ISSN 2161-3974

JeSLIB 2021; 10(3): e1210
https://doi.org/10.7191/jeslib.2021.1210

Journal of eScience Librarianship
putting the pieces together: theory and practice

Full-Length Paper

Active Curation of Large Longitudinal Surveys: A Case Study

Inna Kouper[1,2], Karen L. Tucker[2], Kevin Tharp[2], Mary Ellen van Booven[2], and Ashley Clark[2]

[1] Indiana University, Bloomington, IN, USA
[2] Center for Survey Research, Indiana University, Bloomington, IN, USA

Abstract

In this paper we take an in-depth look at the curation of a large longitudinal survey and activities and procedures involved in moving the data from its generation to the state that is needed to conduct scientific analysis. Using a case study approach, we describe how large surveys generate a range of data assets that require many decisions well before the data is considered for analysis and publication. We use the notion of *active curation* to describe activities and decisions about the data objects that are "live," i.e., when they are still being collected and processed for the later stages of the data lifecycle. Our efforts illustrate a gap in the existing discussions on curation. On one hand, there is an acknowledged need for active or upstream curation as an engagement of curators close to the point of data creation. On the other hand, the recommendations on how to do that are scattered across multiple domain-oriented data efforts.

Correspondence: Inna Kouper: inkouper@indiana.edu

Received: April 8, 2021 **Accepted**: June 4, 2021 **Published**: August 11, 2021

Data Availability: The paper documents a case study of curating an active dataset. The data and related products are under embargo until the end of the project.

Disclosures: The authors report no conflict of interest.

Abstract Continued

In describing the complexities of active curation of survey data and providing general recommendations, we aim to draw attention to the practices of active curation, stimulate the development of interoperable tools, standards, and techniques needed at the initial stages of research projects, and encourage collaborations between libraries and other academic units.

Introduction

Surveys and observations remain primary methods in social science research (Wright and Marsden 2010). As the number of surveys and related data continue to grow, especially, in the context of new data sources and opportunities, so do the needs to optimize their creation, preservation, and distribution (King 2011). Generating quality survey data is labor intensive, and curatorial actions require support that combines technical, social, cultural, and organizational aspects. Better understanding of those aspects of data curation can maximize efficiencies and release untapped resources to the research community.

While many universities have begun thinking about these challenges and developing their research data services, the approaches are relatively isolated, confined to separate units or organizations within institutions. The data ecosystem in academic institutions is diverse, and the units that are involved in data stewardship include research labs and centers, externally funded projects, libraries, computing support centers, and so on. Long-term solutions for data benefit from cross-unit cooperation and coherent institutional frameworks, but before such cooperation and frameworks can be established, we need a better understanding of how these various units approach the creation, curation, and reuse of data (Macdonald and Martinez-Uribe 2010; Rice 2009; Yakel, Faniel, and Maiorana 2019).

In this paper we take an in-depth look at the curation of a large longitudinal survey and activities and procedures involved in moving the data from its generation to the state that is needed to conduct scientific analysis. Using a case study approach, we describe how such surveys generate a range of data assets that require many decisions well before the data is considered for analysis and publication. We use the notion of active curation to describe activities and decisions that happen when the data objects are "live," i.e., when they are still being collected and processed for the later stages of the data lifecycle (Akmon et al. 2017; Goble et al. 2008). In describing the complexities of active survey data curation we aim to a) draw attention to the practices of active curation and the role of various professionals in it, b) stimulate the development of interoperable tools, standards, and techniques needed at the initial stages of research projects, and c) encourage collaborations between libraries and other academic units in building services and workflows that support data across both earlier (live) and more final (published or archived) states.

Background

Data curation is an important step in research, although the professional status and institutional roles of data curators are still under discussion (Higgins 2011; Tammaro et al. 2017; Weber, Palmer, and Chao 2012). It includes maintaining and improving the quality of data, adding metadata, and ensuring that the data is available for others to use through persistent identifiers and viable repositories (Giaretta 2004; Yakel 2007). Data curation overlaps with other terms, including

data management, preservation, and archiving. The choice of terminology depends on organizational expertise and priorities, but the term "curation" is often used as a broader concept that guides and defines the rest (Abbott 2008; Constantopoulos et al. 2009; Steinhart et al 2008).[1]

Many curation approaches are conceptualized along the research lifecycle that starts with project planning and creation of data and continues into its dissemination and re-use (Ball 2012; Higgins 2008; Lord and Macdonald 2003). In survey research the earlier stages of the data lifecycle, especially data processing and administration, are often part of the scientific processes (IFD&TC 2021; Couper 1998). Being "the least glamourous aspects of survey research," data management includes bringing data into an appropriate digital form, editing, coding, transforming, and cleaning the data, and ensuring its quality and access (Davis and Smith 1992; Singleton and Straits 2009). Earlier approaches combined data management with project management and included managing people, budgets, and data (van Kammen and Stauthamer-Loeber 1998). Handbooks on social science research task researchers themselves with managing metadata, deploying databases, and integrating multiple software and data components or gathering and analyzing information about survey processes, or paradata (Groves and Heeringa 2006; Lavrakas 2008).

The early involvement of curators in data production is crucial to alleviating the burden on researchers, promoting the use of data, and making data available for early insights and discoveries (Lord et al. 2004). The value of early and rapid curation has been demonstrated recently during the COVID-19 pandemic, when data on multiple aspects of the virus and associated social, behavioral, and epidemiological variables was needed fast (Johns Hopkins Coronavirus Resource Center n.d.; RDA COVID-19 Working Group 2020). The COVID-19 data example shows multiple interdependencies between research and curation and the need to meet researchers "upstream" (Scientific Data Curation Team 2020). In practice, most of the curation activities still focus on the discovery and preservation layers, leaving data creation, cleaning and other tasks to data creators, analysts, or IT professionals (Beheshti et al. 2018; Chu et al. 2016; Downs and Chen 2010; Johnston et al. 2018; Julkowska et al. 2019; Lee and Stvilia 2017; Wynholds 2011).

Curation is important for all stages of the data lifecycle as these stages are mutually dependent and decisions made at each stage have cumulative effect (Wallis et al. 2008). At the same time, the curation activities that are performed closer to the data origin are different as they need to respond to the "messy and quirky" acquisition of scientific data (Baker and Yarmey 2009). While academic

1 Approaches that consider data management as a broader concept that includes curation can also be found in the literature, see, for example, Qin et al 2014. A detailed discussion about the differences and advantages or disadvantages of either of the approaches is beyond the scope of this paper. We use "curation" rather than "management" to emphasize the importance of "looking after" research assets rather than "dealing with" or "controlling" them as the dictionary definition of the term "management" would suggest.

libraries are actively involved in supporting research data curation, they face an overwhelming number of decisions about how to implement such services (Akers et al. 2014; Bracke 2011; Cox et al. 2017; Tenopir, Birch, and Allard 2012). Moreover, the libraries may not have the infrastructure and workforce capacity to support all data needs of the university across the data lifecycle, particularly, the curation upstream at the point of creation or the oversight of data and documentation retention and destruction (Oliver and Harvey 2016). The case study described below points to the need of developing new approaches to active curation that combine existing expertise with new tools and techniques that facilitate working with multiple data objects and processes and capturing constant change.

The Person to Person Health Interview Study

The survey described in this paper, called the Person to Person Health Interview Study, is part of the Indiana University Precision Health Initiative[2] that develops a personalized approach to prevention and treatment of diseases, taking into account individual genes, environment, and lifestyle. In addition to collecting information about diseases and studying their genetic foundations and interactions, the project launched a longitudinal survey study. The study is a collaboration between a core science team comprised of faculty from diverse social and health science disciplines, a data team comprised of experts from an academic survey research center, and other partners, such as a sampling vendor and a biobank. The study has the following goals:

- To understand the relationship between the person's genetics, their social and physical environments, attitudes, and behaviors, and their ability to respond to and recover from various health-related events.
- To create a multi-level dataset that integrates data on genetics, biology, and the sociocultural and physical environments and can be used in various contexts.

