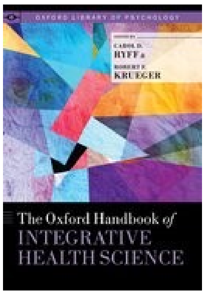
OXFORD LIBRARY OF PSYCHOLOGY
CAROL D. RYFF & ROBERT F. KRUEGER
The Oxford Handbook of INTEGRATIVE HEALTH SCIENCE

The Oxford Handbook of Integrative Health Science

Carol D. Ryff (ed.), Robert F. Krueger (ed.)

https://doi.org/10.1093/oxfordhb/9780190676384.001.0001

Published online: 09 October 2018 **Published in print:** 29 November 2018 **Online ISBN:** 9780190676407 **Print ISBN:** 9780190676384

Search in this book

CHAPTER

2 Behind the Scenes in Integrative Health Science: Understanding and Negotiating Data Management Challenges

Barry T. Radler, Gayle D. Love

https://doi.org/10.1093/oxfordhb/9780190676384.013.2 Pages 23–34

Published: 09 October 2018

Abstract

The complexities of managing the data collection and data products in multidisciplinary, population-based longitudinal research on health are not readily apparent to those who use publicly available data. This chapter explores the behind-the-scenes details involved in conducting the Midlife in the United States (MIDUS) study. The objective is to explicate how such a complex investigation involving *comprehensive biopsychosocial assessments, obtained on the same individuals followed repeatedly over time*, is carried out. The MIDUS study includes multiple data collection projects, covering different domains of science. Such extensive data collection must be carefully sequenced and distributed across different data collection sites, while recognizing that the longitudinal research data life cycle is iterative in nature. The overarching goals are to ensure the integrity of the data collection process while generating data products that are discoverable, understandable, and well documented to promote robust primary and secondary usage of publicly available datasets.

Keywords: Midlife in the United States, MIDUS, biopsychosocial assessment, data collection, longitudinal research, health, data management

Subject: Health Psychology, Psychology

Series: Oxford Library of Psychology

Collection: Oxford Handbooks Online

Introduction

This chapter explores the behind-the-scenes details involved in conducting longitudinal biopsychosocial health research while ensuring that data products are documented to be maximally useful. In doing so, it discusses best practices and lessons learned that account for the Midlife in the United States (MIDUS) study as a valued public resource. Such information is foundational to the larger objectives of integrative science and thus is likely to be beneficial to other similar research pursuits. To unpack the tasks involved, this chapter addresses three key data management challenges posed by integrative health research that uses a longitudinal design. The first pertains to the orchestration of multisite data collection and the maintenance of a viable long-term sample. The second addresses the creation of coherent cross-project and cross-time data and documentation products. The third examines the provision of research metadata and, in particular, of exploiting new technological metadata standards to produce study documentation that is web friendly and interoperable (exchanged and used across different computer systems or software).

Before getting into the three major topics of the chapter, we first examine background issues that inform how we approached the tasks involved. The first involves a distillation of key principles behind our approach to digital data stewardship, and the second underscores the need for data management best practices in managing the complexity of the MIDUS design, involving multiple domains of assessment on the same respondents moving through time. The central message is that MIDUS was designed to provide data products that are of ↳ interest and use to the scientific community, which increasingly is committed to advancing knowledge of health that cuts across wide disciplinary territories. This goal, in turn, has been advanced by the recognition and fulfillment of principles that address the notable new challenges in the collection and management of multidisciplinary data. p. 24

Background Issues

Principles of Digital Data Stewardship

The data management practices and principles of the MIDUS study are one of the keys to its popularity and success with the scientific community. Good data management is not an end in itself but rather a key conduit to knowledge discovery and innovation as well as to subsequent knowledge integration and reuse by the community after data publication. What then constitutes good data management? The FAIR guiding principles provide four desiderata behind effective digital data stewardship: producing research objects that are findable, accessible, interoperable, and reusable (Bechhofer, De Roure, Gamble, Goble, & Buchan, 2010; Wilkinson et al., 2016). For MIDUS, these digital research objects are data and metadata (information about research data): Findable data are described with rich metadata so they can be uniquely identified and systematically searched; accessible (meta)data are freely available using standard communications protocols (e.g., web); interoperable (meta)data are formatted with open or shared standards; and reusable (meta)data are richly documented to facilitate verification and reuse by machines and humans.

The overarching goals of data management in MIDUS have been to ensure the integrity of the complex data collection process and generate FAIR data products to facilitate active secondary usage of publicly available data. While the material in this chapter is often specific to MIDUS, the principles elucidated are broadly applicable to the conduct of integrative health research and can be valuable to the data analyst as well as the project manager and data collector.

The Complexity and Popularity of MIDUS

The MIDUS study is a multicohort, multidisciplinary longitudinal study of health and well-being of a national sample of US adults. As such, the MIDUS study design is inherently complex. The complexities of *doing* the actual research, however, are not readily apparent (and they do not need to be) to users of the publicly available data that MIDUS generates. In particular, the MIDUS study design poses challenges to managing both the data collection and the resultant data products. Several unique characteristics of MIDUS inform its approach to study and data management.

