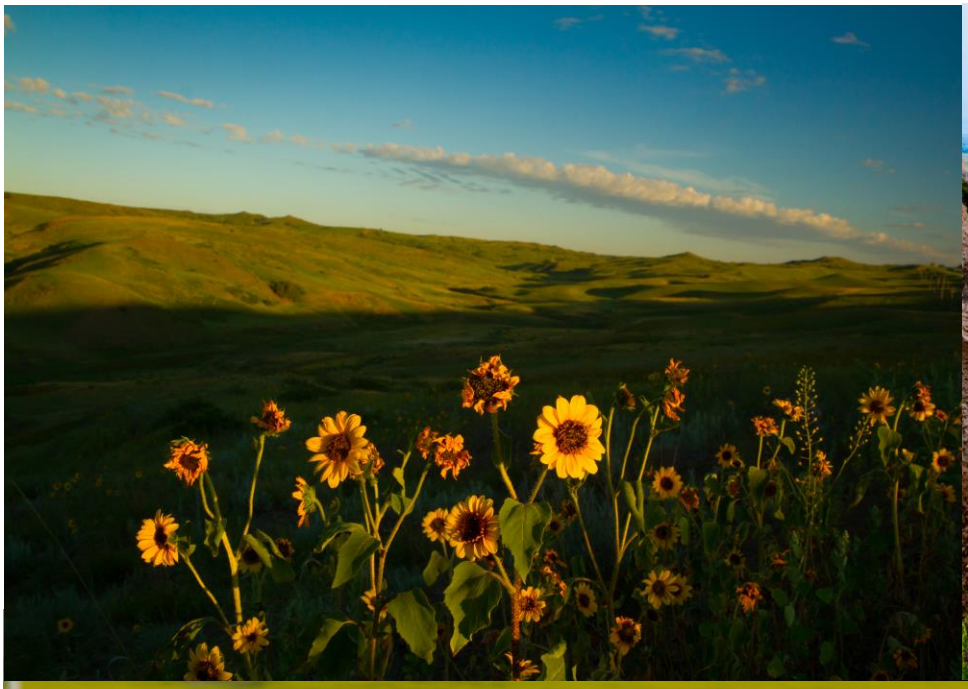
CRCAIH Research Data Management Toolkit



The mission of the Collaborative Research Center for American Indian Health (CRCAIH) is to bring together tribal communities and researchers within South Dakota, North Dakota, and Minnesota. Our goal is to build tribal research infrastructure and transdisciplinary research teams to improve American Indian health through examination of social and environmental influences. Research data is an important part of addressing health concerns in tribal communities as it offers an avenue for exercising tribal sovereignty through data stewardship and the pursuit of health priorities that match the values and needs of the community.

The main objective of the CRCAIH Research Data Management Toolkit is to highlight methods for tribal nations to acquire data from researchers and store data in a way that promotes secondary use (i.e., use of the data for purposes or needs different from those posed in the original research study). The toolkit developed from work that the CRCAIH Methodology Core has done to assist our Tribal Partners in efforts to build research infrastructure and harness the power of data to improve the health of their citizens. It offers practical tools and guidance for tribal research data management, though aspects of the toolkit may also apply to management of other types of data (e.g. data collected for clinical or surveillance purposes).

In addition to serving as a resource for our CRCAIH Tribal Partners, we hope this toolkit can benefit other American Indian and Alaska Native tribal nations, Indigenous nations, and those interested in research data management. We encourage users to review the materials and guidance in the toolkit to identify areas that resonate with the research data management needs of their communities, and visit www.crcaih.org for additional CRCAIH resources, including instructional videos and webinars related to tribal research and our Tribal Institutional Review Board (IRB) Toolkit developed by the CRCAIH Regulatory Knowledge Core.

Table of Contents

I. Importance of Research Data Management

A significant amount of data has been collected from American Indian and Alaska Native (AI/AN) peoples through past research studies, and continues to be regularly collected in the present. Yet, for many tribal communities, it is challenging to access and manage research data in a way that is meaningful and useful for their populations. Issues can arise from a variety of situations. In some instances, findings are not reported back to tribal communities at all, making it difficult to determine whether any community benefit resulted from participation in the research. In other instances, findings are shared with the community, but the format does not match the community's needs or capacity. For example, a research team may only provide summary results (e.g., percentages about the overall population or a particular sub-group), which do not allow further analysis of the data. Or, a research team may only provide raw data (i.e., data that has not been analyzed) contained in a computer software program that requires additional training or financial expense by the community to be of any use.

It is imperative for tribes to be able to build research capacity and infrastructure to ensure that research data can be accessed by the tribe and utilized for other purposes following completion of a study (Cross, Fox, Becker-Green, Smith, & Willeto, 2004). Research data can have many important uses. It can offer a glimpse into key issues for tribal nations by allowing detection of demographic, health, and social trends. It can also help to identify areas for future research and facilitate the establishment of priorities that can be addressed through policies or programs. Research data can even be used as baseline data to apply for additional grant funding or to track indicators over time.