The core science team is led by the Principal Investigator (PI) with support from a research director, project manager, and science advisory team. The core science team collaborates with our data team at the IU Center for Survey Research (CSR), an academic survey research center that is responsible for the implementation of the survey and data collection and processing, in other words, for the active curation activities described below. Our data team includes a project director, two study supervisors who manage 20 - 30 field interviewers across the state, a software developer, and a data curator. The team is the steward of the survey data and all products associated with data collection and preparation; it performs continuous and systematic curation and delivers clean de-identified data to the core science team for further analysis and publications.

The study was designed as a representative survey of a random sample of over

2 https://precisionhealth.iu.edu

2,000 residents of one state. The survey consists of hundreds of questions and, in addition to demographics and employment history, collects biometric information, such as height, weight, blood pressure, and so on, information about health behaviors and attitudes and levels of physical activity and fatigue, and social network information. The study also includes a mental health assessment via an external mental health screening platform and a collection of saliva samples from each participant that are used for DNA sequencing and subsequent matching of the genetic traits with the survey data.

In addition to the information described above, more information is collected to enrich the dataset and enable its future use. This additional information includes audio recordings of respondent concerns using their own words, contact information for current and future contacts and participant tracking, photographic images of participants, informed consent for the survey, audio, and photos, and comments from field staff conducting the interviews.

Active Curation Activities

To ensure the success of this complex survey that combines several modalities, including in-person visits, telephone calls, and computer-assisted adaptive surveying, the data team developed and maintains the infrastructure that supports survey administration and data collection and delivery, field operations, and all other tasks related to survey data and its collection and processing. All these components can be broken down into the following categories: development, collection, management, and delivery and analytics (see Figure 1).

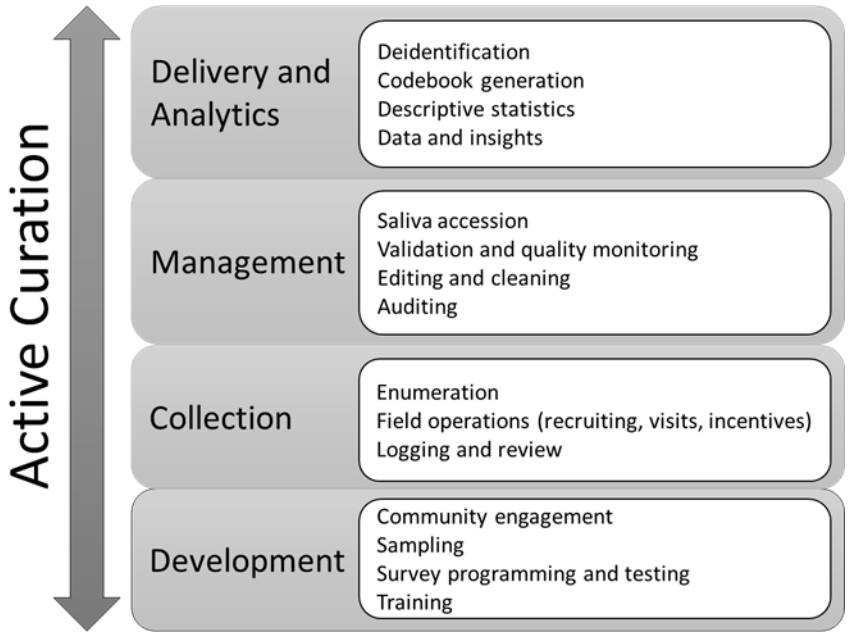

Figure 1: The survey project curation components performed by the survey research center team.

Development includes developing forms and questionnaires, obtaining IRB approval, recruiting and training office and field staff, community engagement, sampling, and design of the IT infrastructure that will support the rest of the components. Community engagement, for example, is a key component to gaining cooperation from research participants and addressing the challenges of decreasing trends in survey response rates and increasing fieldwork (Beullens et al. 2018). Building a strong IT infrastructure is another crucial component of the development stage as most of the curation efforts are grounded in technology. A decision log that documented technical and other decisions throughout the development phase helped the team to develop a shared understanding of the project and was later transformed into the Manual of Procedures (MoP) that the team can reference.

Collection of data includes establishing contacts with the households from the sample, recruiting and selecting eligible participants (enumeration), consenting subjects, administering the survey, establishing field operations, and organizing a system of logging and reviewing all the collected assets. Field operations involve working with and training field interviewers, organizing visits to the households and documenting their outcomes, and managing the incentives (payments for participating in the survey via gift cards). In-person and phone surveys are more complicated than web surveys, especially, when attempts are made to convert non-respondents and refusals and increase the number of participants. All information that gets collected during this stage becomes part of the data collected in the project and helps to ensure the quality of the survey output.

Management includes all the activities of reviewing and transferring data from its original sources (e.g., tablets or third-party platforms) into a unified platform and then editing, cleaning, validating, and re-coding the data. At this stage all the data components, including saliva samples, survey responses, incentives, and other assets get verified for consistency and quality and processed into packages ready for further analysis. Occasional audits of all components of data collection and management processes are part of this stage as well.

Delivery and analytics is the final stage of active survey curation activities performed by the data team, although curation does not end with the delivery and continues as the survey research center keeps a copy of the data and documentation. At the delivery stage, the data is prepared for the delivery to the core science team for their analysis and publications. A de-identified dataset along with the codebook is saved separately and shared with designated collaborators. To facilitate further scientific activities, grant proposals, and dissemination activities, the data team also performs analytics by request and identifies preliminary trends and patterns in the data.

Development

Typically, development begins as soon as the study is approved by an Institutional Review Board or any other committee that is responsible for monitoring and

reviewing research that involves human subjects. However, complex "omnibus" surveys, such as this one, can take a long time to approve, which negatively affects the project timeline and delays all the subsequent stages. To address this, we introduced a staggered time framework, so that the stages have an overlap between them with the later stages beginning before the earlier stages end. For the development stage, our team did preliminary work while the survey itself was still in development. Thus, we reviewed available tools and platforms, pre-programmed parts of the survey using our previous experience with health and social science surveys and began preparing training materials.

Developing an appropriate sampling methodology is part of any survey research as it provides scientific techniques to study populations without full enumeration (Brick 2011). NORC at the University of Chicago was contracted to develop a sampling plan for the survey in collaboration with project staff. In parallel with sampling planning, the data team began programming the survey. At first, it was programmed in REDCap on Android tablets and tested in a pilot study conducted in November 2018 (Project REDCap n.d.). During the pilot study the REDCap mobile app showed slow performance due to the length and complexity of the survey. The mobile app had poor usability and insufficient flexibility in designing transition screens and implementing data and logic checks that were needed for the survey of this complexity. Concerns regarding its flexibility and the requirement of internet connectivity for administering the external mental health module resulted in the team selecting NORC and its proprietary case management system, NSMobile, and a relatively new survey software, Dooblo.

The survey also had to be modified as the result of the pilot as the survey was still being developed. The questions were modified for clarity, additional instructions for interviewers were added, and survey responses were expanded to include some unanticipated categories of responses. While the survey was in pilot testing, the core science team has added more questions and sections, including audio recordings of participants' experiences and opinions, anthropometric measurements and screening questions, the family history, treatment stigma and cognitive assessment. The survey became even more complicated and required more programming and testing efforts. The training materials, informed consent, and data collection forms also had to be modified to reflect the changes.

Another third party was involved in development. Adaptive Testing Technologies (ATT) designed and implemented the instruments that collected information about participants' mental health, CAT-MH and CAT-SA (Adaptive Testing Technologies n.d.). These computerized adaptive instruments, created outside of our project, collect self-reported ratings about depression, anxiety, substance use, and other mental health issues. In the end the instrument provides estimates of the severity of each mental health issue that was evaluated. As part of the development stage, the survey was programmed to navigate to CAT website and return once that part of the survey was over.

The main products of curation at the development stage are the sample

with individual cases, the survey instrument, including its multiple versions and changes, pilot data, and the tools needed throughout the project and their documentation. We load the sample received from NORC into the case management system, review the instrument and make sure it includes all the necessary components and logic. We also connect the programmed survey to case management, and consent and incentives modules. The documentation of piloting and testing becomes part of the data curation record. Its results allow us to not only administer the current survey, but also to build a knowledge base for subsequent surveys. Later, the team comes back to this documentation and updates it with weaknesses that we find or anything that did not work as planned.