The first is the multidisciplinary nature of MIDUS. The baseline study brought together a diverse set of researchers from epidemiological, sociological, and psychological fields and subfields to collaborate on a national survey. The survey data were supplemented at baseline with projects measuring participant cognitive capacities and daily experiences of stress. At the first follow-up, new projects were added to obtain a wider array of cognitive assessments, comprehensive biomarkers, and brain-based assessments of emotion regulation. All of these assessments were organized as distinct data collection projects, distributed over an extended fielding period with different project leaders, co-investigators, and project staff managing them. MIDUS collected data from thousands of participants via five different discipline-specific projects (survey, daily diary, cognitive, biomarker, and neuroscience) that were administered at separate universities, laboratories, and clinics across the country. This situation required a coordinated approach to study and data management that orchestrated distinct but related components of the MIDUS research enterprise.

Second, as a longitudinal study, MIDUS had to manage its sample to maximize efficiency of data collection as well as maintain its long-term viability so that participants could be effectively followed up years later. MIDUS participants ranged in age from 25 to 74 at baseline. This wide range allowed for study of individuals as they journeyed into and out of middle age, and it required interacting with them over extended periods of time. Intervals between data collection waves averaged around 9 years, with the survey data collection period ranging between 7.8 and 10.5 years, thus highlighting the extended field period of the survey alone. Further, individuals who participated in subsequent projects at each wave could potentially be participating in a data collection fielding period lasting years. To both maximize participation and minimize attrition in this complex context, it was critical to establish and nurture a rapport with respondents that grew over decades. Throughout this lengthy research data life cycle, contact information had to be updated and carefully managed to maintain a viable longitudinal sample that showed continued willingness to participate in the MIDUS enterprise.

p. 25 With regard to use of the data, MIDUS has been well received by the scientific community: It is hugely popular, with over 38,000 unique public users (J. McNally, personal email communication, September 26, 2016). MIDUS research products are shared under an open data philosophy and are available online via the Inter-University Consortium for Political and Social Research (ICPSR) website, the largest digital archive of social, behavioral, and health science data in the world. While a chief reason for the popularity of MIDUS is the depth and breadth of its content, another reason for its popularity is that MIDUS has made it a primary goal to render its data and documentation widely accessible without being obtuse, confusing, or disorganized. Lucid, findable, and accessible data significantly bolster usage.

Challenge 1: Orchestrating Multidisciplinary, Multisite Data Collection and Maintaining a Large Sample Over Time

Successful management of a large national sample of US adults who would participate in multiple components of MIDUS required a set of organizing principles and a means of implementing them. These core principles were essential for coordinated sequencing of participants through multiple data collection projects. This required a centralized Administrative Database to organize, schedule, and implement complex data collection while also maintaining the long-term viability of the sample. This section describes the key aspects of managing MIDUS data collection: the guiding principles to distributed data collection across different projects; the design and use of the Administrative Database to facilitate that data collection; and the mechanisms by which participant information was actively managed during and between fielding periods.

Guiding Principles Behind Multiproject Data Collection

Following successful participation in the MIDUS survey project, each respondent became available for recruitment to other data collection projects. The overall goal was to facilitate maximum participation in and across individual projects during an overlapping (shared) fielding period. To achieve these potentially conflicting objectives while avoiding the confusion such a situation presented, participation in MIDUS projects had to be carefully sequenced and scheduled to minimize respondent fatigue, ensure consistent communication to respondents, and maximize response rates. MIDUS managed the sequencing process by implementing three guidelines to optimize success for individual projects while maximizing participation across multiple projects:

1. *Participants could be recruited by only one project at a time*. This minimized confusion and burden among respondents and project staff.

2. *A minimum of 2 months must pass before a subject could be reassigned to the next project in their sequence* after completing (or declining) participation in a prior project. This sought to avoid overwhelming respondents with too many data collection requests in a short period of time.

3. During the fielding period, *all projects needed to have sufficient cases to recruit from* to complete data collection goals within the proposed timelines. Implementation of the first two guidelines was largely driven by the third (i.e., knowledge of what was needed to ensure that all sites had sufficient cases from which to effectively and efficiently recruit).

Purpose and Content of the Administrative Database

Balancing these three principles necessitated a system to store and organize data about project eligibility and assignment, subsample membership, and participation history. It also required knowledge of overall design issues related to multiple data collection projects with specific characteristics that influenced the pace of data collection. Together, this information determined the sequence through which respondents were recruited for each project. With the exception of the survey (which was always completed first), the order of participation could vary.

