Responsible Conduct of Research Program
UT Arlington

Workshop on Data Management

Dr. Jim Grover, Associate Dean of Science

Outcomes

1. Apply principles of research integrity to working with data generated in research studies.
2. Identify phases of working with data and the corresponding best practices for ensuring integrity and utility of research data.
3. Work with organizations, institutions, research teams, and other stakeholders to implement policies, procedures, and research activities that ensure the integrity and utility of research data.

Two useful resources

1. Guidelines for Responsible Data Management in Research, Office of Research Integrity, US DHHS.
2. Data Guidance, Office of Research, UTA.

What is data?

Data embraces any collection of facts, measurements, or observations. Different disciplines have different notions of what constitutes data, ranging from material created in a wet laboratory, such as an electrophoresis gel or a DNA sequence, to that obtained in social-science research, such as a filled-out questionnaire, video or audio recordings, or photographs. Data can be astronomical measurements, microscope slides, climate patterns, cell lines, field notes, soil samples, or results of statistical analyses.

from: *UTA Data Guidance,*
Office of Research

Desirable qualities of data

- Integrity
- Utility

Data management to ensure integrity and utility – in the past

Craft model of science:

- Projects run by principal investigator, few others involved
- Master and apprentice mode of training and management
- High reliance on individual integrity and accountability
- Idiosyncratic variations in practice

Data management to ensure integrity and utility – as it is now

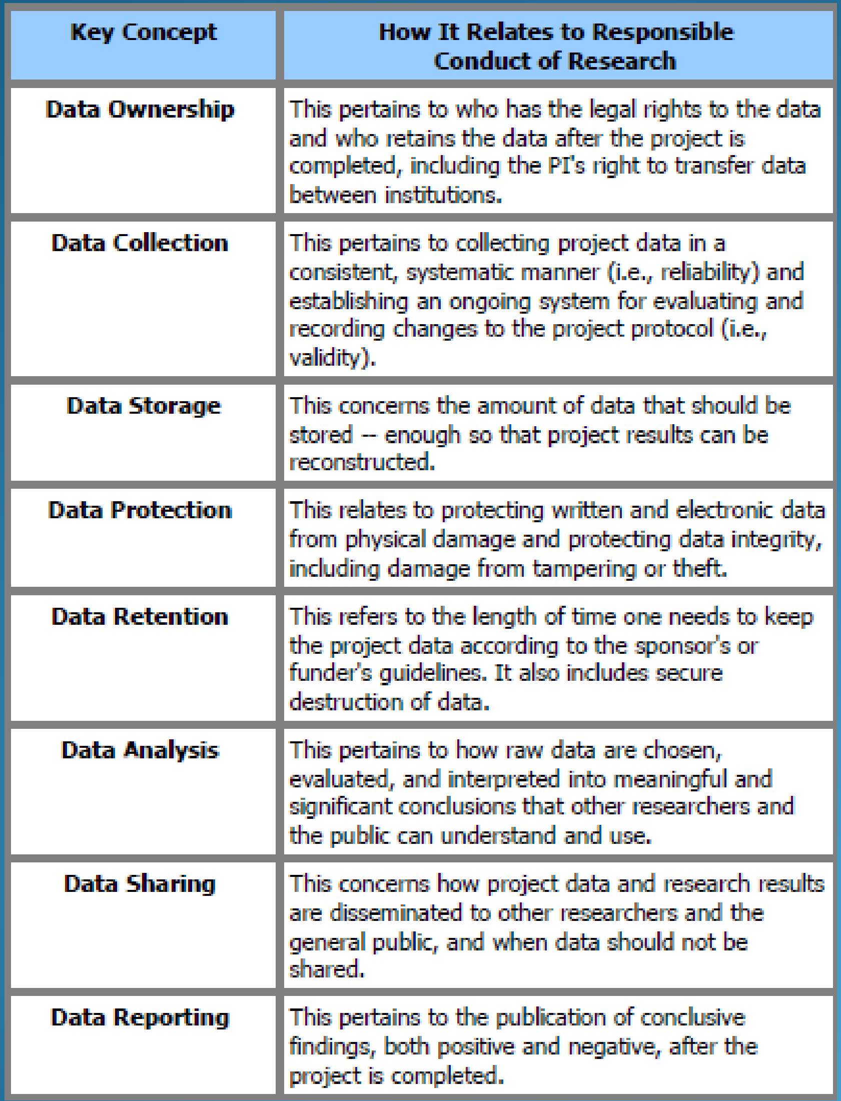
Organizational model of science:

- Projects run by principal investigator, but with many collaborators, employees, and stakeholders
- Formal and organized modes of training and management
- High reliance on organizational accountability and standardized practice
- Role of individual integrity and responsibility can be underemphasized

Data management to ensure integrity and utility – as you move through your career

- You will probably work in several different settings (employers, research group organizations, etc.)
- Each setting will have different rules, regulations, policies, procedures, and practices
- Each setting will assign responsibilities for ensuring research integrity in different ways
- In each setting, it is your professional responsibility to understand the players and the rules, and your personal responsibility to ensure integrity
- This requires understanding general principles and concepts of responsible conduct of research (more than just following the rules)

Phases of working with data



Key Concept	How It Relates to Responsible Conduct of Research
Data Ownership	This pertains to who has the legal rights to the data and who retains the data after the project is completed, including the PI's right to transfer data between institutions.
Data Collection	This pertains to collecting project data in a consistent, systematic manner (i.e., reliability) and establishing an ongoing system for evaluating and recording changes to the project protocol (i.e., validity).
Data Storage	This concerns the amount of data that should be stored -- enough so that project results can be reconstructed.
Data Protection	This relates to protecting written and electronic data from physical damage and protecting data integrity, including damage from tampering or theft.
Data Retention	This refers to the length of time one needs to keep the project data according to the sponsor's or funder's guidelines. It also includes secure destruction of data.
Data Analysis	This pertains to how raw data are chosen, evaluated, and interpreted into meaningful and significant conclusions that other researchers and the public can understand and use.
Data Sharing	This concerns how project data and research results are disseminated to other researchers and the general public, and when data should not be shared.
Data Reporting	This pertains to the publication of conclusive findings, both positive and negative, after the project is completed.

From: Guidelines for Responsible Data Management in Research, Office of Research Integrity

Data Ownership

Ownership implies rights and responsibilities.

Parties with some degree of ownership:

1. Research institution (university)
2. Funding agency
3. Principal investigator
4. Human subjects

UTA Guidance

Ownership issues: Data "ownership" generally refers to both the possession of and responsibility for information. As a legal concept, it embraces the range of rights and obligations with respect to a data collection, including rights and obligations to share. All investigators and research staff should review the institution's policies with respect to data ownership, to make sure their understanding matches the institution's. If a specific third-party sponsor is involved, the sponsor / granting agency may set out the terms of copyright.

Notebooks and journals: Data and data books collected by undergraduates, graduates, and postdoctoral fellows on a research project generally belong to the grantee institution, or the PI under conditions described above. In any case, students should generally not assume that it will be permissible to take "their data" when they leave. Appropriate arrangements need to be made in advance. If the faculty PI does not raise the issue, the student or fellow must. Usually arrangements may be made to take copies of the data when they leave.

Data Collection

Good data collection involves

- Good scientific and technical methodology
- Good record-keeping

Make sure you know what you're doing

When in doubt, record more information rather than less.

Date and cross-reference records, so that you can backtrack.

Bound notebooks are a good "master" record.

UTA Guidance

Different disciplines have preferences for different approaches, and for what constitutes acceptable "rigor" for reliability and validity of results. This is one reason why a careful prior review of the existing literature on a topic is imperative when designing a research protocol.

Data collection methods vary by discipline, and according to the data types of interest; but the emphasis on ensuring accurate and honest collection remains the same.

It is critical that researchers have sufficient methodological skills to assure the quality of data collection efforts. Everyone who participates in the investigative effort should be trained in the methods. Where possible, researchers should try to build checks-and-balances into the collection process.

Data Storage, Protection, and Retention (Disposal)

Data should be stored :

- For use by original investigation team and others
- For as long as required by sponsors or owners for their use
- In a way that is secure, but also available to those who need it
- With adequate backups

Confidential information should be de-identified or securely disposed of when no longer needed

Long-term archiving of (some) raw data may be unnecessary, but reconstruction and validation of a study’s conclusions should be possible.

Journals, scientific societies, and some other non-profit organizations offer data archiving and sharing services.