Collection

Data collection begins with household enumeration, or identification of eligible participants to be enrolled in the study. It then goes through several iterative stages that include obtaining participants' consent and conducting the survey, addressing non-responses and refusals, and documenting the processes, artifacts, and data. A simplified data collection workflow is presented in Figure 2 below:

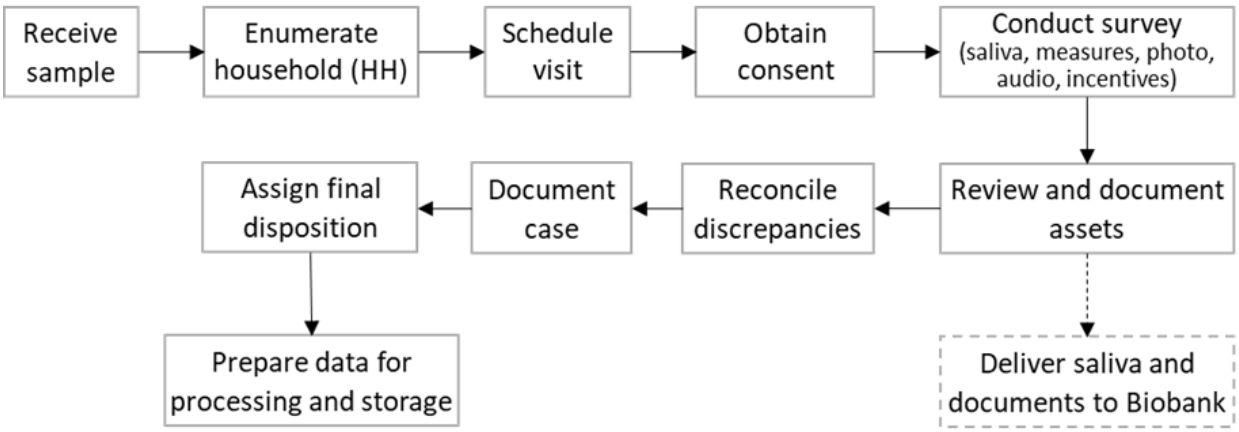

Figure 2: Data collection workflow.

Most of the data collection centers around the actual home visit, which is divided into three stages: pre-visit, visit, and post-visit. During the pre-visit stage the field interviewers review and verify participant history, confirm directions to home and appointment time, and ensure that all document forms and supplies are available and packed according to protocols. This eliminates unnecessary delays during the home visit and helps reduce skipped and rescheduled appointments. The interviewers are provided with checklists that they go through in preparing for visits, which helps to avoid many of missteps (Hales and Pronovost 2006).

During the home visit the interviewers obtain electronic consent, collect information by going through the survey, performing the anthropometric measures (height, weight, etc.), collecting a saliva sample, and providing participants with

gift cards.[3] After the visit is over, the interviewers go over another post-visit checklist, update the case management system, register receipt of saliva and document gift card distribution, and update the inventory system with expended materials.

Case documentation and asset logging are key for good data collection. In addition to the third-party case management system, which allows to track the detailed history of contact attempts, including the invitations, and the date, content, and outcome of each visit, we have developed a system for tracking consent, incentives, and supplementary survey materials, such as audio and photos. The system is an MS Access database with a console that allows multiple users to log and document the assets.

The need for such an elaborate tracking system comes with the complexity of this survey. Several modules in the survey require separate consent, and multiple personnel works with cases and collects and processes all the additional assets. The diagram in Figure 3 provides a schematic representation of the asset logging console, which allows to make sure that nothing in the data collection process is missing or overlooked:

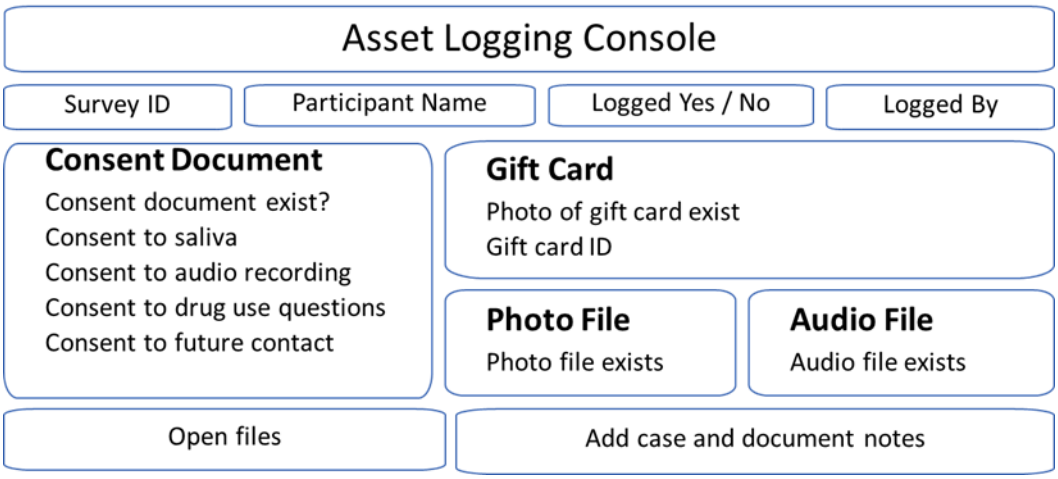

Figure 3: A schematic representation of the survey asset logging console.

As a stage of research lifecycle, data collection is inseparable from curation. Each step in the data collection workflow is tracked daily for completion, and the information about each step defines the next steps and the actions that will be performed, whether its proceeding to the next step, applying corrections, or making changes. The interviewers return to households to enumerate and conduct the survey up to 10 times, and all this information needs to be part of the decision-making process. Refusals and scheduled appointments become part of

3 The COVID-19 pandemic added another layer of complexity, when the interviewers had to screen for COVID-19 and at some point, switch to telephone interviews. This is beyond the scope of this paper.

the curation so that various checks and modifications can be done on an ongoing basis. **The curation at this stage involves study information materials, cases (information about respondents), survey supplies (forms and devices), consent documents, survey outcomes and paradata, including interviewer comments and dispositions, incentives, physical samples (saliva), and supplementary survey materials (photo, audio).**

Management

Data management in curation activities ties together all other activities and ensures that all data assets are documented and monitored for quality. Management of our survey includes saliva accession, validation and quality monitoring, data editing and cleaning, and auditing.

Saliva accessioning follows its own protocol with daily registering and tracking and a delivery to the biobank. Additionally, an encrypted data file containing the identifying data on respondents (i.e., respondent first and last name, DOB, race, ethnicity, and birth sex) is shared electronically via secure file sharing options. Saliva processing has a designated curator who is responsible for making sure that a) all samples are delivered to the biobank, b) the documentation supports daily tracking and reconciliation, and c) each sample is connected to the right respondent.

A validation survey is used to check and verify the correctness of survey responses. It is a small survey that is sent to the original sample participants, re-asking several questions where responses should not have changed from the initial survey. The validation survey is distributed via email or phone using the survey software CASES, a package for collecting survey data developed by the Computer-assisted Survey Methods (CSM) Program at the University of California, Berkeley.[4] The validation survey responses comprise an additional dataset that needs to be stored, reviewed, and used for necessary corrections in the main survey database.

The survey responses from the tablets are first synchronized with the NORC study database, and then are deposited into our servers at the end of each day, with separate folders for consent, incentives, audio files, and saliva tracking documentation. A Bulk Rename Utility[5] is used to copy and rename all the files to a standard convention, while keeping the originals intact. The convention is {identifier}_{type of file}, e.g., id_consent.pdf or id_incentive.png. The raw data from our servers is then imported into the main survey database, a relational SQL database located on a designated Microsoft SQL server. The server is secured with access limited to IT personnel and data analysts and uses the least privilege principle for additional protection of the personally identifiable information (PII). In addition to data from the tablets, the database integrates mental health data from

4 https://cases.berkeley.edu
5 https://www.bulkrenameutility.co.uk/Main_Intro.php

ATT provided via an API. The data is downloaded in JSON format and then loaded into the database along with other survey data.