The Administrative Database helped “proceduralize” and automate the successful implementation of these three guidelines across MIDUS projects. Relational database systems like Access or MySQL are ideally suited to manage the various criteria used in participant sequencing. The database format facilitated creation of multiple modules that afforded MIDUS flexibility in creating graphic interfaces that were customized to different types of users (data collection staff, administrative staff, etc.) and provided the capacity to lock certain fields so they could only be modified by authorized staff. Importantly for MIDUS, such a database system allowed the creation of customized versions of the database that were distributed to each data collection site.

p. 26 The MIDUS Administrative Database included key types of information to provide accurate and timely contact information about MIDUS respondents as well as information about their eligibility for, and unfolding participation in, the different projects:

1. *Basic administrative data* included the respondent identifier along with subsample and family identifiers (for twins and siblings; see Chapter 1 for details on subsamples and waves). The respondent identifier was the primary key linking all respondent information in the database as well as their research data. The sample identifier indicated subsample membership (e.g., national random-digit-dialing [RDD] sample, twins, siblings, Milwaukee, Refresher). The family identifiers were helpful in the recruitment and scheduling process (e.g., scheduling twin pairs' data collection together, tracking "lost" participants through family members, etc.). Other key fields included birthdate, decedent status, address validity, and eligibility for individual MIDUS projects.

2. *Current and historic contact information* included full name(s), addresses, phone numbers, and email and was critical for facilitating ongoing communication with participants, confirming respondent identity, and performing tracking. Given that contact information can change over time, historic contact information was automatically archived.

3. *Details about current and past participation* in various MIDUS projects informed decisions about participation in the current wave of data collection, guided decisions about eligibility for other projects, and determined when a given participant was made eligible for the next project. Among the project-specific information included in participation data were final disposition codes (complete, noncomplete, refusal, noncontact, etc.) and completion dates. Project participation was regularly monitored so cases that completed or declined a given project could be reassigned to their next project in a timely way.

Actively Managing Participation and Sequencing

Along with the administrative, contact, and participation information contained in the Administrative Database, eligibility and sequencing decisions had to take into account other features of individual projects as well as the study design as a whole. This information was not necessarily contained in the database.

Each project had a unique set of objectives that relied on particular modes of data collection (e.g., cognitive, biomarkers, neuroscience). Determining how many cases were sufficient for a given project's recruitment required understanding how differences in project objectives impacted the pace of data collection. Phone or mail-based projects (i.e. survey, cognitive, daily diary) could complete a large number of assessments in a short period of time. Projects that required in-person data collection (i.e., the biomarker and neuroscience projects) could take several hours or days to complete the assessments. Such projects proceeded at a much slower pace because fewer participants could complete the assessments in a given period of time. Understanding the mode and pace of data collection helped determine the distribution and scheduling of cases to specific projects.

Understanding the target sample had implications for prioritizing the order in which participants were sequenced from one project to the next. While all participants in the recruitment pool were important to the study, some were more valuable to certain projects than others. For example, participants who had already completed two waves of data collection for a given project were likely more valuable (from a research standpoint) than those who previously completed just one wave, and individuals who had completed more than one project were also more valuable to ***all*** the projects. Such characteristics were used to differentially prioritize, schedule, and assign cases.

The net effect of balancing these considerations translated to different sequences of cross-project participation. The combination of varying sequences and “lag times” between project completions, and indeed between waves of data collection, not only required a flexible database system to implement, but also involved analytic considerations. For some research questions, this study design had implications for exploring causality in that some assessments could be tested as precursors to other assessments that came later in time. Thus, participation dates (month and year) were included in publicly released datasets so that such sequencing and lag effects could be accounted for in relevant analyses.

Longitudinal Sample Maintenance

p. 27 Ongoing longitudinal sample maintenance was accomplished during and between data ↳ collection events by continued communication with respondents, such as sending them birthday cards and newsletters. Both birthday cards and newsletters generated new contact and status information updates from respondents, next of kin, and the US Post Office. These efforts allowed MIDUS to maintain an up-to-date database of contact and status information, but they also rewarded participants for their cooperation by sharing research findings and keeping them aware of ongoing or upcoming data collection projects.

An important side effect of these maintenance efforts was the receipt of information on participant deaths by next of kin or other household informants. Other more systematic methods augmented this decedent information, including queries of the National Death Index, a centralized database of death record information administered by the National Center for Health Statistics. The advent of the Internet has also provided a wealth of public and private online resources that search legal and public records data, online obituaries, and genealogical websites for decedent information. All of these sources are used to obtain accurate information about participant decedent status and date of death, which is released to the public along with other MIDUS research datasets. These mortality data are profoundly important to aging research, but also serve the practical purpose of maintaining a current and efficient longitudinal sample.

Challenge 2: Creating User-Friendly Documentation and Data for the Scientific Community

One way in which MIDUS made its research products easier to navigate and use was by adopting conventions that specified how research materials, files, and data were organized. Specifying conventions for file naming, variable labeling, missing value specifications, and data formats provided project managers an organizational template that facilitated coherent and accurate data collection as well as efficient communication. These practices also made the diverse projects and products (summarized in Table 2.1) cohere to a MIDUS “brand,” giving the study a common look and feel while helping secondary users understand the various research outputs that comprise the study. Table 2.1 summarizes five projects that comprised the MIDUS study and illustrates the deeply multidisciplinary nature of the enterprise by describing the types of data and modes of assessment for each project. This diversity, in part, drove the need for conventions to tie the various projects’ data and documentation together.