Data Storage, Protection, and Retention (Disposal)

UTA Guidance

Storage and Protection: In information security, it is conventional to speak of three core goals for information protection:

- Confidentiality - limiting information access and disclosure to authorized users;
- Integrity - ensuring that data is not changed inappropriately after recording, whether by accidental or deliberate activity. Also, the notion that the person or entity in question entered the right information - that is, that the information reflected the actual circumstances ("validity") and under the same circumstances would generate identical data (what statisticians call "reliability").
- Availability - refers to the availability of information resources to authorized users.

Everyday risks like fire, water or other environmental damage, or simple technical failures like hard disk crashes, must be considered. It's an essential practice to make frequent, periodic backup copies of a data collection, and store these copies in a secure secondary location that is protected both from intruders and environmental threats.

Retention and disposal: Data handling procedures should describe when, how, and who may handle data for storage, retrieval, sharing, archiving and disposal purposes. These procedures may depend on the nature of the project, the cost of maintaining that data, research sponsors' requirements, etc.

Retaining data on paper files and electronic media long past the end of a project can increase the chances of unauthorized access. Disposal of sensitive data requires care and technical expertise to ensure that the information could not be reconstructed from the storage media.

Data Storage, Protection, and Retention (Disposal)

Data storage options are evolving

- personal or university computers
- detachable drives
- university owned servers
- third-part servers (e.g. “cloud”)

General advice

- university owned or contracted resources are preferred (required for some types of data)
- back up to at least one appropriate server resource (e.g. J-drive, UTA Box)
- seek advice for data needing special consideration (human subjects, medical or educational records, intellectual property)

Data Analysis

Good data analysis requires good scientific and statistical judgment.

Analytical strategies and methods should be considered prior to data collection, in the design of the research.

Exclusion of data from analysis should be based on problems recorded during data collection or other objective indicators.

Use of non-standard analytical approaches requires justification and clear explanation.

UTA Guidance

Like data selection criteria, the choice of statistical analysis methods should always precede data collection. Waiting until later in the research process increases the risk that analytic decisions will be driven by consideration of which produces the most favorable results. Any bias occurring in the collection of the data, or selection of method of analysis, will increase the likelihood of drawing a biased inference. Every field of study has developed its accepted practices for data analysis; if an unconventional approach is used, it is crucial to clearly state this is being done and show how this new and possibly unaccepted method of analysis is being used, as well as how it differs from other more traditional methods. Whether statistical or non-statistical methods are used, researchers should be clear - to themselves and to the persons to whom the analyses are presented - of the limitations and possible biases of their methods.

Data Sharing and Reporting

- Reporting by publication in scholarly literature is expected for most academic research.
- Additional reporting may be required by some sponsors.
- Reporting should be sufficiently complete so that a competent colleague could repeat the study.
- Reporting should not exclude data and results that are “unfavorable” to the investigators.
- Increasingly, the sharing of data (raw or summarized) is also expected, especially after publication.
- Practices for sharing of data are evolving, and involve complex issues, especially for the sharing of data prior to publication.

Data Sharing and Reporting

ORI Guidance:

Sharing Data Prior to Publication

Before publication, there is often no obligation to share any preliminary data that have been collected. In fact, sharing at this stage is sometimes discouraged because of the following reasons:

- The implications for a set of data may not be understood while a project is still in progress. By waiting until a project is ready for publication, researchers ensure that what they share has been carefully reviewed and considered.
- There is fear that less scrupulous researchers will use shared research results for their own gain.

However, in some cases preliminary data should be shared immediately with the public and/or other researchers since it would be of immediate benefit (e.g., if a research project found that a new drug placed subjects at grave risk or greater benefit).

Many researchers find it worthwhile to present preliminary findings in a conference setting before the study is complete to inform peers about their forthcoming research.

Data Sharing and Reporting

ORI Guidance:

Sharing Data After Publication

After a project's research has been published or patented, any information related to the project should be considered open data. Other researchers may request raw data or miscellaneous information related to the project in order to verify the published data or to further their own research project. However, each project should evaluate its ability to share raw data in terms of specific needs and budget constraints.

Data Sharing and Reporting

UTA Guidance:

The practice of ensuring research integrity extends to the stage of documenting and preparing results for publication. Publishing in peer-reviewed journals or presenting in scholarly meetings is the primary mechanism for investigators to disseminate their findings to the research community. This community relies on authors to report the events of a study honestly and accurately. All researchers should be aware of the issues that compromise the integrity of data reporting and publishing:

- Misrepresentation of data quality, or of the data itself.
- Analysis of data by several methods to find a significant result.
- Fabrication or falsification of data.
- Inadequate evaluation of prior research.
- Misleading discussion of observations.
- Reporting conclusions that are not supported.
- Failure to disclose conflicts of interest.
- Plagiarism.
- Unjust attribution of authorship.