Tribes have a vested interest in protecting research data obtained from their communities to make sure that it is being used in ways that can directly benefit them. For tribal communities who choose to participate in research, a research data management plan is one way to exercise tribal sovereignty. It creates an opportunity for data stewardship that aligns with tribal needs and values, and establishes the structure necessary for tribes to harness the power of local data for tribally-driven approaches to community needs.

There is no one-size-fits-all solution to tribal research data management. However, assessment of existing data management needs and capacity, as well as future community research goals, can allow a tribe to put in place the components of a data management plan that will best serve their community. The next few sections of this toolkit are designed to assist with development of procedures related to tribal research data access and management.

II. Starting a Research Data Storage Process

A. Overview of Maintaining Research Data

Decisions about maintaining tribal research data are extremely important to address early in the process of overseeing and regulating research. The CRCAIH Methodology Core has been working with our Tribal Partners on creating and implementing procedures for the preservation of research data in order to increase self-sufficiency and administrative resources.

One concept to consider in maintaining research data is data stewardship. Data stewardship entails caretaking of the data and verifying if the data is accessible and preserved. Data stewards are vital for ensuring the integrity and quality of available data.

It is important to establish procedures for storing and maintaining research data for various reasons. Some benefits of storing research data are listed below.

Benefits of Research Data Storage:

- Properly storing data is a way to safeguard research investment
- Data may be accessed to enhance and guide future research initiatives
- Tribal researchers or entities might wish to evaluate research results

B. What You Need to Know About Research Data

What is Research Data?

In research the term "data" can refer to many different things. Some examples of items that are referred to as data could include completed surveys, aggregated and summarized tables and results, research findings, published results, and laboratory specimens.

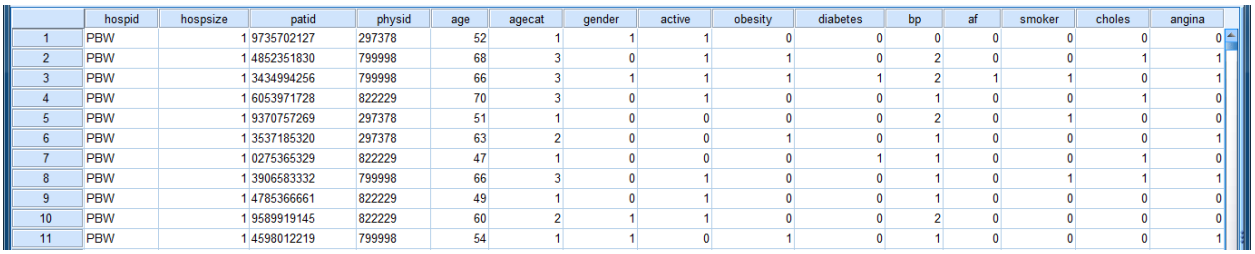
The National Institutes of Health (NIH) defines research data as: Recorded factual material commonly accepted in the scientific community as necessary to document and support research findings (NIH Data Sharing Policy and Implementation Guidance, 2003). This does not mean summary statistics or tables; rather, it means the data on which summary statistics and tables are based. The name for this type of data is raw or primary data.

i. Primary (raw) data:

When the data is first collected, this is called the primary or raw data. This data has not had anything done to it; it is whatever was collected initially. It has not been subjected to processing or any other analysis. Examples of primary (raw) data are listed below.

- Examples:
  - Medical records
  - Completed surveys
  - Database files containing personal information
  - Recordings of interviews or focus groups
  - Unmodified transcriptions of interviews or focus groups

Figure 1: Example of Raw Data File



	hospid	hospsize	patid	physid	age	agecat	gender	active	obesity	diabetes	bp	af	smoker	choles	angina
1	PBW	1	9735702127	297378	52	1	1	1	0	0	0	0	0	0	0
2	PBW	1	4852351830	799998	68	3	0	1	1	0	2	0	0	1	1
3	PBW	1	3434994256	799998	66	3	1	1	1	1	2	1	1	0	1
4	PBW	1	6053971728	822229	70	3	0	1	0	0	1	0	0	1	0
5	PBW	1	9370757269	297378	51	1	0	0	0	0	2	0	1	0	0
6	PBW	1	3537185320	297378	63	2	0	0	1	0	1	0	0	0	1
7	PBW	1	0275365329	822229	47	1	0	0	0	1	1	0	0	1	0
8	PBW	1	3906583332	799998	66	3	0	1	0	0	1	0	1	1	1
9	PBW	1	4785366661	822229	49	1	0	1	0	0	1	0	0	0	0
10	PBW	1	9589919145	822229	60	2	1	1	0	0	2	0	0	0	0
11	PBW	1	4598012219	799998	54	1	1	0	1	0	1	0	0	0	1

ii. Processed data

After the raw data has been collected, it is subject to different levels of processing to help make sense of the data or for protection of research subjects. Two specific types of processed data are described below: de-identified data and summarized data.