Table 2.1 Description of MIDUS Data Collection Projects

Project Description	Instrument or Protocol
Survey	40-minute computer-assisted telephone interview CATI) followed by two mailed 50-page booklets (self-administered questionnaires); wide array of psychological constructs, demographic characteristics, and extensive mental and physical health measures
Daily diary inventory	8 days of daily experiences obtained via CATI interviews and 4 days' worth of salivary cortisol measurements
Cognitive assessments	20-minute CATI interview with seven cognitive tasks, including word recall, modified Stroop test, reaction times, etc.
Biomarkers	2-day clinic visit consisting of neuroendocrine, cardiovascular, immune, and bone biomarkers; physical exam; medical history; current and historical medication use; sleep assessments (subjective and actigraphy); and a psychophysiological challenge experiment
Neuroscience	Affective reactivity and recovery paradigm consisting of baseline and task-related electroencephalography (EEG), task-related electromyography (EMG), structural and functional magnetic resonance imaging (MRI), as well as self-administered questionnaires

At a more technical level, the uniform standards and conventions adopted by MIDUS have supported the efficient and accurate combination of its various data sources. MIDUS produces multiple project-specific datasets within each longitudinal wave that are *designed to be combined* to facilitate the analysis of integrative research questions.

File Types

p. 28 MIDUS conceives of two general file types that are generated by each data collection project. ↳ MIDUS organizes its local folders and subfolder structures based on this conceptualization, and ICPSR follows a similar file-type organization:

1. Documentation—Any information that explains the methodology or procedures of data collection, processing, and transformation is contained in documentation files. These files can and should include copies of data collection instruments.
2. Data—Data files contain quantitative and qualitative information collected from or obtained about study participants via protocols described in study documentation. These data are recorded in a standardized structured format that lends itself to analysis usually performed with specialized statistical software.

File-Naming Conventions

File-naming conventions are a critical data management function that helps manage and organize the materials produced by different MIDUS projects. These conventions become increasingly valuable as a study like MIDUS becomes more complex by the addition of waves, samples, and projects. The file and formatting specifications described in the material that follows are made available thru ICPSR as a piece of documentation in their own right.

The MIDUS file names are designed to convey numerous pieces of information in abbreviated form, including the sample, wave, project, file type, and date (version). MIDUS also employs shorthand and abbreviations to refer to different waves and projects; these are easily incorporated into file names. Two examples of documentation file types follow:

> **Examples: Documentation/Instruments**
>
> **MIDUS 1 Survey project: M1_P1_PHONE INSTRUMENT_5-8-12**
>
> **MIDUS 2 Cognitive project: M2_P3_READMEFIRST_20100819**

Data file names also include information on the number of cases (rows) in the dataset. Generally, the more specific the information included in a file name, the easier it is to understand its contents at a glance.[1]

> **Examples: Datasets**
>
> **MIDUS 2 Neuroscience project: M2_P5_DATA_N331_11-18-10**
>
> **MIDUS 3 Survey project: M3_P1_DATA_N5000_20120508**

Providing Sufficient Documentation for Projects

The multidisciplinary nature of MIDUS requires a variety of complex and often technical data collection protocols. To make the data produced by these different procedures independently understandable to secondary users, sufficient documentation of the data collection procedures must be associated with each dataset.

A README file is mandatory documentation for each MIDUS project; the file serves as introduction to the project and lists all the files available to the user, including datasets, instruments, and other documentation. It provides further information about the structure of the dataset (wide vs. long); its format (SPSS, SAS, etc.); and the number of variables (columns) and cases (rows).

The next most important piece of the documentation is a copy of the instrument or data collection tool itself. For survey projects, the instrument includes full question text, response options, skip directions, and interviewer instructions or prompts. But, for projects entailing nonsurvey data capture, the "instrument" may consist of descriptions of clinical or laboratory procedures, experimental protocols or conditions, and (in the context of physical measurements) observations, readings, and instrument settings. All of this information should be sufficiently recorded so that the data produced by these protocols can be appropriately understood and analyzed.

MIDUS encourages the publication of any information that will make MIDUS data independently understandable to secondary users. Each project must make clear the number of cases in a dataset, their completeness, and the origin of those cases. For the survey project, a final field report (which includes information on fielding, sample design, response rates, and weighting) is critical to understand the characteristics and representativeness of the core sample, on which other projects depend for a participant recruitment pool.

MIDUS also produces many constructed, derived, and transformed variables. These might be summary scale scores, such as the Center for Epidemiological Studies—Depression (CES-D) scale (Radloff, 1977); derived variables, such as respondent age, which is calculated with birthdates and interview dates; or standard code lists, such as the US Department of Labor's Dictionary of Occupational Titles. The ↳ provenance of such variables must be made clear by detailing the formulas, procedures, source, criteria, and so on used in producing them.

Finally, a data dictionary or code book must accompany each dataset. A code book is a document containing not only the code lists and response options used in the research, but also a detailed description of each variable contained in a dataset. A code book summarizes all of the metadata associated with and represented by each piece of raw data in a data file. The critical importance of metadata and code books to such integrative health datasets is discussed more fully in a further section.

