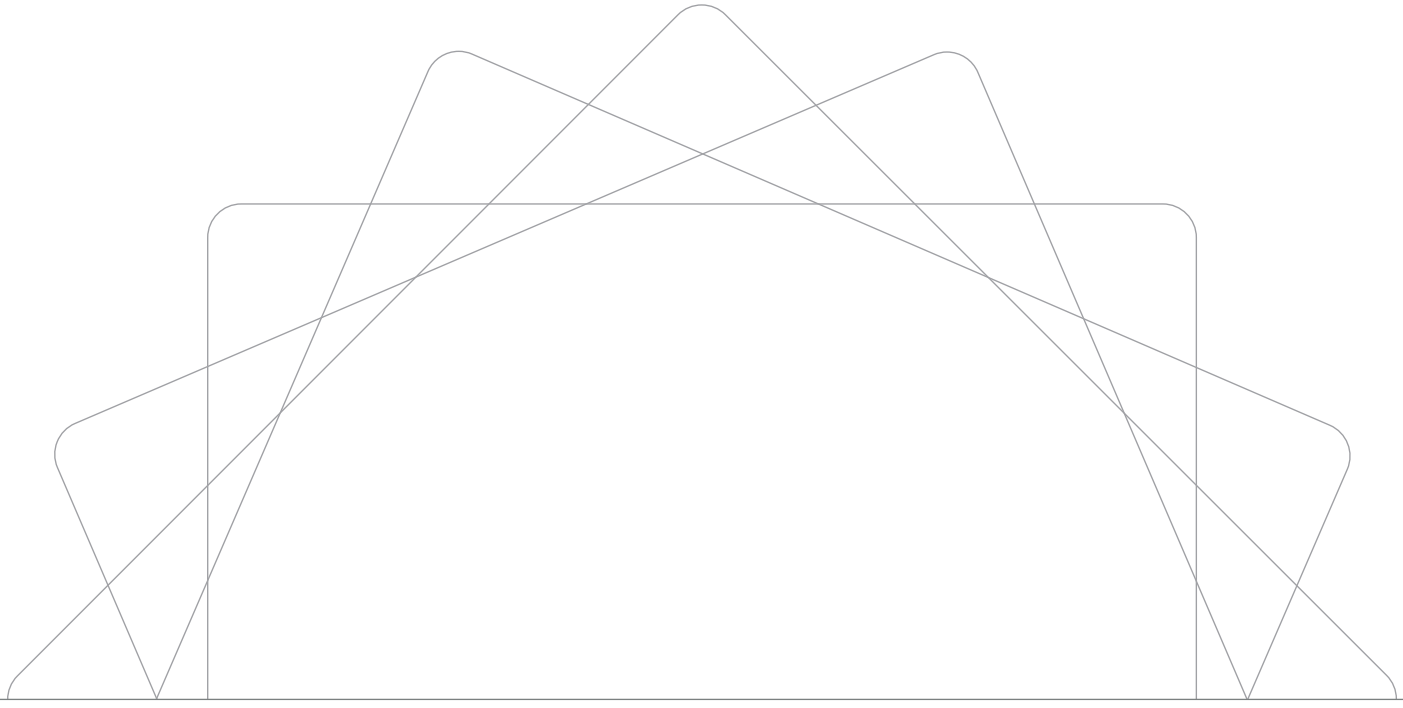
BOLD
THINKERS
DRIVING
REAL-WORLD
IMPACT

ABT WHITE PAPER

Digital for Research: A Taxonomy of Data Collection Methods and Solutions

Abt Associates
Brian Sokol, Vice President, Digital Delivery



Choosing the right data collection solution is critical to the success of policy research and program evaluation. Sometimes, researchers or clients accustomed to a particular approach or familiar with a certain tool might want to re-use that tool, even if it is not necessarily right for the current project. While it's often possible to tailor a tool to meet needs even when it is a sub-optimal choice, it's better to be familiar with the range of solutions and their core uses and choose the tool that best fits the need.

At Abt Associates, we work with many types of front-end tools. One significant distinction among tools is that some are fully **research-driven**, used to collect data generated only for the specific purposes of a research study; others are best for obtaining **activity-based**, secondary data not generated with research in mind. A second key distinction is that some tools are intended to collect data based on what research subjects **self-report**, in contrast to others intended for **tracking** subjects' interactions, experiences, or conditions.

Combining these two distinctions we can organize data collection methods and tools into four categories, wherein a taxonomy emerges. This taxonomy encompasses both a variety of types of data systems and alternative methods of obtaining second-hand data—jointly referred to as "approaches" —which are not ordinarily organized and compared within the same conceptual framework.