De-identified data: When research data contains information on human subjects, protecting the privacy of research participants is very important. One of the ways that researchers protect privacy is by de-identifying data. This is done by removing all of the information that could be used to identify an individual so that the data cannot be linked back to the person who provided it. Examples of information that should be removed to de-identify data are listed below.

- Examples:
  - Names
  - Addresses or other physical locations
  - Email addresses or phone numbers
  - Account numbers
  - Medical record numbers
  - Date of birth
  - Dates of medical services or other identifying dates

Summary data: The goal of processing data is generally to understand the data and obtain information about the research. To do this, the data is processed into summary data. Summary data includes aggregating the data by group (e.g. treatment vs. placebo, male vs. female) or describing the data set as a whole. The goal is to provide an overall picture of the group or groups on which the data was collected. Examples of this type of summary data include averages, counts, and percentages (e.g. 25% of participants reported smoking during the study). Graphs are often used to summarize research data in a meaningful way to assist with interpretation. Sometimes patterns and trends are more apparent in a graph than just in a table even if they are based on the same information. Another way data is summarized and interpreted is through statistical analysis. This is a way of processing the data that answers specific questions or hypotheses. This is often the primary goal of a research study. Questions like if a new treatment or drug is better than an old treatment or drug can be answered through statistical analysis.

Information about the data: Just having a data set is not enough to understand the data. In addition to the data sets themselves a description of the data along with code book or data dictionary can help those unfamiliar with the data to understand the context of the data. This type of data is sometime called meta-data. Formal descriptions of the data include the rationale behind the project, who contributed to the project, when and where the project was conducted, what type of data was collected and how was it processed and analyzed. In addition to this basic data, a codebook or data dictionary provides more specific information about the data and may include variable names and labels, a description of how certain variables were created, and codes or classifications used in the data (e.g. Sex: Female = 1, Male = 2).

III. Data Return

A. Importance of Data Return

Following the conclusion of a study, research data returned to a tribal nation can have many uses. Tribes have a vested interest in protecting research data to make sure it is used in a way that aligns with tribal values and directly benefits their citizens. Research studies conducted within a tribal nation can offer important insights by supplementing data already collected by other programs and services, or providing data in cases where information is not being collected at all through existing tribal infrastructure.

Data returned to a tribal nation can allow a tribe to determine what is happening in their communities and shape next steps for addressing challenges or bolstering strengths. It can help establish future research directions or assist in priority setting for tribal policies and programs in a variety of sectors. Research data returned to a tribal nation can also function as a springboard for additional studies, provide an overview and means for tracking health factors over time, and serve as baseline data to apply for additional grant funding.

B. Data Return Process

Early on in the process, a decision needs to be made on what data the tribe would like returned from researchers working in their area. Developing a plan for communicating with researchers should be done prior to a research project being conducted and include clearly written instructions for researchers to utilize when gathering and storing data. Items to think about when developing this plan include defining what, how, when, and where data should be returned as well as items such as the privacy of research participants and other data sharing policies.

Additionally, staffing issues and storage procedures need to be considered. There should be an individual identified that is available to keep track of past and ongoing studies, verify that the data is returned correctly, and make sure that it can be correctly stored according to the storage process. If staffing and storage is not sufficient to maintain data internally, tribes can also require that researchers be responsible for storing their own research data for a study and returning it to the tribe upon request. However, this approach is contingent on being able to locate the researcher when data is requested.

Developing a process for return of research data is necessary to ensure that all data is returned to the tribe according to specifications that are set by the tribe as well as maintaining the confidentiality of research participants. This will help maintain consistency of research data over time for a large number of studies.

i. Returned Data

Before data is returned to the tribe, the principal investigator (i.e., the lead researcher for the study) should be informed of exactly what data should be returned. The term data has many different meanings, so requesting specifically what data needs to be returned is extremely important. Research studies can collect data on small number of variables (e.g., basic participant demographics and a score on a single mental health scale) or on a very large number of variables (e.g., participant demographics, scores on multiple mental health scales, laboratory specimens, and medical records). Questions to consider include:

- Do we want data on every variable?
- Is a database file sufficient or should the researcher return completed paper surveys?
- Should a Codebook or Data Dictionary be included with the data?
- How can we link the data to regulatory documentation including a copy of the study consent form?