Producing User-Friendly Datasets

All MIDUS projects produce SPSS-formatted datasets (.sav files) that are reviewed for adherence to MIDUS conventions. After a period of internal review (called the “shakedown cruise” after the term in which a ship’s performance is tested before entering service), these datasets are then submitted to ICPSR, which in turn produces datasets in a variety of proprietary (SPSS, SAS, Stata, Excel) and open (ASCII, CSV) formats. With some exceptions, these MIDUS datasets are flat rectangular files comprising numeric-formatted variables.

Regardless of project, every MIDUS dataset includes a core set of common administrative and demographic variables, such as participant identifier, interview date, age, sex, and so on. These variables not only are fundamental for many analyses but also serve as quality control checks for dataset integrity and facilitate accurate merges among multiple MIDUS datasets.

Variable Naming and Labeling.

MIDUS uses a consistent and simple variable-naming scheme that conveys important variable information at a glance, such as distinguishing the longitudinal wave, project, and source of the variable. Variable-naming conventions also make cleaning and programming new variables more efficient (old code or syntax can easily be repurposed) and ensure compatibility across different software programs (which can have different limitations on variable names).[2]

Figure 2.1 shows examples of two MIDUS variable names, which include information on wave, sample, project, and question number or assessment.

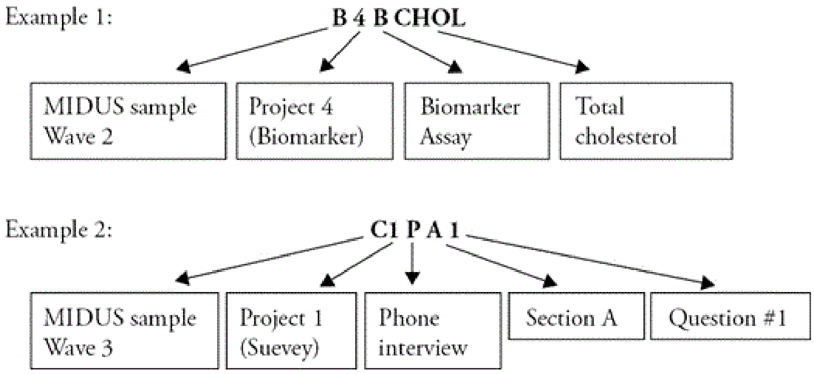

Figure 2.1 MIDUS variable naming examples.

Variable Formats.

MIDUS datasets utilize numeric variable formats whenever possible. The inclusion of string variables is discouraged, and open-ended responses, text, and verbatim data are converted into categories or codes where possible. Variable formats are precise: Variable lengths should not exceed the maximum number of digits possible for a code, value, or response, and decimals should not exceed three places due to an ICPSR data standard. Time and date variables require special consideration. Because of conflicting or proprietary formatting that hinders their interoperability among different software, date and time data are recorded as separate numeric variables. For example, date information in MIDUS is recorded in separate month and year variables, and temporal information is recorded in separate hour, minute, second, and meridian (am/pm, etc.) variables. One alternative for temporal variables is using 24-hour clock or military time, in which hours and minutes can be represented as a six-digit numeric variable (HHMMSS).

Variable Labels.

Additional description of each variable is included in its label. Variable labels provide a more detailed verbal description of the variable, and new technical metadata standards can take advantage of the text contained in a label to identify, search, and understand specific variables.

Value Labeling.

MIDUS value labels (within a variable) are displayed in UPPERCASE and can describe response options and p. 30 categorical or nominal data types. There are no character limits on ↳ the value labels. MIDUS encourages the standard coding and labeling of common variables, such as dichotomous indicator variables (YES=1, NO=2). Such data are common enough that standardizing them across MIDUS datasets ensures coding consistency and the ease of combining variables.

Missing Values.

MIDUS has coding conventions to indicate different types of item nonresponse. Such “missing” values are used to populate otherwise empty cells within a variable. The presence of empty cells in a variable introduces ambiguity regarding their “missing-ness.” By descriptively labeling specific values and designating them as “missing,” the researcher is provided more information about their status, and software programs can then ignore these values during statistical and mathematical computation. There are three general categories of missing values in MIDUS: DON’T KNOW (used only when there is an explicit DK response option), REFUSED/MISSING (used when valid response is not provided or chosen), and INAPPLICABLE (indicating an intentional skipping pattern). Some MIDUS projects include multiple data collection instruments and require codes for missing instruments, while other projects include specific filter codes or variables that indicate invalid or incomplete data.

Documenting Data Management With Code.

One of the most important principles of good data management is the ability to reproduce any research output (King, 1995). This principle is greatly advanced through the use of programming code wherever possible. Programming code (also known as syntax) contains all the commands that instruct statistical software to execute the steps involved in cleaning, transforming, analyzing, and managing datasets (Ball & Medeiros, 2011). Documenting all data work with programming code—as opposed to using the point-and-click drop-down menus in software programs—ensures that data processing can be replicated, modified, and reviewed. This is even more important in a longitudinal context where programming code can reinforce consistency, minimize errors, and facilitate corrections to data collection, management, and analytic processes across multiple waves of data.