The database is designed using an Entity–Attribute–Value data model (EAV). This model, which is also known as a vertical database model or open schema model, encodes data into few columns, namely, the entity (the item or case that is being described), the attribute (the name of the variable and any other parameters, such as variable range, type, and timestamps for recorded values), and the value of the attribute (Marenco et al. 2003). Such a design accommodates varying sparseness of large surveys and allows to avoid database redesign in an evolving data collection situation.

Prepared SQL code (stored procedures) runs regularly to check for new or changed responses in the raw data and import all detected changes into the database. Variable names and response codes are converted to match the final survey specification definitions, as unintended differences cropped up during survey programming. For items lacking a response, standardized nonresponse codes are applied to indicate when items were skipped (not asked based on certain conditions), not answered, or not available due to technical issues (e.g., tablets not synching, etc.)

Storing all the data in a centralized database allows us to generate various reports for early identification of any gaps and discrepancies in data collection. Integrating multiple components of data, including paradata, metadata, supplementary data, and survey data allows for continuous monitoring, identification of any missing data, and coordination across all stages of the project.

For quality purposes the following procedures are established as part of our active curation approach. All attempts to contact, recruit, and interview participants are recorded and reviewed for completeness and accuracy. A random subset of completed household enumerations (5% per interviewer) is reviewed to make sure that enumeration, appointments, and participant information are recorded as needed. Another random subset of participants (10% per interviewer) is reviewed to verify that contact and tracking information, as well as consent and questionnaire information are appropriately recorded.

The contact tracking and tracing system is another component of active curation. This component has been developed in REDCap to track respondents' contact information over time. Such a system is critical for longitudinal studies for retaining participants and recruiting them in follow-ups and future studies. We are now actively tracing those lost to contact to obtain new addresses and phone numbers by searching free, online services such as fastpeoplesearch.com and verifying with two sources before changing an address in the system. In the future, such a system would integrate multiple studies and provide a searchable database with the latest most accurate contact information of current and potential survey participants.

Data cleaning and editing are two other significant curation activities that are performed in two stages. First, we check all supplementary data, such as audio, field interviewer comments, and survey disposition codes. We check them for quality, remove unnecessary or sensitive information, and reconcile discrepancies to make sure refusals, partial completions, and full completions are coded consistently across all survey components. Full and partial completions become part of the final aggregated dataset delivered to the core science team.

The second stage of cleaning and editing involves checking the main survey data. The data in the long EAV format from the database is pivoted into a wide format and exported into the SPSS format. The data team checks all survey components for correctness and completeness, check that survey logic and skipping have been handled correctly, verify there are no unintentional missing responses, check for outliers and responses that are out of range. Any unusual pattern or outlier values are investigated, and data is edited as appropriate. We also verify that variables labels (survey questions) match exact instrument wording as closely as possible and edit and code open-ended questions.

To perform the cleaning, all answers to text variables are exported into csv format and loaded into OpenRefine, a standalone open-source desktop application for data cleanup and transformation. The csv file exported from the main database contains a case ID, the name of the variable (item), and the text to be cleaned (value_text, see Figure 4 below).

rows

ıs: **rows** records Show: 5 10 **25** 50 rows

	su_id	item	value_text
1.	61758458	CONSTRUCTION_SPECIFY	Architect
2.	69902638	CONSTRUCTION_SPECIFY	building structures
3.	34646870	CONSTRUCTION_SPECIFY	carpentry
4.	29997160	CONSTRUCTION_SPECIFY	demolition
5.	39633101	CONSTRUCTION_SPECIFY	insurance work- a little of everything
6.	28801290	CONSTRUCTION_SPECIFY	labor on highway
7.	27827130	CONSTRUCTION_SPECIFY	Landscaping
8.	23090270	CONSTRUCTION_SPECIFY	moving dirt
9.	27746840	CONSTRUCTION_SPECIFY	owner of construction company
10.	61070218	CONSTRUCTION_SPECIFY	pipe fitting
11.	24384050	CONSTRUCTION_SPECIFY	remodel
12.	61421908	CONSTRUCTION_SPECIFY	vinyl siding
13.	63866148	CROPS_SPECIFY	Flowers
14.	38587120	CROPS_SPECIFY	oats
15.	63040168	CROPS_SPECIFY	rice, vegetables
16.	39633101	CROPS_SPECIFY	Sod, blueberries
17.	27473630	CROPS_SPECIFY	strawberries, cucumbers, tomatoes, potatoes, blackberries
18.	61674878	CROPS_SPECIFY	tomatoes
19.	20843440	CROPS_SPECIFY	vegetables
20.	35717050	EMOT_ELSE_SPEC	change diet
21.	22972400	EMOT_ELSE_SPEC	gene type test
22.	27473630	EMOT_ELSE_SPEC	go for a drive

Figure 4: Free text cleaning preparation.

To keep the original data as a reference, we duplicate the "value_text" column and create another column called "value_text_cleaned". We apply multiple text clustering algorithms that OpenRefine provides, including four key collision algorithms and two nearest neighbor algorithms, which allows to catch many misspellings and differences in capitalization (see Figure 5):

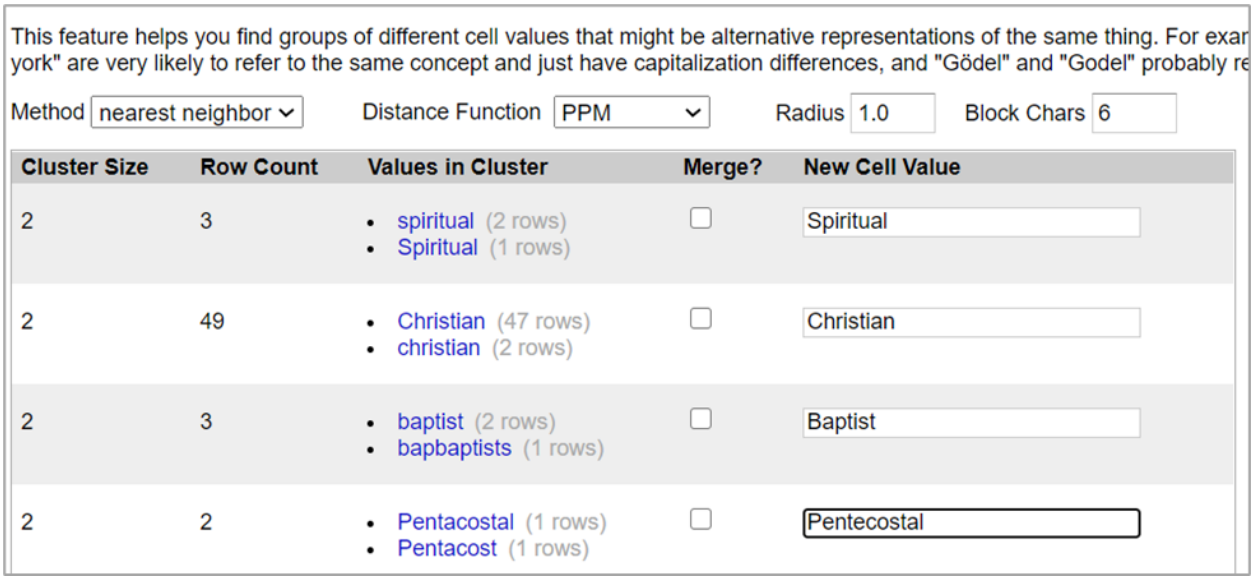

Figure 5: Cleaning free text via clustering in Open Refine.

When no additional automated clustering can be performed, we manually standardize spelling variations to bring the number of open-ended responses to a manageable number of categories. When responses, such as specifying emotional problems, do not allow for a simple standardized taxonomy without consulting a mental health professional, the answers are standardized as much as possible without assigning categories. Figure 6 below illustrates standardization to a form "condition" or "condition1, condition2", which enables easier processing in the future:

39.	26364970	EMOT_PROB_SPEC	anxiety, binge eating	ANXIETY, BINGE EATING
40.	22972400	EMOT_PROB_SPEC	anxiety, depression	ANXIETY, DEPRESSION
41.	33352710	EMOT_PROB_SPEC	anxiety-depression	ANXIETY, DEPRESSION
42.	65303908	EMOT_PROB_SPEC	Anxiety/Depression	ANXIETY, DEPRESSION
43.	58770238	EMOT_PROB_SPEC	Anxiety/Paranoia	ANXIETY, PARANOIA

Figure 6: Free text standardization.