The following is an example of a statement about what data should be returned:

> The investigator shall return all electronic databases containing all of the primary de-identified participant information along with all analysis files, project summaries, and manuscripts based on the research study. These must be accompanied by meta-data including a complete description of the project (rationale behind the project, who contributed to the project, when and where the project was conducted, what type of data was collected and how was it processed and analyzed), codebook or data dictionary, and IRB approvals and participant consent forms.

ii. Format of Returned Data

Some file formats stand up over time better than others. If a researcher sends in a file containing a database that requires expensive software that the tribe does not own then that data is not useful. Make sure to be very specific about the format that is requested. Universal data formats are ideal for data preservation. Universal formats should:

- be accessible in the future
- be non-proprietary
- be commonly used

The following table (Table 1) includes some examples of different types of files and file types of universal formats and proprietary formats. Depending on which computer operating system is being used or which software is installed, a user may not be able to open all proprietary files without additional software.

Table 1: Universal File Format Examples

Type of File	Example of Universal Format	Example of Proprietary format
Spreadsheet	Comma Separated Values (.csv) Tab-delimited text file (.txt, .tsv)	Microsoft Excel
Report	Portable Document Format (.pdf)	Microsoft Word
Image	PNG (.png) or JPEG (.jpg, .jpeg)	Bitmap (.bmp) or Photoshop (.psb)
Audio	MP3 (.mp3)	Windows Media Audio (.wma)
Video	MP4 (.mp4)	Windows Media Video (.wmv) or QuickTime File (.mov)

iii. Timeline of Data Return

Data can be returned at many points during a study. A best practice is that data is returned upon completion of the study. This practice guards against unintended duplication of data that could arise from data returned at multiple time points, and ensures data returned to the tribe matches the data that the researchers used to arrive at their final results.

What is defined as the completion of the study may differ depending on specific criteria. For example, some IRBs close their oversight of a study after the data is collected and de-identified while other IRBs consider the study open until all study activities are complete and final manuscripts are published, even if the data is de-identified. Thus is it necessary for the data return policy to specifically identified at what point they consider the study complete.

iv. Additional Considerations

Other things to think about when developing a data return policy include other storage and data sharing policies associated with the research study and the confidentiality of the research participants.

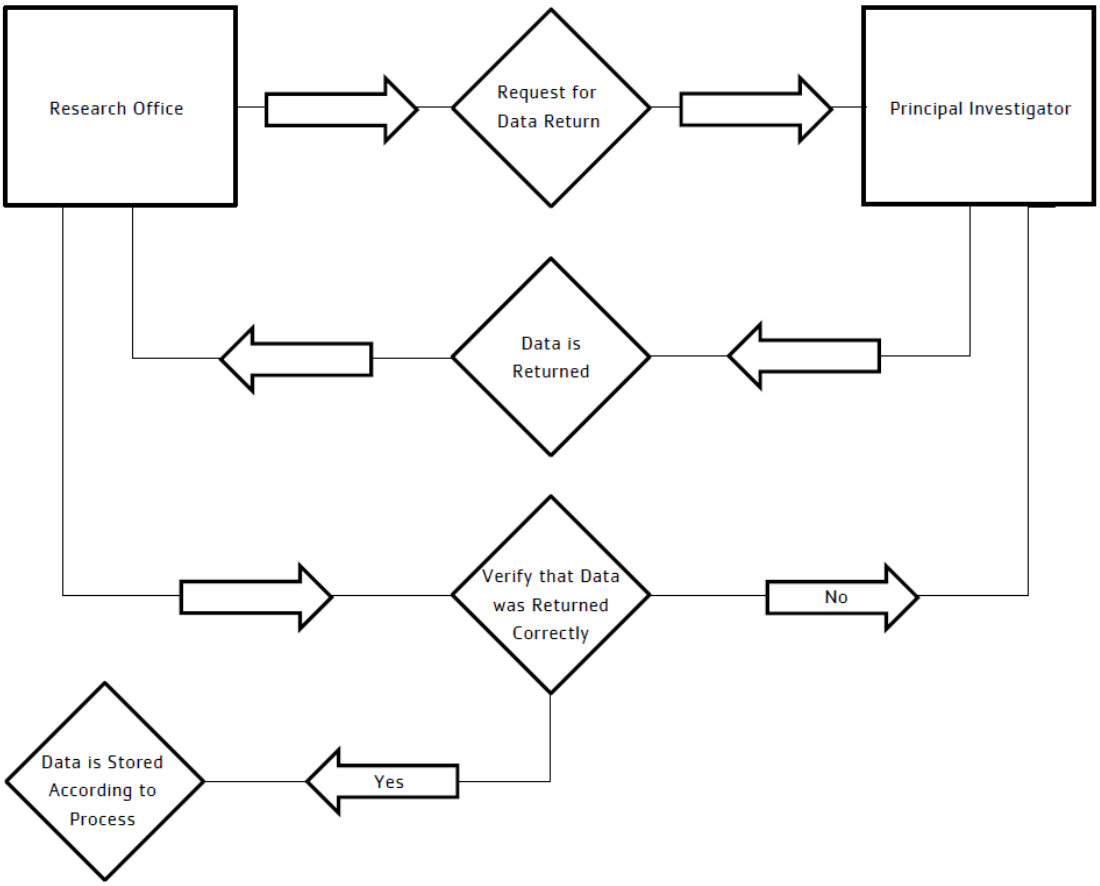
It is common practice in human subjects research to create de-identified data sets and keep the data indefinitely. Tribes need to inform researchers if they do not want the researchers to maintain a separate copy of the research data. In addition, data sharing policies with entities such as the NIH should be understood and agreed to by all parties prior to the start of a research project. These data sharing practices should be consistent with the tribes requirement for research projects. More information about

