Data Rescue at the U.S. National Agricultural Library: Case Studies of 3 Hybrid Collections

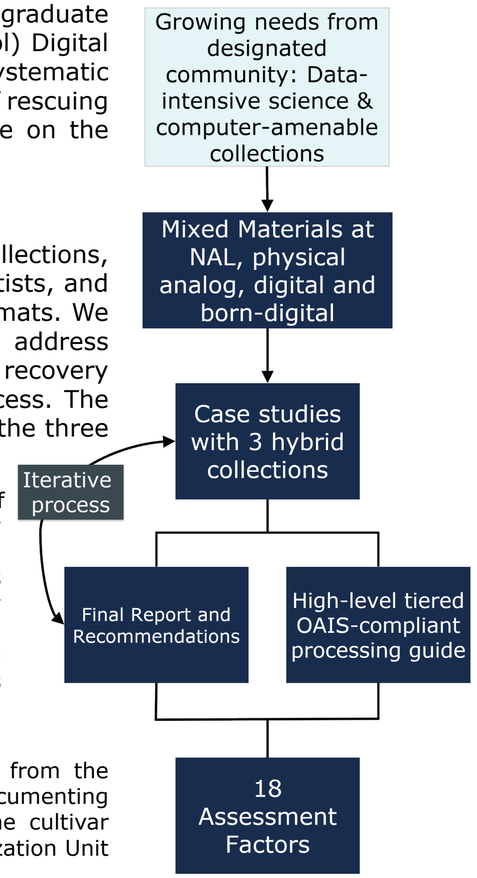
drum.lib.umd.edu/handle/1903/26475

authors
Hilary Szu Yin Shiue, Cooper Clarke,
Katrina Fenlon, PhD (PI)

Background

While the open science movement has facilitated discussions in managing, preserving and curating data from active and ongoing research, data rescue efforts respond to the growing recognition of the value and reuse potential of data biding in unpublished records and collections of legacy research materials. We define data rescue as 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.

The Data Rescue project at the National Agricultural Library (NAL) was led by graduate fellows in the University of Maryland College of Information Studies (iSchool) Digital Curation Fellowship, a cooperative agreement, to assist NAL in creating systematic approaches to appraise and process legacy collections with the specific goal of rescuing reusable or historically valuable scientific data, and making them available on the Department's open data repository, Ag Data Commons.