Quadrant Exemplars

The choice of approaches affects data content, reliability, and the level and types of researcher effort. For example, **research-driven** approaches require more effort for upfront system design and development and recruitment; **activity-based** systems require more effort in data transformation, validation, and cleaning. Let's first discuss the key exemplars of each quadrant—then expand the taxonomy to discuss approaches that fall in between these lines.[1]

Research-Driven, Self-Report

In *Surveys*, data collection is done exclusively for the research purpose. The text of the question; the available answers and how they are ordered; the patterns and logic; the frequency of the surveys themselves; and the sample of who's invited to take the survey all serve the research effort. Surveys are fully self-reported. The researcher asks questions and relies exclusively on the answer provided by the respondent—even if some responses may be self-contradictory or contradict what the person administering the survey observes or can verify. For example, when asking a question about gender, survey staff should record what the person says their gender is, regardless of how the person appears to the surveyor. These two fundamental characteristics of surveys hold true no matter the method: whether the survey is taken by phone, in person, over the web, by SMS, or by other means.[2]

Figure 1: Approaches in Each Quadrant

Research-Driven, Tracking

Electronic Data Capture (EDC) solutions are optimized for when the researcher is studying a particular program or treatment and wants to observe interactions and results, rather than rely on a self-report. An example of this is in health research, when a researcher obtains results for medical tests administered as part of a study conveying the effectiveness of an intervention. It can also refer to a researcher recording attendance at a particular training, or mentoring session, administering a pre- or post-test, or recording observed behavior in a specific scenario constructed by the researcher. In these cases, research drives the observation; the data would not exist otherwise.

Activity-Based, Self-Report

We use an **activity-based, self-report** approach when the activity itself involves interviewing participants. This frequently occurs upon a participant's enrollment into a social service program. For grant reporting, eligibility screening, and service delivery purposes, the program staff may ask the new participant questions about their identity, household, health, and income. Such questions —and more in-depth ones—may also be asked as part of *Case Management* activities to provide the most appropriate direct service or referrals. Researchers can piggy-back on the data collected from these activities. But the solution is built and optimized to help the service providers do their job effectively, even if some of the data is not directly useful for research.

Activity-Based, Tracking

Activity-based, tracking occurs when the researcher obtains data on subjects' activities from sources other than the subjects' own self-report, such as *Administrative Data, Electronic Health Records, Social Media analytics, Internet of Things (IOT) data, or Web-Scraping.* For example, a researcher can obtain Medicaid claims records to study how the program is being used and by whom, as well as the prevalence and duration of particular conditions. We can also use electronic health records for studying health conditions, prevalence, disparities, and the outcomes of specific interventions and treatments. Web scraping or analytics is often useful to track behavior trends. IOT data may include location or health data tracked on people's phones or other devices. Virtually all programs and conditions needing study accumulate data on their own and can be used secondarily for research. This data may be comprehensive and accurate, but it is generally the "messiest" and "noisiest" and requires the most cleaning and data processing. The corresponding technology solutions must have strong integration and Extract, Transform, and Load (ETL) capabilities.

Activity-Based, Tracking Example

The Summer Electronic Benefit Transfer program provides grocery assistance to households with eligible children during summer months when millions of U.S. children lose access to free or reduced-price school lunches.

Abt analyzed the grocery transactions made by Summer EBT participants to learn what foods they purchase, how much, and who makes the purchases, how often benefits are used, and the impact of household distances to grocery stores.

The In-Betweens

Often, the research requires a solution that doesn't fit neatly into one of the four categories. Here are some examples:

Between Self-Report and Tracking

- **Research-Driven:** *Retrospective Surveys* are self-reported, but the **self-report** is focused on a particular event, activity, or experience. The researcher is not asking for the subject's thoughts but a true history of what has occurred. It is a way to do "tracking" when direct tracking is not possible but

is prone to recall bias. An example of this, slightly closer to **tracking** than to **self-report,** is the use of a *Participant Diary*, in which they log an activity at the time they do it, rather than trying to recall it in the future. This retains the event-driven nature of **tracking**, while still relying on the participant to report the event.

- **Activity-Based:** *Data Portals and Grant Management Systems* encompass both **self-report** and **tracking** data. We use these systems to send data on program outputs and outcomes to project funders and other stakeholders. They aggregate and cross-cut individual self-reported data and service activities to report services provided and outcomes by key demographics. They are also often supplemented by **self-reports** of program staff on the status of program activities. The main difference between the two is that data portals are used primarily for this data reporting and are thus somewhat closer to having a research function, while grant management systems are used for other purposes as well, such as tracking deliverables, contracts, and funds.

