doi.org/10.17723/aarc.73.1.m577353w31675348

Greene, M. A. (2011). Doing less before it's done unto you: Reshaping workflows for efficiency before the wolf is at the door. *RBM: A Journal of Rare Books, Manuscripts, and Cultural Heritage*, *12*(2), 92–103. https://doi.org/10.5860/rbm.12.2.356

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

Griffin, E. R. (2015). When are old data new data? *GeoResJ*, *6*, 92–97. https://doi.org/10.1016/j.grj.2015.02.004

Harling, A. (2014). MPLP as intentional, not necessarily minimal, processing: The Rudolf W. Becking Collection at Humboldt State University. *The American Archivist*, *77*(2), 489–498. https://doi.org/10.17723/aarc.77.2.563004228307n2m3

Hawkins, S. J., Firth, L. B., McHugh, M., Poloczanska, E. S., Herbert, R. J. H., Burrows, M. T., Kendall, M. A., Moore, P. J., Thompson, R. C., Jenkins, S. R., Sims, D. W., Genner, M. J., & Mieszkowska, N. (2013). Data rescue and re-use: Recycling old information to address new policy concerns. *Marine Policy*, *42*, 91–98. https://doi.org/10.1016/j.marpol.2013.02.001

Higgins, S. (2008). The DCC curation lifecycle model. *The International Journal of Digital Curation*, *3*(1), 134–140.

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

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

Johnston, L. R. (Ed.). (2017). *Curating research data volume two: A handbook of current practice* (Vol. 2). Association of College and Research Libraries.

Johnston, L., Farrell, S., Herold, P., & Stewart, C. (2019). *Final report of the research data services strategic planning task force delivered March 13, 2018*. 51.

Kaufman, D. S. (2018). Technical note: Open-paleo-data implementation pilot – the PAGES 2k special issue. *Climate of the Past*, *14*(5), 593–600. https://doi.org/10.5194/cp-14-593-2018

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 Edition)*. Digital Preservation Coalition. https://doi.org/10.7207/twr14-02

McCann, L. (2013). Preservation as obstacle or opportunity? Rethinking the preservation-access model in the age of MPLP. *Journal of Archival Organization*, *11*(1–2), 23–48. https://doi.org/10.1080/15332748.2013.871972

McCrea, D. E. (2006). Getting more for less: Testing a new processing model at the University of Montana. *The American Archivist*, *69*(Fall/Winter), 284–290.

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

Meissner, D., & Greene, M. A. (2010). More application while less appreciation: The adopters and antagonists of MPLP. *Journal of Archival Organization*, *8*(3–4), 174–226. https://doi.org/10.1080/15332748.2010.554069

Ness, C. (2010). Much ado about paper clips: "More product, less process" and the modern manuscript repository. *The American Archivist*, *73*(1), 129–145. https://doi.org/10.17723/aarc.73.1.v17jn363512j545k

Niu, J. (2016). Organisation and description of datasets. *Archives and Manuscripts*, *44*(2), 73–85. https://doi.org/10.1080/01576895.2016.1179585

Oden, J. T., Ghattas, O., King, J. L., Schneider, B. I., Bartschat, K., Darema, F., Drake, J., Dunning, T., Estep, D., Glotzer, S., & Gurnis, M. (2011). *National Science Foundation Advisory Committee for Cyberinfrastructure Task Force on Grand Challenges*. https://www.nsf.gov/cise/oac/taskforces/TaskForceReport_GrandChallenges.pdf

Palmer, C. L., Weber, N. M., & Cragin, M. H. (2011). The analytic potential of scientific data: Understanding re-use value. *Proceedings of the American Society for Information Science and Technology*, *48*(1), 1–10. https://doi.org/10.1002/meet.2011.14504801174

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

Persaud, B., Sookoo, N., Bocaniov, S., Szigeti, K., Van Wychen, W., & Van Cappellen, P. (2019, December 1). *Data rescue: No pain, no gain - rescuing historical USSR meteorological data*. American Geophysical Union, Fall Meeting 2019. https://agu.confex.com/agu/fm19/meetingapp.cgi/Paper/494793

Piazza, P., Cummings, V., Guzzi, A., Hawes, I., Lohrer, A., Marini, S., Marriott, P., Menna, F., Nocerino, E., Peirano, A., Kim, S., & Schiaparelli, S. (2019). Underwater photogrammetry in Antarctica: Long-term observations in benthic ecosystems and legacy data rescue. *Polar Biology*, *42*(6), 1061–1079. https://doi.org/10.1007/s00300-019-02480-w