The use of programming code also ensures the provenance of data as they are cleaned and transformed. In this way, the original raw data file can be preserved unchanged, while any subsequent alterations are performed on copies of the original data using programming or syntax. Doing so produces an audit trail of the work performed. Code files can provide rich documentation of procedures by "commenting out" work, that is, inserting human-readable explanatory text into the code that is ignored by the software but makes it easier for people to understand.

Using Repositories and Archives

In recent years, several national scientific organizations have begun requiring the studies they fund to have a data management plan in place that specifies how research products will be archived and shared (ICPSR, 2012). The National Institutes of Health (NIH), for example, now requires a data-sharing plan for large projects like MIDUS, and compelling data management plans often benefit from storing and sharing data via an official data archive.

Data distribution, curation, versioning, and preservation are critical data management tasks that can be handled very efficiently by research data archives. MIDUS has benefited from ICPSR's expertise in these areas as its primary archive and distribution channel for data and documentation. ICPSR ensures the long-term preservation of MIDUS data by normalizing submitted SPSS datasets into nonproprietary formats that are more likely to be accessible over time. In addition, they create a variety of distribution formats in commonly used media types. ICPSR also assigns persistent identifiers to the MIDUS study project, enables citation of data downloads, and provides version history by documenting and publishing updates to the data. Additionally, ICPSR also ensures the appropriate use of MIDUS data by conducting confidentiality reviews of the data, requiring data use agreements from users, and offering restricted access protocols to sensitive information. The curation and preservation services provided by ICPSR help make the MIDUS data and documentation more FAIR.

Finally, data archives such as ICPSR are often involved in the development and promotion of metadata standards that facilitate the discovery, interpretation, and use of digital data through enhanced documentation technology. The structured documentation that MIDUS has adopted to make the study independently understandable to secondary users has been easily incorporated into existing metadata schemes used by data archives such as ICPSR. Indeed, ICPSR has been closely involved in the development and support of a metadata standard that has provided MIDUS an ↳ enhanced documentation system that highlights the integrated nature of its data. p. 31

Challenge 3: Making Sense of MIDUS Data With Metadata

"Research outputs (data, code, etc.) that are poorly documented are like canned goods with the label removed: the contents may be something desirable, but it is impossible to tell without metadata" (Stasser, 2015, p. 7). This observation is relevant to research outputs of all kinds but is exponentially more important in complex studies that attempt to manage a wide array of integrated domains of data.

MIDUS recognizes that richly structured and well-formed metadata are central to facilitating multidisciplinary longitudinal research and producing research outputs that are discoverable, interpretable, and usable to a large group of public users. Adequate metadata are crucial to realizing the analytic potential and meaning represented in studies, datasets, and variables. This section unpacks the term *metadata*, discusses its importance, and describes metadata standards, including how MIDUS has benefited from one in particular.

Definition and Types of Metadata

Metadata—literally "data about data"—can be defined broadly as any "structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource" (Understanding Metadata, 2001). Although the idea of metadata has existed for a long time (library card catalogues, for example), the term as used in current research contexts describes information that supports the discovery, management, and interpretation of research data (Mark & Roussopoulos, 1986).

There are two general types of research metadata: study level and variable level. Study-level metadata include broad descriptions of the research objectives, such as those found in the MIDUS documentation or README files. This type of metadata can also describe study administration, data collection methodology, field reports, funding sources, and sampling and weighting strategies where appropriate.

Variable-level metadata include more granular information and in the social sciences have traditionally taken the form of a hard copy code book or data dictionary. Code books have often been considered the best single source of variable-level metadata about a dataset. Social science code books—especially survey research code books—often include the following variable-level metadata (ICPSR, 2017):

- full question text
- interviewer instructions
- skip directions, routing, or question universe
- response options, code lists
- variable names, variable and value labels
- variable groupings
- missing value specifications or schemes
- physical format (numeric, string, data, etc.)
- column identifiers and physical layout of data file
- frequency distributions or summary statistics

How Metadata Benefit Datasets

Metadata provide the bridges between the producers of data and their users, conveying information that is essential for secondary analysts (Ryssevik & Musgrave, 2001) while making the data independently understandable. Sufficient metadata also reduce the burden on the data producer or manager of responding to secondary users' questions and requests. Well-formed and comprehensive research metadata provide the following additional benefits to secondary analysis:

- Discovery. Metadata facilitate the discovery and access of data by associating identification, labels, and "handles" to it that can then be searched. This is especially true in a searchable digital environment like the Internet.
- Preservation. Metadata support the long-term preservation of data by ensuring that relevant information remains with the data for its future use or for conversion into new archival formats.
- Exchange. Common metadata language and structures are essential for supporting the exchange and interoperability of information between agencies, individuals, systems, and networks (Gregory, Heus, & Ryssevik, 2009).
- Harmonization. The assumption of cross-time equivalence in methods and procedures is one that lends longitudinal research and repeated measures designs advantages over cross-sectional studies. But, changes can occur due to methodological, technological, and organizational differences that threaten equivalence. Variable cross-walks and concordance tables are critical pieces of longitudinal metadata that document such changes and offer solutions for harmonization or reconciliation.

p. 32

Data Documentation Initiative and the MIDUS DDI Portal

In the past two decades, advances in computing technology and the advent of the Internet have made available ever-increasing amounts of research information. This situation has placed a premium on practical approaches to the efficient management of these growing data resources, and metadata standards have evolved as organizational frameworks to facilitate the conduct of all aspects of research. In a digital environment, metadata standards that are platform and software independent—such as the Data Documentation Initiative detailed below—are more likely to be widely adopted by the user community (and thus become a standard).