NIH data sharing policies can be found at http://grants.nih.gov/grants/policy/data_sharing/.

Confidentiality of research data is a concern for all researchers and tribes approving research projects. When implementing a data return policy this confidentiality must be protected. This may include return of de-identifying data with additional consideration for de-identification in small groups. For example, even if all identifying information is removed it may be possible to identify individuals by rare diagnoses in small communities.

The following diagram (Figure 2) shows an illustration of a sample data return process that involves data returned to a tribal research office. In this example, the tribal research office first sends a data return request to the principal investigator of the study. The investigator should then send the data back to the research office according to the timeline and specifications provided. The personnel at the research office verify that the data was returned correctly according to the requested specifications. If the data was not returned correctly, the principal investigator should be contacted and informed of the policy violation and be requested to re-submit the data according to the correct specifications. When the data is returned correctly it moves into the data storage process.

Figure 2: Data Return Process



The data return process illustrated is only an example. There is not a one-size-fits-all solution to data return. Each tribe should carefully develop their own process that fits with the personnel and resources that are available as well as the overall anticipated use of the data by the tribe. A template of a data return checklist to be utilized by tribes can be found on the next page.

CRCAIH Research Data Return Checklist Template

CRCAIH

Research Data Return Checklist

Investigator INFORMATION			
Name of Principal Investigator:			
Mailing Address:			
Telephone number:		Email Address:	
Project Title:			
Time Period Data was Collected			
Data Returned Date:			
Keywords:			

REQUIRED CHECKLIST	
☐	The flash drive has content
☐	There is a data sub-folder
☐	The Data folder contains de-identified data (raw data with identifying variable removed) in a universal format(like .txt or .csv)
☐	There is a file containing all required study information in a universal format (like .txt or pdf.) to fully explain all of the variables in the data set.
☐	There is a Codebook and Instrument folder containing a Codebook or Data Dictionary in a universal format (like .txt or pdf) to fully explain all of the variables in the data set.
If any of the boxes above were not checked a letter should be sent to the Principal Investigator to request that the missing information be submitted.	
OPTIONAL CHECKLIST	
☐	The Data folder contains de-identified data (raw data with identifying variable removed) in a software specific format (SPSS, SAS, Excel, or other)
☐	The flash drive contains a copy of the study grant
☐	The flash drive contains relevant IRB forms
☐	The flash drive contains publications and reports in a universal format (like .txt or.pdf)

Research Office Staff Reviewer:	
Date:	

IV. Data Storage

Good data stewardship requires that data is properly maintained and stored once it is returned. Beyond the simple act of holding data, tribal data stewardship acknowledges that information collected through research studies with tribal communities may cover a range of topics from low-risk to highly sensitive individual and cultural information, all of which is valuable by virtue of the fact that is has been provided by consenting tribal citizens and has the potential to impact tribal wellbeing. This data must be securely held and managed in a way that preserves it, protects the research subjects and the tribe, and allows for efficient access for future use.

If the tribe determines that it will collect research data, there should be a secure location to store the data. There may be an Information Technology (IT) department available for assistance with storing electronic data or it can be done by an individual with a minimal infrastructure.

A. Data Access and Security

In order to secure research data and prevent it from being altered or corrupted, access to the data should be limited. Allowing only certain people to access the data will hopefully prevent accidents like erasing or moving files, changing data or filenames, or improper usage of data. Identifying the individuals that can access data files is essential. While access should be limited, there should always be more than one individual that has access to files. It is necessary to have more than one person with control over the data storage just in case the original holder loses accessibility.

Data should be password protected, with access to the password limited to the individuals that have access to the data. Data could be rendered unusable if it is stored on a computer that is infected with a virus or other malicious software, otherwise known as malware. Anti-virus and anti-malware software should also be taken into consideration to prevent against potential cyber-attacks. Other possible protections to think about include a firewall and data encryption scheme. For more information on computer protection, take a look at the following websites:

https://www.getsafeonline.org/protecting-your-computer/

http://www.consumer.ftc.gov/media/video-0056-protect-your-computer-malware

http://www.consumer.ftc.gov/media/video-0081-computer-security

http://www.online-tech-tips.com/computer-tips/how-to-protect-your-computer-from-hackers-spyware-and-viruses/

B. Location

Storage location is also important to keep in mind when working with data from research studies. Some types of data files, such as videos, audio recordings, and images take large amounts of storage space.