All the cleaning is then incorporated back into the main database.

The management stage of active curation includes numerous curation objects, including data from the survey, saliva samples, full contact history for each case, time cost of data collection, e.g., travel time, hours per each complete, and so on, reports from field interviewers, and statistics on field interviewer performance. While some of this information may be considered internal project data used for internal monitoring and evaluation, most of it is eventually shared with professional communities via presentations and publications. The curation of these items is not only important for the quality of project deliverables, but also for the improvement of professional data services and shared expertise.

Delivery and Analytics

Clean deidentified data for finalized cases (complete or partial) are delivered to the core science team on a quarterly basis or as requested. We share data via an institutional storage option so that the data can be used in multiple analyses and papers. The data set consists of multiple files that follow the structure of the survey: each section of the survey corresponds to one file. Each file contains a column with respondent ID for easy merging and transformations. Data files are provided in three formats, SPSS, Stata, and csv, to accommodate varying software preferences and analytical skills of the core science team.

In addition to the data files, the delivery includes field interviewer comments, audio recording files, and case report forms. To provide context for the dataset, each delivery also includes a data processing document that describes all editing, coding, and data handling decisions as well as any modifications to the processes that happened during the last quarter. The documentation folder also includes descriptive statistics for each main survey file to assist in interpreting and understanding the data.

Finally, we have created a codebook that is updated with every delivery and shared with the core science team. The codebook describes the number of observations, number of variables in each module, missing value codes, and variable names, types, labels, and values and their ranges and can be used in future data re-use or for sharing data with external collaborators (see Figure 7).

In addition to the delivery of data to the core science team, deidentified data is available for sharing with outside researchers. Currently, the data is available only upon request and requires approval of the core science team PI. Upon such an approval, which also includes filling out a data use agreement, the data team adds the approved researchers to the institutional storage folder that contains deidentified data. Availability upon request will continue to be the main option for data sharing until the data collection finishes and the core science team answers the research questions that were part of the initial study design (see also the discussion on sharing and preservation in the "Discussion and Recommendations" section).

Module: Demographics

Number of observations: 1680

Number of variables in this module: 56

Missing value codes (unless specified otherwise):
94 NOT ASKED DUE TO CHANGE IN QUESTIONNAIRE DURING FIELD PERIOD
95 NOT ASKED DUE TO BREAK-OFF
96 NOT ASKED DUE TO SKIP LOGIC
97 REFUSED
98 DON'T KNOW

Label	Variable	Type	Value Labels / Range
Case ID from NorcSuite	SU_ID	numeric	8 digits
Demographics Module - Start Time	DEMOGRAPHICS_STARTTIME	date	dd-mmm-yyyy hh:mm:ss
What is your date of birth?	DOB	character	8 characters
What is your date of birth? YYYY/MM/DD FORMAT	DOB_DATE	date	yyyy/mm/dd
What are the last four digits of your social security number?	FOURDIGIT_SSN	character	4 digits
What is your current sex or gender?	GENDER_ID	numeric	1 MALE 2 FEMALE 3 TRANSGENDER 4 NON-BINARY/GENDER FLUID 5 GENDERQUEER 6 MALE TO FEMALE 7 FEMALE TO MALE 8 INTERSEX 9 ANOTHER IDENTITY, NOT LISTED, SPECIFY

Figure 7: Survey codebook excerpt.

Active curation of all data assets enables not only consistent and timely quarterly deliveries, but it also allows the data team to respond to analytical requests that serve outreach or exploratory purposes or provide information for developing future research hypotheses and proposals. For example, the outreach team requested the latest findings about substance use and the associated social stigma. Or, the core science team inquired about the rates of obesity in rural and urban counties in Indiana, USA. To generate a report for this request, we identified cases that had both height and weight measurements available, calculated the Body Mass Index (BMI) using the CDC formula for the Metric System,[6] and assigned respondents to categories of obesity based on the CDC definitions.[7]

6 Calculating BMI Using the Metric System, available at https://www.cdc.gov/nccdphp/dnpao/growthcharts/training/bmiage/page5_1.html

7 Defining Adult Overweight and Obesity, available at https://www.cdc.gov/obesity/adult/defining.html

Another variable was constructed to assign state counties to categories within the urban-rural continuum. The classification used the 2013 US Census statistical area maps and a classification of counties or county equivalents into metropolitan, micropolitan, and noncore as defined by the Office of Management and Budget (OMB).[8] Once the appropriate dataset was constructed, we create a report with relevant statistics and visualizations and the corresponding dataset.

As can be seen from the examples above, working on such analytical requests create additional intermediary or derived data products that become part of the project. The coding and cleaning procedures, as well as additional sources of information require documentation and management so that they become integrated into the database and the project assets. Some of it, for example, the calculated variables, are added to the survey database and enrich it for future use. Others, such as the cleaning and processing scripts or a review of urban-rural classification schemas, become part of the curated archive of research objects that can be used by others.

Discussion and Recommendations

Active curation is a laborious process that goes beyond the application of curation activities to the final data products. Within an academic institution the generation of such final products is often shared between the researcher, units with expertise in data collection and analysis, and the units that are responsible for data sharing and long-term preservation. In addition to internal collaborators, data production also relies on third party vendors and organizations that provide tools, instruments, equipment, or additional data.

Despite the existing recommendations for best practices that try to bring order into curation during all stages of research, our experience indicates that working with live data objects continues to be the "Wild West" experience (Plale and Kouper 2017). Active data work happens in a space where organization and descriptions keep changing, the work exists on multiple devices and computers, using multiple software with an ever-increasing number of files that have inconsistent names. While lifecycle thinking helps to break down some of the processes of working with data, as soon as we started working on this complex project, the stage sequencing fell apart and we realized that the stages do not neatly follow one after the other. In our work we rely on a variety of guidelines, both from the industry and academia, and most of our activities are cross-cutting rather than sequential (Faundeen et al. 2013).

As our case analysis demonstrates, active curation of a large longitudinal survey involves researchers (the core science team) delegating many curation responsibilities to other units and organizations, including the survey research center (the data team), NORC (the sampling vendor), biobank, and others. The researchers rely on the expertise of those units in generating and handling the

8 US Census Statistical Area Classification https://www.census.gov/geographies/reference-maps/2013/geo/statecbsa.html

data. The curation activities are performed during survey development, data collection, data management, and delivery and analytics. They include multiple iterations of working with various processes, tools, and live data assets, integrating them into the main survey database, and assisting in the production of final data products (see Figure 8).

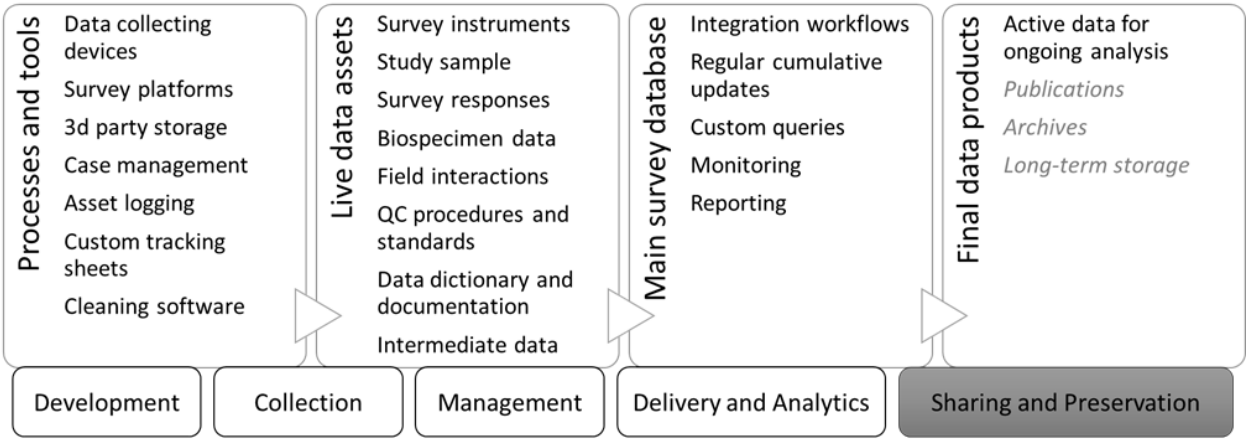

Figure 8: Active curation of the survey.