Png, I. P. L., Chen, Y.-N., Chu, J., & Feng, Y. (2019). *China weather, 1912-49: Data rescue* (SSRN Scholarly Paper ID 3454857). Social Science Research Network. https://papers.ssrn.com/abstract=3454857

Poole, A. H. (2016). The conceptual landscape of digital curation. *Journal of Documentation*, *72*(5), 961–986. https://doi.org/10.1108/JD-10-2015-0123

Pryor, G. (2012). Why manage research data? In G. Pryor (Ed.), Managing Research Data (1st ed., pp. 1–16). Facet. https://doi.org/10.29085/9781856048910.002

Punzalan, R., Kriesberg, A., Daniels, M., & Gucer, K. (2016). *National Agricultural Library: Digital Curation Plan*. National Agricultural Library and University of Maryland College of Information Studies. https://drum.lib.umd.edu/handle/1903/26358

Rountree, R. A., Perkins, P. J., Kenney, R. D., & Hinga, K. R. (2002). Sounds of western north Atlantic fishes—Data rescue. *Bioacoustics*, *12*(2–3), 242–244. https://doi.org/10.1080/09524622.2002.9753710

Sanz, F., Pognan, F., Steger-Hartmann, T., Díaz, C., Cases, M., Pastor, M., Marc, P., Wichard, J., Briggs, K., Watson, D. K., Kleinöder, T., Yang, C., Amberg, A., Beaumont, M., Brookes, A. J., Brunak, S., Cronin, M. T. D., Ecker, G. F., Escher, S., … Zamora, I. (2017). Legacy data sharing to improve drug safety assessment: The eTOX project. *Nature Reviews Drug Discovery*, *16*(12), 811–812. https://doi.org/10.1038/nrd.2017.177

Searcy, R. (2017). Beyond control: Accessioning practices for extensible archival management. *Journal of Archival Organization*, *14*(3–4), 153–175. https://doi.org/10.1080/15332748.2018.1517292

Slonosky, V., Sieber, R., Burr, G., Podolsky, L., Smith, R., Bartlett, M., Park, E., Cullen, J., & 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

Specht, A., Bolton, M., Kingsford, B., Specht, R., & Belbin, L. (2018). A story of data won, data lost and data re-found: The realities of ecological data preservation. *Biodiversity Data Journal*, *6*, e28073. https://doi.org/10.3897/BDJ.6.e28073

Thompson, M., & Ramsey, M. H. (1995). Quality concepts and practices applied to sampling—An exploratory study. *Analyst*, *120*(2), 261–270. https://doi.org/10.1039/AN9952000261

Vearncombe, J., Riganti, A., Isles, D., & Bright, S. (2017). Data upcycling. *Ore Geology Reviews*, *89*, 887–893. https://doi.org/10.1016/j.oregeorev.2017.07.009

Waugh, D., Russey Roke, E., & Farr, E. (2016). Flexible processing and diverse collections: A tiered approach to delivering born digital archives. *Archives and Records*, *37*(1), 3–19. https://doi.org/10.1080/23257962.2016.1139493

Weideman, C. (2006). Accessioning as processing. *The American Archivist*, *69*(2), 274–283. https://doi.org/10.17723/aarc.69.2.g270566u745j3815

Williams, S. F., Rilling, R., & Stossmeister, G. (2019). NCAR's Earth Observing Laboratory Legacy Field Campaign Data Rescue. *AGU Fall Meeting Abstracts*, *21*. http://adsabs.harvard.edu/abs/2019AGUFMED21A..07W

Wippich, C. (2012). Preserving science for the ages—USGS data rescue. In *Preserving science for the ages—USGS data rescue* (USGS Numbered Series No. 2012–3078; Fact Sheet, Vols. 2012–3078). U.S. Geological Survey. https://doi.org/10.3133/fs20123078

Wyborn, L., Hsu, L., Lehnert, K., & Parsons, M. A. (2015). Guest Editorial: Special issue: Rescuing legacy data for future science. *GeoResJ*, *6*, 106–107. https://doi.org/10.1016/j.grj.2015.02.017

Yu, X., Lamačová, A., Shu, L., Duffy, C., Krám, P., Hruška, J., White, T., & Lin, K. (2020). Data rescue in manuscripts: A hydrological modelling study example. *Hydrological Sciences Journal*, *65*(5), 763–769. https://doi.org/10.1080/02626667.2019.1614593