There are many options available for storage of data contained within electronic or computer files. Three possible options include storage on a local device, cloud storage, and file servers storage. Each option has advantages and disadvantages and each tribe will have to determine which option is best for its community (Table 2).

i. Local Storage

Many files are stored on local devices such as the hard drive of a personal computer, external hard drive, flash drive, or a CD/DVD. Storing on a local device introduces risk since these drives are likely to have a limited lifespan and can be easily damaged.

ii. Cloud Storage

Cloud storage is an emerging tool in which a third party stores the data. This data can then be accessed through the internet. Cloud file storage is usually managed by a hosting service that may charge a monthly service fee for utilization. Some examples of services that provide cloud storage include Dropbox, Apple iCloud, Amazon Web Services, and Microsoft OneDrive. Cloud storage requires a secure internet connection to upload, download, and manage data files.

iii. File Server Storage

A file server is a dedicated local system that is usually accessed via a network by multiple computers. File servers typically contain more storage space than personal computers or external hard drives and space can be allocated for specific purposes. Servers are usually maintained by an IT department.

C. Data Backup

What would happen to valuable research data if the data storage device is damaged? In order to protect this valuable resource, data should be backed up on a regular schedule using a well-documented process. Backing up data is making multiple copies of the data and storing them in separate locations in case something bad happens to the original data files. The simplest example of this would be copying the entire contents of a disk drive to a separate disk drive. It is typically recommended that there are at least two backup copies of the data plus the original. It is also a good idea to store at least one of the backup copies of the data in a separate physical location so that it is not harmed if there is physical damage, like a fire or flood, in the location of the original.

The data backup process needs to include a specific timeline that is followed. Data could be backed up daily, weekly, or monthly. The more frequently that data is backed up, the more likely it is to be recovered if needed. However, data backup takes time so it is recommended to make a plan and stick to it. Unfortunately one missed backup can lead to data loss.

Documentation of the procedures is required when backing up data. It is also recommended that the process be tested to ensure that data can be successfully retrieved when the backup is needed. Testing a data backup procedure could be as simple as periodically utilizing one of the backup copies to attempt to access data files, verifying that all files are able to be located and the data can be utilized.

The following table (Table 2) highlights advantages and disadvantages of using cloud storage, a dedicated server, and local storage for storage and backups.

Table 2: Data Storage Options

Storage Type	Advantages	Disadvantages
Cloud	• Easy to use • Backup/Recovery is covered by service provider	• Data is not local • Monthly service fee
Server	• Locally managed by trusted professionals • Backup/Recovery is responsibility of IT	• IT department or staff needed to support/maintain servers • Equipment is expensive to purchase and maintain
Local Storage (Hard Drives)	• Data is very accessible • Data is local	• Disks can be corrupted or damaged • Backup takes time

D. File/Folder Naming Convention

When data is going to be stored, it should be organized in a systematic fashion. An example of a systematic organization scheme is the Dewey Decimal System that is used to organize books in a library. This system allows librarians to place books into sections of the library that contain other books that have similar subjects. If a library patron wants to locate a specific book, they are able to navigate to the appropriate section and locate the book based on the number that is assigned to it.

A method of organizing the data files from research studies will allow the tribe to locate and utilize the data in an efficient manner. Personnel responsible for storage and retrieval of data may change over time, and one person's organization scheme may not be useful to someone else. If there is a consistent naming scheme that is used it will be much easier to locate data when it is requested.

Begin by setting up a file directory (folder) structure that consistently organizes files from studies. Using descriptive terms in folder names can help identify the contents of the folder.

A clearly defined file naming scheme for research data storage should also be implemented and documented. File names should describe what is contained in the file. For example, the filename "document1.txt" does not give any indication of what is in the document but "data_management_plan.txt" does describe the file contents. Each file name also should contain a file extension at the end. In the examples listed ".txt" is the file extension. A file extension informs a computer's operating system which program can be used to open the file. Below are some recommendations for file naming (Organizing Your Data):

- Try to keep the file names relatively short (under 25 characters)
- Use the underscore character (_) rather than spaces in file names
- Avoid using symbols in file names
- There should only be one period (.) in a filename (part of the file extension)

V. Secondary Data Analysis

A. Identifying Useful Data

While the primary use of research data is to perform the original research project, it also could have potential for other uses. Research data could be useful for tribal entities for items such as reporting purposes, additional research projects, or as a baseline for applying for additional grant funding. Besides specific research data, aggregate data from reports, publications, and publicly available databases can be readily adapted to meet tribal needs. Many of these resources are available online.