As the diagram above illustrates, the earlier stages of research and data collection are characterized by the multiplicity of tools and data assets that need curation. We worked with various software and devices and, in addition to survey responses and biosamples, processed many assets, including the survey instrument, field operations data, quality procedures and standards, and data documentation. Every software-device-data asset connection required making decisions about a) what to keep and what to discard, b) how to document the asset and its relationship to other assets, c) how to balance cost and usability, and d) how to facilitate access to needed assets by stakeholders within the designated timeframe. Despite our best effort and decades of experience, the decisions sometimes remained ad hoc as there were no accepted best practices or standards for active curation.

One of the main achievements of the team was the development of an integrated database that simplifies tracking of the data assets and integration between several tools and software. We developed automated workflows to transfer data from server to server and from vendor to vendor. The standardized structure of the database and the integration of all information into it allowed to support many activities and combine information across assets. Bringing all the data into the database also let us to maintain integrity and quality of the data, quickly adapt to changes, and automate interviewer and respondent reminders, data aggregation, and other tasks.

The diagram above also illustrates a disconnect between active curation and what can be called traditional curation, or curation at the end of the data lifecycle in an

academic setting. During the active phase of the study, our work ends with the delivery of data for ongoing analysis by a core science team; that is why the stage of "Sharing and Preservation" and the associated final data products that support publications, archival, and long-term storage are greyed out on the diagram. The survey research center has its own policies of storage and retention of research data that apply to all data the center generates or curates, but these policies are not coordinated with other units on campus involved in data work.

The project does not yet have established procedures for where, when and for how long the data will be archived after it moves out of the delivery and analytics stage. The complexities of the survey could not be captured in advance in a data management plan. As the survey continues, so does the discussion about the best ways to preserve the data. The plans include depositing data into an archive that specializes in health data, such as the Regenstrief Institute, preserving it in an institutional repository, or creating a dedicated virtual enclave to support ongoing research. The goals of supporting ongoing research and archiving paper copies of data assets are beyond the scope of many institutional repositories or even data archives, including our university library.

Our university library is working on expanding its capacity to accept datasets, but currently, its data repository is at a pilot stage. While a deidentified archival version of our data can be deposited there, preservation of the rich integrated database that was developed for the project will most likely remain the responsibility of the survey research center. The library offers consultations on how to curate data, but there are no services yet that address active curation and preservation of such databases, especially, with privacy and security controls.
Currently there are no accepted practices of how to connect active and archival data products and no workflows that support the transition from active to the final stages of research products. We hope our paper can be a first step in stimulating a discussion about it between academic units that share responsibilities of taking care of data at various stages of the lifecycle, including campus data centers, survey research centers, libraries, and others.

Active curation is also a battle between craftsmanship and economics. On one hand, data curators want to do a thorough job and address every aspect of taking care of people, products, and processes around research, but at the same there is always a consideration of cost. While it is tempting to promise the "golden standard" of curation no matter what, the reality of working on research projects always includes a trade-off between quality, cost, and time. There is also a delicate balance between maintaining the quality of products, which is visible to the user, and the quality of processes, which is not visible to the user and, as a result, is subject to less scrutiny. And yet, ignoring internal quality is short-sighted because external and internal quality are connected. In the long run poor internal quality affects the external quality. Given these challenges, researchers and data curators can use our case study as a reference for recommendations on best practices in data management and data curation, and plan accordingly within their budget and resource allocations.

A longitudinal survey of this complexity requires dedicated staff and resources to implement, monitor, and curate. However, we believe that our case study provides valuable lessons even for projects of smaller sizes or with fewer resources available. As our case study illustrates, active data curation requires a broad range of skills and technologies. In larger projects different individuals contribute different skills and form a larger team with broad expertise. Larger teams can also afford to purchase and maintain multiple tools and services, both specific and generalized. Smaller projects would have to look for individuals who combine various skills and expertise or rely on collaborations with other units. They would also spend more time planning and selecting multifunctional tools and technologies.

Our recommendations below can also be scaled up or down depending on the size of the project. In addition to providing some practical guidance, they offer a future research agenda in data curation.

Table 1: Recommendations and considerations in building support for active curation.

Recommendation	***Benefits***	***Possible considerations***
Develop a consistent approach to working with active or "live" data	• Support for new studies and new products • Increased efficiency in training and management • Integrated assets for consistent logging, organization, and documentation • Effective time management • A library of tools for re-use	No one solution will support all active curation needs, therefore often this will require technical expertise in programming and databases or review and evaluation of various existing solutions and decisions about how to combine them.
Design curation for current and future data work	• Alignment of the current project goals with the larger goals of the unit and the research agenda • A consistent approach regarding original and derived data based on the research questions and the need for reuse and reproducibility • Streamlined new studies design	A uniform data integration and aggregation platform, such as the database, makes curation modular and easy to adapt to the future needs. It allows to separate survey programming from storage and subsequent processing and to keep track of new and derived variables.
Consider working with humans as part of curation	• Dedicated team that includes data specialists, but also interviewers, supervisors, and IT personnel • A robust curation ecosystem • Improved sample curation and work with all the stakeholders • Increased support for longitudinal efforts	The sample of participants needs active curation too. A tracing / tracking system is a "live" system that is constantly updated as more information is obtained.
Standardize time and cost tracking	• Sufficient time for curation activities • Resource allocation and its measures of efficient use become part of curation as basic requirements • Improved information about labor and costs	Developing systems to track how much time and other resources curation activities require can be time consuming, it needs to fit with the larger university structures.
Develop and adopt standards for active curation	• Integration of active curation into the research lifecycle and curation activities • Optimized work of data-generating organizations • Better alignment between the goals of research and preservation and organizational cultural and technical resources	Each of the units involved in data work will need policies and standards about data storage, access, retention, and destruction, but those policies will need to be coordinated with each other.

The last recommendation to develop and adopt standards for active curation applies at multiple levels, including academic and service-oriented units, such as survey centers, departments and schools, libraries and archives, but also larger entities, such as the domains of social sciences and their professional organizations. Some professional organizations have recognized that guidelines for data curation and management vary among professions, institutions, organizations, or even research groups. They began to provide guidelines for researchers, often organized in the form of questions that researchers need to answer (ICPSR 2012; Kalichman 2016). These guidelines need to be expanded to address the messiness of active curation and the shared nature of data stewardship responsibilities. Close cooperation and coordination between academic units with data responsibilities will help to develop consistent institutional and professional frameworks.

Conclusion

Despite the ongoing research into the concepts and practices of curation, academic institutions are still facing many challenges in supporting the growing needs of research data work. The challenges include the changing roles of experts to address dynamic and complex data problems, the multiplicity of tools that vendors suggest will solve all the problems, and the lack of communication and collaboration across units that are involved in data production and curation. Curating data for integration and interdisciplinary use is another long-term challenge that only a few big data centers have begun to address.

In this paper we discussed the challenges of generating data for a large longitudinal survey and argued that these challenges are better addressed through active curation. Our efforts illustrate a gap in the existing discussions on curation. On one hand, there is an acknowledged need for active or upstream curation as an engagement of curators close to the point of data creation. On the other hand, the recommendations on how to do it and technologies that support that are scattered across multiple domain-oriented data efforts and projects.

Our paper proposes a broader view on active curation that focuses on the curation of live data objects and includes the curation of people, data and instruments, code, derived products, and all materials and procedures. This broader view expands the current understanding of data and what is being curated as part of adding value to research and making the products fit for future purposes. Viewed as all information that provides support for generating insights, data includes such information as sampling probabilities, questionnaire development procedures, training materials, metadata, paradata, auxiliary data as well as the cost and decision-making that surrounds this project.

We also propose to view active curation in the context of data work within academic institutions and conceptualize curation along the stages of development, collection, management, and delivery of data. We described how these stages often occur concurrently and require decisions with regard to tools and software,

multiple data assets, human resources, and integration workflows. As academic institutions expand their data services, they need to think about models of collaboration and division of labor between various units that provide those services. Aligning various units along the notions of upstream and downstream curation is one such model, designing a concierge service to field various data requests is another one (Collura et al. 2019).