Between Research-Driven and Activity-Based

- **Self-Report:** *Omnichannel Call Center* solutions can be used to gather self-reported data both to provide services and fulfill a research need. In this case, participants are asked to contact the researcher when a particular event occurs. Multiple channels, such as phone, web, SMS, or chat could be used to report an address change, a change in earnings, family or housing situation, or health or vaccination status. This report could trigger a referral to a service provider or a benefits adjustment, and the data are also used for research purposes.
- **Tracking:** When a researcher needs to gather **tracking** data, but that data is not collected organically as part of the activity, one approach is researcher *Observation*. For example, a researcher can go into a classroom and observe teacher techniques. A second approach, closer to **activity-based**, tracking on the continuum, is leveraging *Audit* data. In this scenario, an auditor is already observing program operations and collecting data to ensure quality and compliance, and the data can also be used for research purposes. An audit-focused solution can be used that is optimized to answer a series of questions about a particular program while on-site. A third flavor—closest to the activity-based quadrant—are *Mobile Applications*, which can independently track where and what staff are doing during their day (e.g., in a time and motion study). The solutions in this category are generally at the program level, rather than at the individual participant level.

Grant Management System Example

The Family and Youth Services Bureau (FYSB) is required by Congress to develop and deploy a Homeless Management Information System (HMIS) for its Runaway & Homeless Youth Program (RHY). The agency wants to ensure it collects high data quality—and provides statistical analyses—from FYSB's more than 600 grantees.

The data intake and consolidation system used to support data submission by FYSB-funded grantees feeds into a RHY-HMIS Dashboard, which provides access to the data collected and the information necessary for decision making to enhance outcomes at the program level and nationally.

All of the Above

- In many cases, using an existing **activity-based** system is the best way to gather the data, but the solution that is in place is not configured to collect all the data that the researchers need. In this case, the

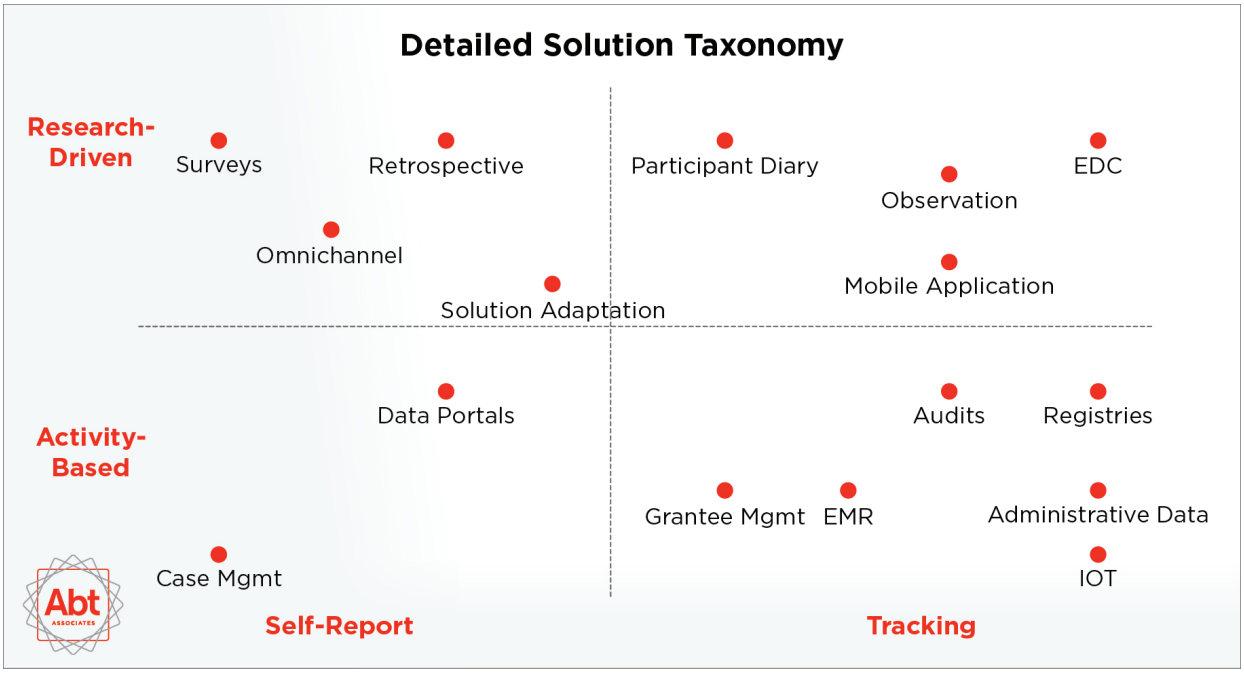
research team can work with the program to adapt the solution currently in use (*Solution Adaptation*), by asking additional questions or modifying the wording or available answer choices to meet the research needs. This could be necessary both to collect the data in the first place or make the data collected consistent across multiple programs. In this case, consulting services and programming support are needed, rather than a particular solution. Suppose various solutions (e.g., from different sites) are in use, each of which is contributing data; we can combine that solution adaptation with data integration while merging data from different systems for analysis. We should prioritze facilitating both the core **activity** and **research** as part of the solution design. This approach can apply both to collection of **self-reported** data and to **tracking** service delivery activities.