It is necessary to know what type of data is available before that data can be used. The Regulatory Knowledge Core of CRCAIH has helped tribal partners catalog previous research to provide a listing of studies that have taken place. This method of cataloging studies allows the tribe to identify research studies that have been completed, along with information about those studies.

When analyzing data returned from prior research studies it might be useful to contact the original researcher. They are familiar with the data and can help to identify whether data from their study will be useful for the potential secondary use. They also may have other potential resources that could assist with the proposed analysis.

B. Regulatory Concerns

Using research data to explore topics or questions that were not included in the original research project can lead to several questions.

- Can this data be used to explore additional topics and questions?
- What are the terms of the consent form from the original research project?
    - Is additional consent required?
- Are there additional regulatory approvals that are needed to use the data?

In addition to knowing the type of data that will be used, it is also important to know how data is going to be used for secondary analysis.

Data from completed research studies can potentially be repurposed to provide additional information for tribes. However, before this data is analyzed for another purpose, the intent of the original research study must be thoroughly examined. Additional use of individual level data needs to be consistent with the original consent. If it is not consistent with the original consent, then the secondary analysis cannot take place without approval of the study both by review boards and the individual participants.

Approval for secondary analysis of individual level data also depends on whether or not participants can be identified, how data is shared, who owns the data, as well as how research policy is enacted. If personal information that could be used to connect data with a specific person has been removed from the data set, then it is referred to as de-identified data.

- If data is de-identified:
    - No longer meets the definition of human subjects research from the Office for Human Research Protections
    - May not need approval from some review boards
    - May need tribal community approval
- If data is identified:
    - May need additional approvals for use – would likely need to go back to all original review boards
    - May need tribal community approval
    - May need additional consent from the study participants

C. Uses for Prior Research Data

Aggregate data can assist in development of community assessments and grant proposals. When multiple data sets are compiled over time, it allows the tribe to create a snapshot of what is happening in health research. If there are collections of data that track certain factors over a long period of time, a compiled data set could be used to examine trends over time.

Community assessments involve community level data and can include a number of different topic areas such as health, environment, housing, etc. The most common way to collect data for a community assessment involves a survey of individuals within the community. Previous research data can help inform the design of the study and highlight important issues that should be addressed by items on the survey. One example of this is developing questions for a community assessment by using data from a prior study that performed focus groups on health care availability.

Additionally, aggregate data can also be used as preliminary data for grant proposals. Community level data can identify areas of importance for future research and help the grant writer present a case that a research study is needed in a particular area in a particular community. Having local data helps create a strong grant proposal.

D. Software for Data Analysis

Research data is usually stored in databases with multiple records and variables. In order to analyze the data to discover patterns or calculate statistics, computer software is generally used. There are many different software packages available for data analysis. Choosing a software package that meets the data analysis needs of the tribe is dependent on many factors. One of these factors is price. Statistical software packages can range in price from free to thousands of dollars per license. Another factor is the ease of use. This can depend on the user's level of comfort with the software program, the availability of training and support, and much more. The Methodology Core has several trainings on using computer software for data analysis available on the CRCAIH website. Methodology Core members also have expertise with many statistical software packages and can help provide guidance in choosing the package that would be a good fit.

VI. Creating a Data Management Plan for Research Data, Data Return, Data Storage & Secondary Data Analysis

Once decisions have been made regarding research data, it is important to document those decisions in a data management plan. This document is valuable for maintaining consistency over time with respect to overall management of research data. The document should contain a rationale section for the policies and procedures that are specific to the tribe. This will help to maintain a historical perspective for present and future maintenance of research data. A data management plan should be consistent with tribal IRB policy and procedures, as well as tribal codes or laws regarding research. An example of the sections to include in a data management plan is presented in Figure 3 on the next two pages.

Specific procedures should also be outlined in the plan. Procedures should contain information about people responsible for specific tasks and a timeline. We recommend creating sections of the data management plan to cover the definition of data, return of data, storage of data, and secondary data use. Additional sections can be added if necessary. In the plan, individuals with data responsibilities should be identified by position instead of by name. This leads to ease of transition if titles or personnel change over time.

The plan should also be a working document that can be easily changed if needed. Policy and technology changes may influence the way procedures are implemented over time, and the data management plan should be easy to modify to address these changes. If changes are made, careful consideration must be given to the effects that the changes have on existing data. A periodic audit can help verify that each of the procedures listed in the plan align with procedures that are actually followed.