Acknowledgements

The project is supported by the Indiana University Grand Challenge Precision Health Initiative through the *Person to Person Health Interview Study (P2P)*. The *P2P* was supported by an award to the Indiana Clinical and Translational Science Institute's Precision Health Initiative from Indiana University Grand Challenges and through a grant from the National Institutes of Health (Award Number UL1TR002529). The content of this publication is solely the responsibility of the authors and does not necessarily represent the official views of the Principal Investigators of the *P2P* or of the *Precision Health Initiative.*

Data Availability

The paper documents a case study of curating an active dataset. The data and related products are under embargo until the end of the project.

References

Abbott, Daisy. 2008. "What Is Digital Curation?" Edinburgh: Digital Curation Centre. http://www.dcc.ac.uk/resources/briefing-papers/introduction-curation/what-digital-curation

Adaptive Testing Technologies. n.d. "CAT-MH Modules." *Adaptive Testing Technologies (blog)*. Accessed March 31, 2021. https://adaptivetestingtechnologies.com/cat-mh-modules

Akers, Katherine G., Fe C. Sferdean, Natsuko H. Nicholls, and Jennifer A. Green. 2014. "Building Support for Research Data Management: Biographies of Eight Research Universities." International *Journal of Digital Curation* 9(2): 171–191. https://doi.org/10.2218/ijdc.v9i2.327

Akmon, Dharma, Inna Kouper, Margaret L. Hedstrom, James D Myers, and Anna Ovchinnikova. 2017. "Building Tools to Support Active Curation: Lessons Learned from SEAD." In *International Digital Curation Conference IDCC-17*, 20-23 February. Edinburgh.

Baker, Karen, and Lynn Yarmey. 2009. "Data Stewardship: Environmental Data Curation and a Web of Repositories." *The International Journal of Digital Curation* 4(2). https://tsc.library.ubc.ca/index.php/christineslais/article/view/44

Ball, Alex. 2012. "Review of Data Management Lifecycle Models." http://opus.bath.ac.uk/28587/1/redm1rep120110ab10.pdf

Beheshti, Amin, Kushal Vaghani, Boualem Benatallah, and Alireza Tabebordbar. 2018. "CrowdCorrect: A Curation Pipeline for Social Data Cleansing and Curation." In *Information Systems in the Big Data Era, edited by Jan Mendling and Haralambos Mouratidis*, 24–38. Lecture Notes in Business Information Processing. Springer International Publishing. https://doi.org/10.1007/978-3-319-92901-9_3

Beullens, Koen, Geert Loosveldt, Caroline Vandenplas, and Ineke Stoop. 2018. "Response Rates in the European Social Survey: Increasing, Decreasing, or a Matter of Fieldwork Efforts?" *Survey Methods: Insights from the Field (SMIF)*. https://doi.org/10.13094/SMIF-2018-00003

Bracke, Marianne Stowell. 2011. "Emerging Data Curation Roles for Librarians: A Case Study of Agricultural Data." *Journal of Agricultural & Food Information* 12(1): 65–74. https://doi.org/10.1080/10496505.2011.539158

Brick, J. Michael. 2011. "The Future of Survey Sampling." *Public Opinion Quarterly* 75(5): 872–888. https://doi.org/10.1093/poq/nfr045

Chu, Xu, Ihab F. Ilyas, Sanjay Krishnan, and Jiannan Wang. 2016. "Data Cleaning: Overview and Emerging Challenges." In *Proceedings of the 2016 International Conference on Management of Data*, 2201–2206. SIGMOD '16. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/2882903.2912574

Collura, Michael A., Virginia A Dressler, Michael Hawkins, and Michael Kavulic. 2019. "Served on a Silver Platter Working Towards an Academic Research Data Concierge Service." *DESIDOC Journal of Library & Information Technology* 39(06): 271–279. https://doi.org/10.14429/djlit.39.06.14774

Constantopoulos, Panos, Costis Dallas, Ion Androutsopoulos, Stavros Angelis, Antonios Deligiannakis, Dimitris Gavrilis, Yannis Kotidis, and Christos Papatheodorou. 2009. "DCC&U: An Extended Digital Curation Lifecycle Model." *International Journal of Digital Curation* 4(1): 34–45. https://doi.org/10.2218/ijdc.v4i1.76

Couper, M. P. 1998. "Measuring Survey Quality in a CASIC Environment - WebSM." In *Proceedings of the Survey Research Methods Section*, American Statistical Association, 41–44. http://www.websm.org/db/12/336/Bibliography/Measuring_Survey_Quality_in_a_CASIC_Environment/?menu=1&lst=&q=search_1_1_-1&qdb=12&qsort=1

Cox, Andrew M., Mary Anne Kennan, Liz Lyon, and Stephen Pinfield. 2017. "Developments in Research Data Management in Academic Libraries: Towards an Understanding of Research Data Service Maturity." *Journal of the Association for Information Science and Technology* 68(9): 2182–2200. https://doi.org/10.1002/asi.23781

Darch, Peter T., Ashley E. Sands, Christine L. Borgman, and Milena S. Golshan. 2020. "Library Cultures of Data Curation: Adventures in Astronomy." June. https://escholarship.org/uc/item/2kw90334

Davis, James A., and Tom W. Smith. 1992. The NORC General Social Survey. SAGE.
Downs, Robert R., and Robert S. Chen. 2010. "Designing Submission and Workflow Services for Preserving Interdisciplinary Scientific Data." *Earth Science Informatics* 3(1): 101–110. https://doi.org/10.1007/s12145-010-0051-6

Faundeen, J. L., T. E. Burley, J. A. Carlino, D.L. Govoni, H. S. Henkel, S. L. Holl, V. B. Hutchison, et al. 2013. "The United States Geological Survey Science Data Lifecycle Model." USGS (US Geological Survey). http://pubs.usgs.gov/of/2013/1265/pdf/of2013-1265.pdf

Giaretta, David. 2004. "DCC Approach to Digital Curation." UK: Digital Curation Centre. http://www.dcc.ac.uk/sites/default/files/documents/DCCApproachtoDigitalCuration-20040827.pdf

Goble, Carole, Robert Stevens, Duncan Hull, Katy Wolstencroft, and Rodrigo Lopez. 2008. "Data Curation + Process Curation=data Integration + Science." *Briefings in Bioinformatics* 9(6): 506–517. https://doi.org/10.1093/bib/bbn034

Gray, Jim, Alexander S. Szalay, Ani R. Thakar, Christopher Stoughton, and Jan vandenBerg. 2002. "Online Scientific Data Curation, Publication, and Archiving." In *Virtual Observatories*, 4846:103–107. International Society for Optics and Photonics. https://doi.org/10.1117/12.461524

Groves, Robert M., and Steven G. Heeringa. 2006. "Responsive Design for Household Surveys: Tools for Actively Controlling Survey Errors and Costs." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 169(3): 439–457. https://doi.org/10.1111/j.1467-985X.2006.00423.x

Hales, Brigette M., and Peter J. Pronovost. 2006. "The Checklist—a Tool for Error Management and Performance Improvement." *Journal of Critical Care* 21(3): 231–235. https://doi.org/10.1016/j.jcrc.2006.06.002

Higgins, Sarah. 2008. "The DCC Curation Lifecycle Model." *The International Journal of Digital Curation* 3(1): 135–140. https://doi.org/10.2218/ijdc.v3i1.48

Higgins, Sarah. 2011. "Digital Curation: The Emergence of a New Discipline." *International Journal of Digital Curation* 6(2): 78–88. https://doi.org/10.2218/ijdc.v6i2.191

International Field Directors and Technologies Conference (IFD&TC). "Challenges with Complex Studies." 2021. Presented at *International Field Directors and Technologies Conference (IFD&TC)*, Session 2C - Field. https://ifdtc.org/events/2c-field

Johns Hopkins Coronavirus Resource Center. n.d. "How to Use Our Data." Accessed March 24, 2021. https://coronavirus.jhu.edu/about/how-to-use-our-data

Johnston, Lisa R, Jacob Carlson, Cynthia Hudson-Vitale, Heidi Imker, Wendy Kozlowski, Robert Olendorf, and Claire Stewart. 2018. "How Important Is Data Curation? Gaps and Opportunities for Academic Libraries." *Journal of Librarianship and Scholarly Communication* 6(1). https://doi.org/10.7710/2162-3309.2198

Julkowska, Magdalena M., Stephanie Saade, Gaurav Agarwal, Ge Gao, Yveline Pailles, Mitchell Morton, Mariam Awlia, and Mark Tester. 2019. "MVApp—Multivariate Analysis Application for Streamlined Data Analysis and Curation." *Plant Physiology* 180(3): 1261–1276. https://doi.org/10.1104/pp.19.00235

Kammen, Welmoet Bok van, and Magda Stauthamer-Loeber. 1998. "Practical Aspects of Interview Data Collection and Data Management." In *Handbook of Applied Social Research Methods*, edited by Leonard Bickman and Debra J. Rog. SAGE.