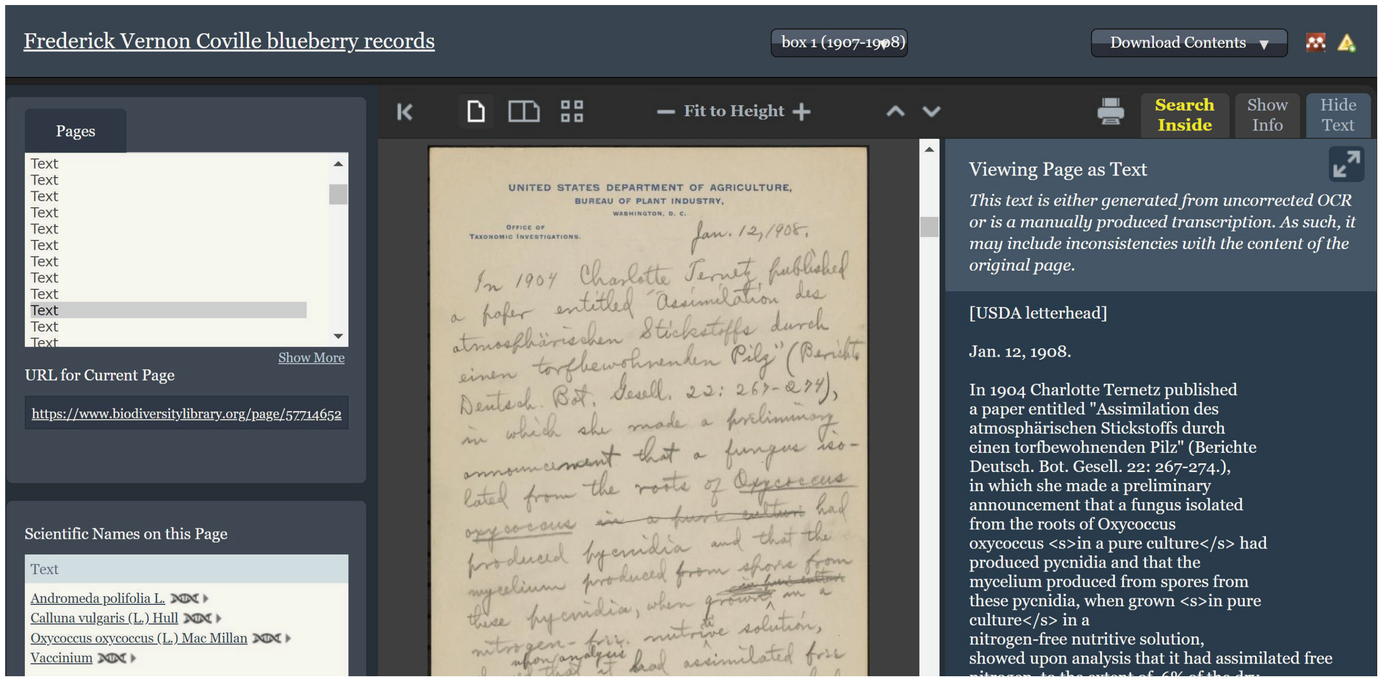


Methodology

NAL is confronted with the opportunity and challenge of extensive donated collections, such as scientific records in myriad formats from lab closures, retiring scientists, and accumulated historical collections of materials largely in physical analog formats. We conducted three case studies with both analog and digital collections to address diverse processing considerations, such as metadata collection, data recovery challenges and methods. The flowchart on the right demonstrates our process. The table on the far right presents the applications of 18 assessment factors on the three case studies:

1. **Frederick Vernon Coville Blueberry Records (1907-1938)**: This collection of hand-written research notes and other documents represents the USDA blueberry records of Coville, documenting the earliest crosses of commercial blueberries.
2. **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.
3. **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.

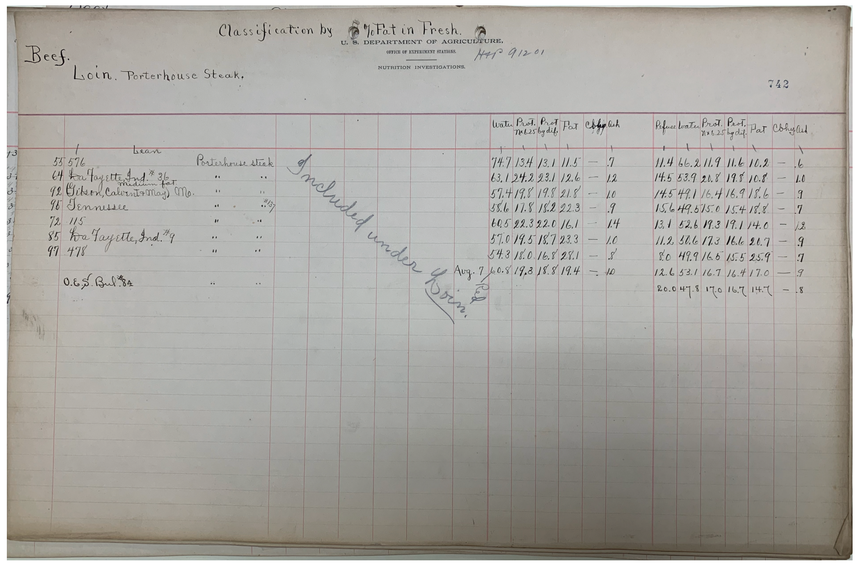
Below: Example of the available transcription and linked scientific data extracted from the Frederick V. Coville Blueberry Records (MS 413, USDA-NAL Special Collections), documenting Coville's blueberry cultivar improvements, this data resulted in the creation of the cultivar Bluecrop which is still consumed today. The transcription was completed by the Digitization Unit of NAL and is available at biodiversitylibrary.org/page/57714652



Assessment Factors	Frederick V. Coville's Blueberry Notes Collection (1907-1938)	Wilbur Olin Atwater Papers (1891-1906)	Rufus Chaney born-digital collection (1989-2014)
Extent	6 linear feet; 24 customized boxes.	900 handwritten sheets	262 files
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
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	Significant contributions to the study of heavy metals present in soil and their uptake in crops, the application of biosolids to cropland, and phytoextraction of contaminated soil.
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, FDA and EPA
Reuse objects	Detailed pedigree information for both released and unreleased cultivars, plant characteristics and inheritance	Raw data set (subject to ongoing scientific citations)	Raw data set, analytics system files (.sas)
Historical objects	The whole collection of Coville's blueberry notes are of historical value as it is a century-old collection.		
Fit for purpose	Additional processing and transcription are 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.	Verification of variable names; connection between data set and publications, further interpretation and documentation
Obstacles for recovery	Mix of analog and digitized materials; Fragility of analog materials; Loose leaf pages vulnerable to loss of original order	Fragility of analog materials; Handwritten tabular data; Inclusion of handwritten margin notes, handwritten strike-throughs	Missing context and metadata require expert consultation; Determining completeness within files; Access to outmoded software originally used to create the files

Other assessment factors		
Data objects	Associated publications	Relevant collections
Completeness	Sensitivity	Access and use restrictions
Priorities	Reproducibility	Rarity and uniqueness

Below: Example of a handwritten data sheet (no. 742) from the Wilbur O. Atwater Papers (MS 261, USDA-NAL Special Collections), documenting Atwater's studies of food nutrition and caloric composition, this data fundamentally changed the Department's approach to food nutrition and popularized the use of calories to measure the energy in food.



Supported by 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.
Poster created in November 2020 for the RDA 16th Plenary poster session.

COLLEGE OF INFORMATION STUDIES

USDA National Agricultural Library U.S. DEPARTMENT OF AGRICULTURE

Acknowledgements: 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, 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.