In recent years, metadata standards in many different fields have adopted Extensible Markup Language (XML) to express their standard on the web. XML is a web-based data-publishing language that renders data as "machine-actionable" digital objects that can be processed by computers in an automated fashion (Brown, 2003; W3C, 2013). XML accomplishes this by "tagging" text, descriptions, and other objects so that semantic meaning is expressed in the markup language (ICPSR, 2012). This tagging embeds "intelligence" in the metadata, facilitating its manipulation by computers and permitting flexibility in rendering the information for display on the web and in a variety of presentation languages. XML has provided a web-friendly basis for transforming code books into an interactive and machine-actionable format.

The integrative and longitudinal characteristics of MIDUS, along with the data management conventions it has adopted, have made the study an exemplar of metadata standards' ability to comprehensively document a complex investigation. Beginning in the mid-1990s, ICPSR was instrumental in supporting the development of the Data Documentation Initiative (DDI), an international XML-based standard for the compilation, presentation, and exchange of documentation for datasets in the social and behavioral sciences (Vardigan, Heus, & Thomas, 2008). Since 1999, nearly all MIDUS researchers have obtained their data and documentation through the ICPSR website. Indeed, the entire life cycle of most research conducted on MIDUS occurs in a digital environment, so it was an advantage for MIDUS to create electronic code books and interactive metadata using a standard endorsed by its online archive.

In 2005, MIDUS began using DDI to create electronic code books for its publicly available datasets at ICPSR. Since then, the DDI standard continued to mature from supporting the creation of relatively simple machine-readable code books to documenting complex longitudinal studies in machine-actionable ways. MIDUS kept pace with these developments by expanding the use of DDI from creating electronic code books to designing fully integrated DDI-compliant systems. Specifically, MIDUS created an interactive DDI-based online portal (http://midus.colectica.org/) that provides enhanced access to MIDUS data and documentation. The portal supports the integrative analysis of the variety of different datasets available in MIDUS by allowing researchers to search for variables of interest from across projects or waves and place them in a virtual variable basket for eventual download. The portal then creates an automatically merged custom dataset that is accompanied by an individualized code book. This system produces a tighter relationship between research metadata and analytic data, allowing researchers to concentrate on analysis instead of managing multiple data and documentation files.

The portal's well-structured and interactive DDI metadata have provided dataset producers and consumers a number of benefits:

- **Longitudinal harmonization.** DDI has facilitated the harmonization of MIDUS longitudinal and cross-cohort variables. The portal accomplishes this by linking or relating versions of the same variable across datasets and by providing information when they might not be strictly equivalent. The portal also incorporates visualization tools in its approach to harmonization. XML's flexibility in rendering information in a variety of formats and presentation languages means that cross-time frequency distributions and other descriptive statistics can be easily expressed via tables, graphs, or figures for longitudinal comparison.
- **Custom documentation.** The portal generates customized code books to accompany downloaded datasets; these code books include additional metadata not contained in the dataset, including full question text, universe, skip directions, and information on comparability, provenance (origins and transformations), and versioning (changes and updates). Each code book also incorporates links ↳ to stand-alone documentation that provides more extensive information about the study. p. 33
- **Comprehensive metadata and documentation.** DDI supports the use of Universal Resource Identifiers (names used to identify material on the web, like http://midus.wisc.edu/), so MIDUS DDI code books contain hyperlinks to other files, documentation, or instruments (hosted at ICPSR) that provide more detailed information. All MIDUS documentation and metadata are concentrated in one resource.
- **Intelligent search.** Because all MIDUS metadata are marked up in DDI, keywords can be searched in different fields across diverse datasets. A Boolean search function is able to search variable names, labels, question text, instructions, skip patterns, and concepts across over 25,000 MIDUS variables. Found variables can be tagged for eventual merge and download.
- **Quality control.** Marking up MIDUS data in DDI automates the review—and thereby increases the quality—of its data and metadata. Because variable-level metadata are tagged in DDI XML, software is able to identify errors or conflicts in variable coding, formatting, or labeling.

By adopting the DDI metadata standard, MIDUS has produced tools that support integrative health research by maximizing the utility of multidisciplinary longitudinal datasets. More broadly, using a metadata standard like DDI has provided a framework that supports the data management goals in MIDUS, ensuring the integrity of a complex data collection process and generating data products that adhere to FAIR principles. These goals have made MIDUS an eminently approachable digital resource for scores of researchers while ensuring the future viability of the study and providing an exemplar for the management of integrative health science research projects.