King, Gary. 2011. "Ensuring the Data-Rich Future of the Social Sciences." *Science* 331(6018): 719–721. https://doi.org/10.1126/science.1197872

Lavrakas, Paul J. 2008. *Encyclopedia of Survey Research Methods*. SAGE Publications.

Lee, Dong Joon, and Besiki Stvilia. 2017. "Practices of Research Data Curation in Institutional Repositories: A Qualitative View from Repository Staff." *PLOS ONE* 12(3): e0173987. https://doi.org/10.1371/journal.pone.0173987

Lord, Philip, Alison Macdonald, Liz Lyon, and David Giaretta. 2004. "From Data Deluge to Data Curation." In *Proceedings of the UK E-Science All Hands Meeting*. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.111.7425

Lord, Philip, and Alison Macdonald. 2003. "Data Curation for E-Science in the UK: An Audit to Establish Requirements for Future Curation and Provision." The Digital Archiving Consultancy.

Macdonald, Stuart, and Luis Martinez-Uribe. 2010. "Collaboration to Data Curation: Harnessing Institutional Expertise." *New Review of Academic Librarianship* 16(sup1): 4–16. https://doi.org/10.1080/13614533.2010.505823

Marenco, Luis, Nicholas Tosches, Chiquito Crasto, Gordon Shepherd, Perry L. Miller, and Prakash M. Nadkarni. 2003. "Achieving Evolvable Web-Database Bioscience Applications Using the EAV/CR Framework: Recent Advances." *Journal of the American Medical Informatics Association* 10(5): 444–453. https://doi.org/10.1197/jamia.M1303

Oliver, Gillian, and Ross Harvey. 2016. *Digital Curation*. American Library Association.

Plale, Beth, and Inna Kouper. 2017. "The Centrality of Data: Data Lifecycle and Data Pipelines." In *Data Analytics for Intelligent Transportation Systems*, edited by Mashrur A. Chowdhury, Amy Apon, and Kakan Dey, 91–111. Cambridge, MA: Elsevier Inc. https://doi.org/10.1016/B978-0-12-809715-1.00004-3

Project REDCap. n.d. "About – REDCap." Accessed March 24, 2021. https://projectredcap.org/about

Qin, Jian, Kevin Crowston, and Arden Kirkland. 2014. "A Capability Maturity Model for Research Data Management." Syracuse, NY: School of Information Studies, Syracuse University. https://surface.syr.edu/cgi/viewcontent.cgi?article=1191&context=istpub

RDA COVID-19 Working Group. 2020. "RDA COVID-19 Recommendations and Guidelines for Data Sharing." https://zenodo.org/record/3932953#.YFtc851KgmJ

Rice, Robin. 2009. "DISC-UK DataShare Project: Final Report." Programme/Project deposit. University of Edinburgh. https://repository.jisc.ac.uk/336

Scientific Data Curation Team. 2020. "Metadata Record for: Epidemiological Data from the COVID-19 Outbreak, Real-Time Case Information." figshare. https://springernature.figshare.com/articles/Metadata_record_for_Epidemiological_data_from_the_COVID-19_outbreak_real-time_case_information/11974344/1

Singleton, Royce A., Jr., and Bruce C. Straits. 2009. "Data Processing and Elementary Data Analysis." In *Approaches to Social Research*, 5th Edition. Oxford University Press.

Steinhart, Gail, John Saylor, Paul Albert, Kristine Alpi, Pam Baxter, Eli Brown, Kathy Chiang, et al. 2008. "Digital Research Data Curation: Overview of Issues, Current Activities, and Opportunities for the Cornell University Library." https://ecommons.cornell.edu/handle/1813/10903

Stinson, Linda, and Sylvia Kay Fisher. 1996. "Overview of Data Editing Procedures in Surveys." Swanberg, Stephanie M. 2017. "Inter-University Consortium for Political and Social Research (ICPSR)." *Journal of the Medical Library Association* 105(1): 106–107. https://doi.org/10.5195/jmla.2017.120

Substance Abuse and Mental Health Services Administration (SAMHSA). 2006. "Standards and Guidelines for Statistical Surveys." Office of Management and Budget. https://www.samhsa.gov/data/sites/default/files/standards_stat_surveys.pdf

Tammaro, Anna Maria, Krystyna Matusiak, Frank Andreas Sposito, Vittore Casarosa, and Ana Pervan. 2017. "Understanding Roles and Responsibilities of Data Curators: An International Perspective." *Libellarium: Journal for the Research of Writing, Books, and Cultural Heritage Institutions* 9(2). https://doi.org/10.15291/libellarium.v9i2.286

Tenopir, Carol, Ben Birch, and Suzie Allard. 2012. "Academic Libraries and Research Data Services: Current Practices and Plans for the Future." Association of College and Research Libraries. http://www.ala.org/acrl/sites/ala.org.acrl/files/content/publications/whitepapers/Tenopir_Birch_Allard.pdf

Tiessen, Jan, Claire Celia, Tom Ling, Helen Ridsdale, Maiwënn Bareaud, and Christian van Stolk. 2010. "Evaluation of DG SANCO Data Mamagement Practices." The RAND Corporation.

Vardigan, Mary, and Peter Granda. 2010. "Archiving, Documentation, and Dissemination." In *Handbook of Survey Research*, edited by James D. Wright and Peter V. Marsden. Emerald Group Publishing.

Wallis, Jullian C., Christine L. Borgman, Matthew S. Mayernik, and Alberto Pepe. 2008. "Moving Archival Practices Upstream: An Exploration of the Life Cycle of Ecological Sensing Data in Collaborative Field Research." *International Journal of Digital Curation* 3(1): 114–126. https://doi.org/10.2218/ijdc.v3i1.46

Wang, Xiaoming, Carolyn Williams, Zhen Hua Liu, and Joe Croghan. 2019. "Big Data Management Challenges in Health Research—a Literature Review." *Briefings in Bioinformatics* 20(1): 156–167. https://doi.org/10.1093/bib/bbx086

Weber, Nicholas M., Carole L. Palmer, and Tiffany C. Chao. 2012. "Current Trends and Future Directions in Data Curation Research and Education." *Journal of Web Librarianship* 6(4): 305–320. https://doi.org/10.1080/19322909.2012.730358

Welch, Matthew, Oliver Dupriez, Mahmood Asghar, Michael Sharp, and Scott Pontifex. 2019. "Open Source Solutions for Curation and Dissemination: Introducing New Multi-Data and Multi-Standard World Bank Tools and Schemas." Presented at *The IASSIST 2019: Data down under: Exploring "data firsts,"* Sydney, Australia. https://doi.org/10.5281/zenodo.3600562

Wright, James D., and Peter V. Marsden. 2010. "Survey Research and Social Science: History, Current Practice, and Future Prospects." In *Handbook of Survey Research*. Emerald Group Publishing.

Wynholds, Laura. 2011. "Linking to Scientific Data: Identity Problems of Unruly and Poorly Bounded Digital Objects." *International Journal of Digital Curation* 6(1): 214–225. https://doi.org/10.2218/ijdc.v6i1.183

Yakel, Elizabeth, Ixchel M. Faniel, and Zachary J. Maiorana. 2019. "Virtuous and Vicious Circles in the Data Life-Cycle." Text. University of Borås. http://www.informationr.net/ir/24-2/paper821.html

Yakel, Elizabeth. 2007. "Digital Curation." *OCLC Systems & Services* 23(4): 335–340. https://doi.org/10.1108/10650750710831466