Data Sharing and Reporting

Data sharing is evolving

- major federal agencies and other research sponsors now require data management plans that address all aspects of data handling
- major federal agencies and other research sponsors now require sharing post-project data
- specific repositories are required by different sponsors
- UTA Library provides guides for data management plans and data sharing (Scholarly Communications Division)

Discussion questions…

True or False: In scientific research, only the information and observations that are made as part of scientific inquiry are considered data.

__True

__False

From: Guidelines for Responsible Data Management in Research, Office of Research Integrity

Discussion questions…

True or False: In scientific research, only the information and observations that are made as part of scientific inquiry are considered data.

__True

__False

What ORI says:

Answer: False. In fact, data also include the materials, products, procedures, and other data sources that are part of the research project. Essentially, data are considered to be anything and everything that informs the way in which individuals are able to understand and to process their world.

From: Guidelines for Responsible Data Management in Research, Office of Research Integrity

Discussion questions…

Dr. Smith works at The University and is the Principal Investigator on a large research project that is funded by the National Institutes of Health (NIH). However, while Dr. Smith wrote the original grant proposal, he does very little day-to-day work on the project. Instead, the Research Director, Betsy, oversees all aspects of the project, including staff supervision and all data management activities. In addition, Betsy has been lead author on several publications about the project's research findings.

Who owns the project and its data?

__ The PI, Dr. Smith
__ The Research Director, Betsy
__ The University
__ The National Institutes of Health
__ No one person or organization

From: Guidelines for Responsible Data Management in Research, Office of Research Integrity

Discussion questions…

Dr. Smith works at The University and is the Principal Investigator on a large research project that is funded by the National Institutes of Health (NIH). However, while Dr. Smith wrote the original grant proposal, he does very little day-to-day work on the project. Instead, the Research Director, Betsy, oversees all aspects of the project, including staff supervision and all data management activities. In addition, Betsy has been lead author on several publications about the project's research findings.

Who owns the project and its data?

__ The PI, Dr. Smith
__ The Research Director, Betsy
__ The University
__ The National Institutes of Health
__ No one person or organization

What ORI says:

Answer: The University. Despite the PI's and the Research Director's work on the project, the sponsoring institution typically maintains ownership of a project's data as long as the PI submitted the grant through that institution and is employed by them. However within the sponsoring institution, a PI is generally granted stewardship over the project data; he/she may control the course, publication, and copyright of any research, subject to institutional review.

From: Guidelines for Responsible Data
Management in Research, Office of
Research Integrity

Discussion questions…

Part of the data collection methodology for Dr. Smith's study includes distributing a 12-page self-administered questionnaire to participants; they must fill out and initial each page of the questionnaire to confirm completion. One day on his way home from conducting an interview with a subject, the Research Assistant, Joel, needed to write directions for a friend and he reached in his bag and grabbed the first piece of paper that he could find. Joel accidentally ripped the back page off of one of the completed questionnaires to write the directions, which he then gave to his friend. He didn't realize this until a few hours later, when he was reviewing the data that he had collected that day. Joel thought that he remembered the participant's answers on the last page of the survey, since they were mostly demographic questions.

What should Joel do?

__ Staple on a new page and fill out the subject's responses, since he remembers them.

__ Contact the subject and ask her to complete the last page of the questionnaire again.

__ Omit the participant's questionnaire from the study, his/her partial data is invalid.

__ Just pretend like he doesn't know what happened to the last page.

Discussion questions…

Part of the data collection methodology for Dr. Smith's study includes distributing a 12-page self-administered questionnaire to participants; they must fill out and initial each page of the questionnaire to confirm completion. One day on his way home from conducting an interview with a subject, the Research Assistant, Joel, needed to write directions for a friend and he reached in his bag and grabbed the first piece of paper that he could find. Joel accidentally ripped the back page off of one of the completed questionnaires to write the directions, which he then gave to his friend. He didn't realize this until a few hours later, when he was reviewing the data that he had collected that day. Joel thought that he remembered the participant's answers on the last page of the survey, since they were mostly demographic questions.