Figure 2 shows a more continuous view of the four quadrants, with the various approaches described above represented in-between the purer exemplars displayed in the corners.

Hybridization and Combination

None of these solutions are necessarily fully exclusive, as most product developers are aware that often a combination of data collection activities are required. So, for example, it is often possible to do tracking using a solution designed primarily for surveys or to pull administrative data into a survey system. That is, one solution can be used for hybrid purposes. But many solutions are built with one of the above use cases in mind, leaving research teams and developers working around the core design of the system. It often makes sense to do this rather than use multiple systems. However, in projects where optimal solutions are needed for divergent data types, the ideal solution is often to use multiple best-in-breed solutions of different types and then merge the data together in the data integration and management phases.

Figure 2: Digital RME Detailed Solution Taxonomy [3]



Combination Example: Surveys and Electronic Health Records

To monitor potential safety issues of COVID-19 vaccines for pregnant people, the Centers for Disease Control and Prevention (CDC) created the v-safe COVID-19 Vaccine Pregnancy Registry to collect data on adverse outcomes such as miscarriage, stillbirth, and pregnancy complications.

Abt's call-center staff enroll vaccinated pregnant people into the registry to collect information about adverse outcomes following vaccination. Interviewers administer telephone surveys to participants once during each trimester, at the end of pregnancy, and when the newborn is three months old.

Interviewers seek information about healthcare providers for the pregnant person and infant, which CDC uses to collect medical record data for a subset of registry participants.

Data Integration and Management

Regardless of the data-gathering approaches, the data needs to be processed, cleaned, and analyzed. Key tasks that often occur after data collection include the following:

- **Data integration**, which includes automatically downloading, transforming, and merging data in multiple formats and protocols.
- **Data warehousing** for tracking and versioning raw, intermediate, and final datasets.
- **Data validation and quality checks** to identify both large scale issues like duplicate data or systemically missing values, as well individual cell-level errors.
- **Free-text review** to analyze, categorize, and possibly de-identify information that is otherwise unrestricted to particular values.
- **Implementation monitoring** to track the progress of the study itself such as the number of people enrolled and withdrawn, or the number of tests conducted, or records obtained. This might also include metrics on the quality of the data obtained.
- **Analytic file development**, including constructing variables that are more suitable for analysis—such as transforming multiple choice options into binary values for each option—and merging rows and implementing rules for the "best" value (e.g., earliest or most recent, or most plausible).

Depending on the type of research all these tasks can occur within minutes or over the course of a month. Of course, all of these tasks are in service of the final step, the end goal of it all, which is to analyze the data, gain insights, and disseminate it as needed.

Conclusion

There are, of course, many practical factors to consider when determining the best data collection approach, including availability of data, privacy issues, cost issues, training, standardization, development timelines—and, last but not least, cost. But, from a research perspective, the ideal course of action is to determine the most comprehensive, representative, and valid data to answer a research question and then find the most appropriate tool to meet the need.

Learn more about how Abt applies technology to our Research, Monitoring, and Evaluation work.

References

1 This paper focuses on collecting quantitative data and does not consider the varieties of qualitative data collection approaches, though some of the key categories are transferable. It also does not consider secondary literature reviews and primarily considers data collection on human subjects, rather than environmental, zoological or other types of scientific research.

2 These categories are not "perfect." There are some types of survey data (e.g., customer or patient satisfaction surveys) that are conducted primarily for non-research purposes. but are also used by researchers. In that case, we would consider it "activity-based, self-report," instead of research-driven.

3 This table is illustrative and represents the ideas of the author around general placement of these approaches in the overall schema. It does not depict any underlying scoring.