Future Directions

The conduct of MIDUS data collection and management has been and will continue to be an exercise in managing complexity. The current data-rich information technology environment has placed a premium on making sense of publicly available research data, and the data management field is quickly evolving to address this need, as evidenced by the recent publication date (Wilkinson, 2016) of the FAIR digital data stewardship principles. Yet, the FAIR principles only address individual digital objects, and the challenges for longitudinal research projects attempting to integrate diverse health domains are an entirely different dimension of these guidelines. MIDUS will continue demonstrating leadership by adapting new technologies and practices in this changing environment to ensure that good data management contributes fully to the research goals of integrative health science.

Notes

1. File management minutiae. The examples include date stamps as a type of version control. MIDUS supports two different types of calendar date formats. The first is an eight-digit format (YYYYMMDD) with no special characters. This format is recommended by the International Organization for Standardization (ISO; 2016) as an internationally accepted way to represent dates using numbers. There are many advantages to formatting dates this way, including that they are easily read by software, are language/software independent, and are easily sorted. The second format (MM-DD-YY) is conventionally popular, but lacks the advantages of the ISO standard. Regarding specific characters and symbols, not all software programs are able to read special characters or symbols in file names (such as =, &, %, etc.), so it is best to avoid them in file names. From the perspective of study curation and data management, the goal is to make file naming, labeling, and formatting interoperable across platforms or software systems by using machine-readable characters and syntax. Doing so increases the robustness and longevity of electronic files and all digital objects.

2. MIDUS adopted variable-naming conventions for its baseline data in the 1990s when most statistical software programs had strict character limits on variable names and labels. While contemporary software programs have evolved so that there are substantially fewer technical or memory limits to names and labels, MIDUS still tries to adhere to its original eight-character variable names for brevity and longitudinal consistency.

References

Ball, R. J., & Medeiros, N. (2011, July 11). *Teaching students to document their empirical research*. Retrieved from http://dx.doi.org/10.2139/ssrn.1892168
WorldCat

Bechhofer, S., De Roure, D., Gamble, M., Goble, C., & Buchan, I. (2010). Research objects: Towards exchange and reuse of digital knowledge. *Nature Precedings*. Retrieved from doi:10.1038/npre.2010.4626.1
Google Scholar WorldCat

Brown, P. J. (2003). *Information architecture with XML: A management strategy*. Hoboken, NJ: Wiley.
Google Scholar Google Preview WorldCat COPAC

International Organization for Standardization. (2016). *Date and time format—ISO 8601*. Retrieved from http://www.iso.org/iso/home/standards/iso8601.htm
WorldCat

Gregory, A., Heus, P., & Ryssevik, J. (2009). *German Council for Social and Economic Data Working Paper Series on Metadata* (Working Paper No. 57). Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1447866
WorldCat

Inter-university Consortium for Political and Social Research. (2017). *ICPSR: A Case Study in Repository Management: Metadata*. p. 34 Retrieved from http://www.researchconnections.org/icpsrweb/content/datamanagement/lifecycle/metadata.html ↳
WorldCat

Inter-university Consortium for Political and Social Research. (2012). *ICPSR guide to social science data preparation and archiving: Best practice throughout the data life cycle* (5th ed.). Ann Arbor, MI: Author. Retrieved from http://www.icpsr.umich.edu/files/deposit/dataprep.pdf
Google Scholar Google Preview WorldCat COPAC

King, G. (1995). Replication, replication. *Political Science and Politics*, *28*, 444–452.
Google Scholar WorldCat

Mark, L., & Roussopoulos, N. (1986). Metadata management. *IEEE Computer Magazine*, *19*(12), 26–36.
Google Scholar WorldCat

Radloff, L. S. (1977). The CES-D scales: A self-report depression scale for research in the general population. *Applied Psychological Measurement*, *1*, 385–405.
Google Scholar WorldCat

Ryssevik, J., & Musgrave, S. (2001). The social science dream machine: Resource discovery, analysis, and delivery on the web. *Social Science Computer Review*, *19*(2), 163–174.
Google Scholar WorldCat

Stasser, C. (2015). *Research data management*. Retrieved from https://groups.niso.org/apps/group_public/download.php/15375/PrimerRDM-2015-0727.pdf
WorldCat

Riley, J. (2017). *Understanding metadata: What is metadata, and what is it for?: A primer*. Retrieved from https://www.niso.org/publications/understanding-metadata-2017
WorldCat

Vardigan, M., Heus, P., & Thomas, W. (2008). Data Documentation Initiative: Toward a standard for the social sciences. *International Journal of Digital Curation*, *3*(1), 107–113.

Google Scholar WorldCat

W3C. (2013). W3C Extensible Markup Language (XML). Retrieved from https://www.w3.org/XML/(
WorldCat

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., . . . Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3, 160018. Retrieved from doi:10.1038/sdata.2016.18
Google Scholar WorldCat