What should Joel do?

__ Staple on a new page and fill out the subject's responses, since he remembers them.
__ Contact the subject and ask her to complete the last page of the questionnaire again.
__ Omit the participant's questionnaire from the study, his/her partial data is invalid.
__ Just pretend like he doesn't know what happened to the last page.

What ORI says:

Answer: Omit the participant's questionnaire from the study, his/her partial data is invalid. This is Joel's best option - if he were to attempt to collect the data again from the subject, the subject would be responding in a different time and mood than when the original interview occurred. As part of responsible data management, honesty about the mishap is the best way to maintain the validity of the data and to clarify that the data were not tampered with or falsified in any way.

From: Guidelines for Responsible Data Management in Research, Office of Research Integrity

Discussion questions…

With the recent emergence of electronic databases, more scientific researchers are storing their data on their computer networks. However, data protection is an issue for both paper- and computer-based data. So what is the best way to protect data?

__ Strip identifiers from human subjects data.

__ Limit who has access to the data.

__ Use an encrypted password system and assign new passwords quarterly.

__ Destroy the written data after transferral to an electronic database.

From: Guidelines for Responsible Data Management in Research, Office of Research Integrity

Discussion questions…

With the recent emergence of electronic databases, more scientific researchers are storing their data on their computer networks. However, data protection is an issue for both paper- and computer-based data. So what is the best way to protect data?

__ Strip identifiers from human subjects data.

__ Limit who has access to the data.

__ Use an encrypted password system and assign new passwords quarterly.

__ Destroy the written data after transferral to an electronic database.

What ORI says:

Answer: Limit who has access to the data. This is the best way to protect data. Simple measures -- like keeping written data in a locked filing cabinet for which there is only one key -- will help minimize the chance that data could be corrupted or stolen. However, this is a complex issue and employing a multifaceted security approach is the best way to ensure that your data is protected.

From: Guidelines for Responsible Data Management in Research, Office of Research Integrity

Discussion questions…

After completing the first phase of data analysis, 1 of the 3 main hypotheses of Dr. Smith and the research team was proven correct. However, the team also found some results from another facet of the project that they were not expecting. While these secondary results do not directly impact Dr. Smith's primary research questions, they may affect at least 3 other investigators' research. The results appear to be pretty definitive, but data analysis is still being conducted on other parts of the project. The 2 Research Associates working on the project, Samantha and Enrique, are insistent that the team should immediately publish their findings in a journal, since the results may have implications on other PIs' work. Dr. Smith and Betsy, the Research Director, do not intend to publish any results for at least another year, since the research is ongoing and some questions are still unanswered.

What should the research team do?

__ They should publish the results in a journal as soon as possible.

__ They should tell the funding agency about the findings, and let the agency disseminate the information if it wants.

__ They should contact the other researchers to let them know the preliminary results.

__ They should do nothing; they aren't legally allowed to share their results until all data have been fully validated.

From: Guidelines for Responsible Data Management in Research, Office of Research Integrity

Discussion questions…

After completing the first phase of data analysis, 1 of the 3 main hypotheses of Dr. Smith and the research team was proven correct. However, the team also found some results from another facet of the project that they were not expecting. While these secondary results do not directly impact Dr. Smith's primary research questions, they may affect at least 3 other investigators' research. The results appear to be pretty definitive, but data analysis is still being conducted on other parts of the project. The 2 Research Associates working on the project, Samantha and Enrique, are insistent that the team should immediately publish their findings in a journal, since the results may have implications on other PIs' work. Dr. Smith and Betsy, the Research Director, do not intend to publish any results for at least another year, since the research is ongoing and some questions are still unanswered.

What should the research team do?

__ They should publish the results in a journal as soon as possible.

__ They should tell the funding agency about the findings, and let the agency disseminate the information if it wants.

__ They should contact the other researchers to let them know the preliminary results.

__ They should do nothing; they aren't legally allowed to share their results until all data have been fully validated.

What ORI says:

Answer: They should contact the other researchers to let them know the preliminary results. If Dr. Smith believes that the results would have implications on other researchers' work and he does not intend to publish for quite some time, he could send his fellow researchers some information about the preliminary results as a professional courtesy and to promote collegiality. However, according to the guidelines of responsible data management, the researchers are not obligated to share their findings while the research is ongoing.

From: Guidelines for Responsible Data Management in Research, Office of Research Integrity