Figure 3: Example Data Management Plan for Research Data

CRCAIH

Example Data Management Plan for Research Data

1. Overview of Research Data

Maintaining data for research projects will be primary the responsibility of the Research Office.

a. Policies regarding Research Data
   Policy gives the Tribal Research Office responsibility to manage research data

b. Roles and Responsibilities
   The Research Data Manager is the person in the Tribal Research Office that is responsible for gathering research data from completed studies, verifying that requested data has been returned, storing data according to specifications, and backing-up the data according to the documented process.
   The Secondary Data Manager is the person identified by the Tribal Research Office that has access to all research data and is familiar with all policies and procedures for maintaining research data. This person is not primarily responsible for research data.

2. Return of Data

Upon completion of a research study, the

a. Types of Data
   Researchers are responsible for data collection and analysis for their research project. They shall carefully keep any identifiable information confidential and remove any personal identifiers according to regulations. The researcher will store all physical and electronic data in secure locations that can only be accessed by identified personnel. Upon termination or completion of the study, the principal investigator is responsible to return all physical and electronic data to the Research Data Manager with the Tribal Research Office.

b. Standards for Data and Metadata
   Any physical data shall be returned using a secure option through a reputable delivery service. All electronic data shall be returned in duplicate on identical password protected flash drives that are mailed securely through a reputable delivery service. The data that shall be included on the flash drive includes:
   - A Directory (folder) named DATA which contains de-identified data in a universal format. Software specific format can also be included but is not required.
   - A Directory named CODEBOOK AND INSTRUMENT that contains a codebook or data dictionary in a universal format (.PDF and/or .TXT) as well as an original copy of the instrument that was used to collect the data.
   - A Directory named FORMS that contains other relevant IRB forms related to the project.
   - A Directory named RESULTS that contains any presentations or publications that were a direct result of the project.

3. Data Storage

a. Data Access and Security
   The Research Data Manager, Secondary Data Manager, and IT Director have access to storing and retrieving all research data.

b. Location
   Physical data will be stored in a secure storage location in a fireproof/waterproof safe. Electronic data will be stored on the secure file server located in the Research Office.

c. Data Backup
   Physical data will not be backed up.
   Electronic data will be backed up utilizing a cloud-based system. Backup will be scheduled for every Friday night using the scheduler included in the service. The Research Data Manager and IT Director will verify on Monday morning that the backup for the previous week occurred properly and manually backup to the service if there were problems.

d. File/Folder Naming
   For organization and ease of retrieval, the following file and folder naming scheme will be implemented:
   i. All information will be stored in a directory named PROJ0000000_InvestigatorName where the 0000000 is the unique identifier for the project in the catalog
   ii. The contents of the returned flash drive containing the folders DATA, CODEBOOK and INSTRUMENT, FORMS, and RESULTS shall be copied directly into the directory for a study.

4. Secondary Data Analysis

In order to utilize prior research data, a request letter should be sent to the Institutional Review Board Chair (IRB) as well as the Research Data Manager. The letter should request specifically what data is requested to be used, the purpose for the request, and other relevant information. The IRB Chair and the Research Data Manager will then determine whether to allow the data to be accessed and utilized, as well as if any additional regulatory actions are required to use the data.

VII. Summary

Tribal research data is extremely valuable and can be a useful tool for assisting with making decisions, tracking outcomes over time, and establishing baseline values while applying for grant funding, among other uses. In order to utilize research data it is important to communicate with researchers that are collecting the data. Researchers should be made aware of what needs to be returned to the tribe and when it needs to be returned. There should be a process that is clearly defined for storing the data, including individuals that have access to the data, security measures, storage location, data organization processes, and data backup procedures. Uses of data beyond the original research project should also be considered. Once all of these items have been addressed, they should be documented in a data management plan.

The CRCAIH Methodology Core has many resources and trainings that are available to assist with this entire process. Please contact the CRCAIH Methodology Core with questions about any items addressed in this toolkit. Contact information can be found on the CRCAIH website at www.crcaih.org.

References

Cross, T., Fox, K., Becker-Green, J., Smith, J., & Willeto, A. A. (2004). *Case Studies in Tribal Data Collection and Use.* National Indian Child Welfare Association.

NIH Data Sharing Policy and Implementation Guidance. (2003, March 5). Retrieved from National Institutes of Health Office of Extramural Research: http://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm

Organizing Your Data. (n.d.). Retrieved from Boston University Libraries Research Data Management: http://www.bu.edu/datamanagement/outline/elements/organize/

Suggested citation: Birger, C., Puumala, S., Around Him, D., & Villalobos, J. (2016). *CRCAIH Research Data Management Toolkit*. Collaborative Research Center for American Indian Health.

Collaborative Research Center for American Indian Health • Center for Health Outcomes and Prevention Research • 2301 East 60th Street North • Sioux Falls, SD 57104-0589

info@crcaih.org • www.crcaih.org • (605) 312-6232

Project is supported by the National Institute on Minority Health and Health Disparities (NIMHD) of the National Institutes of Health (NIH) under Award Number U54MD008164